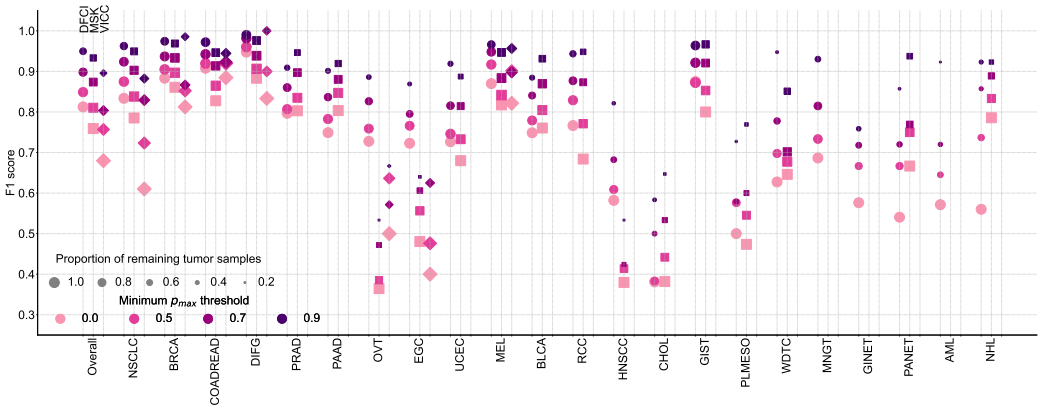
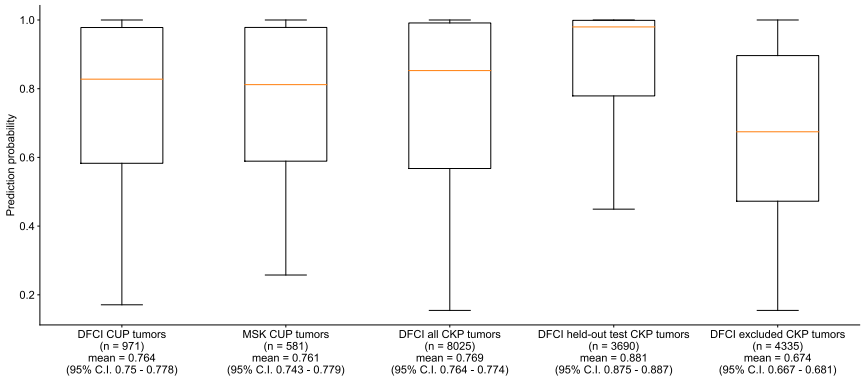
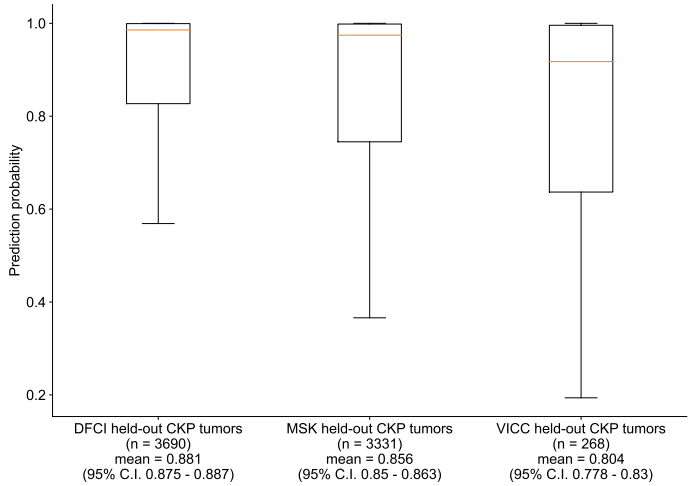
# Supplementary Materials




(a)




(b)




(c)

Figure S1

Figure S1: **OncoNPC prediction performances and confidences (i.e., $p_{\max}$) across chosen cohorts and centers. (a)** Center-specific OncoNPC performance (in weighted F1) on the test CKP tumor samples (n = 7,289). The figure is a breakdown of Fig. 2c based on cancer center (DFCI: ○, MSK: □, VICC: ◇). The performance was evaluated at 4 different prediction confidences (i.e., minimum $p_{max}$ thresholds). Each dot size is scaled by the proportion of tumor samples retained. See Table S1 for the center-specific number of test CKP tumor samples broken down by cancer types and prediction confidence thresholds. **(b), (c)** Box plots of prediction confidences ($p_{\max}$) across **(b)** DFCI CUP tumors, MSK CUP tumors, all DFCI CKP tumors, DFCI held-out CKP tumors, and DFCI excluded CKP tumors, and **(c)** DFCI held-out CKP tumors, MSK held-out CKP tumors, and VICC held-out CKP tumors. The figures display medians, lower and upper quartiles, as well as the mean and 95% confidence intervals, along with the number of tumor samples.

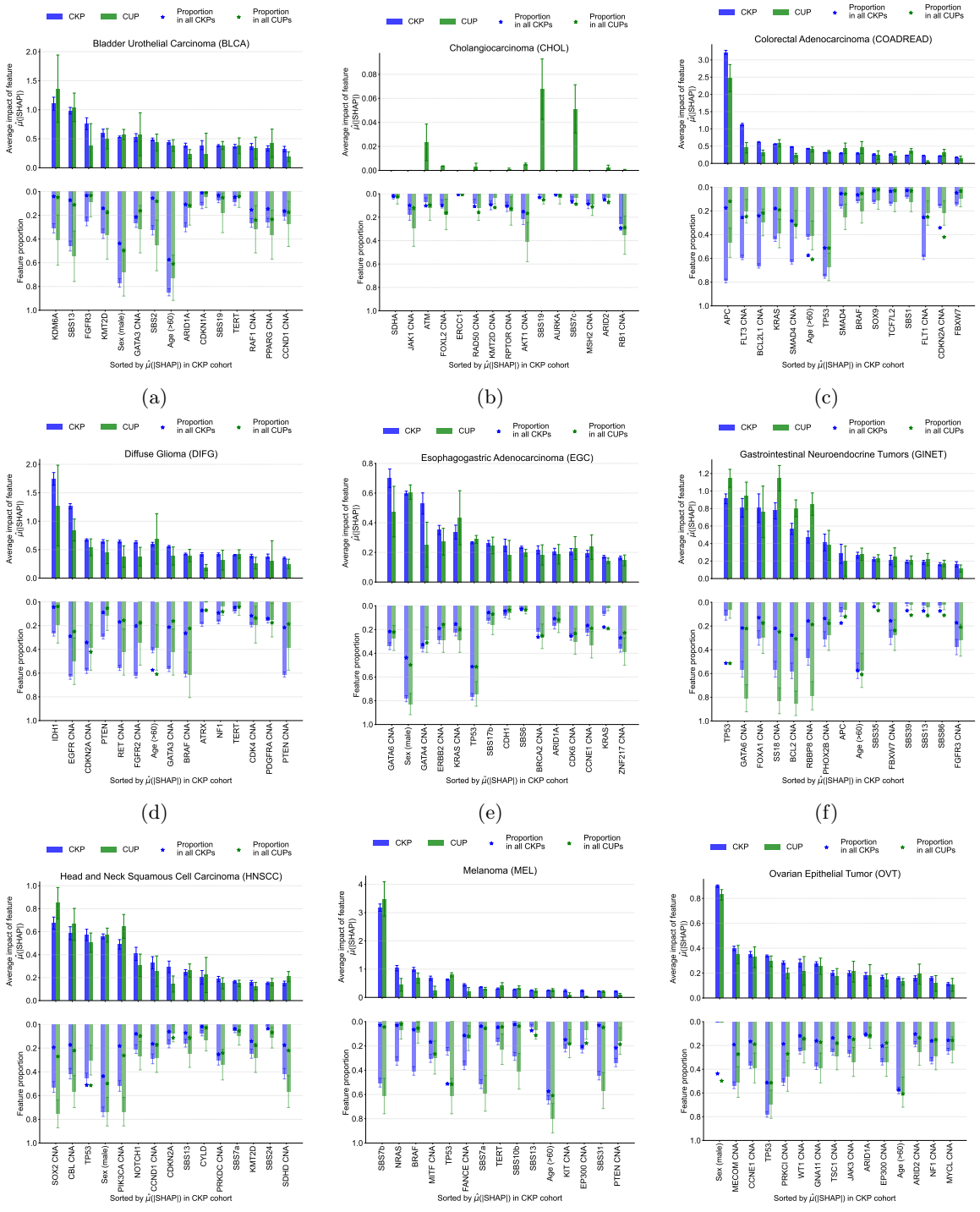
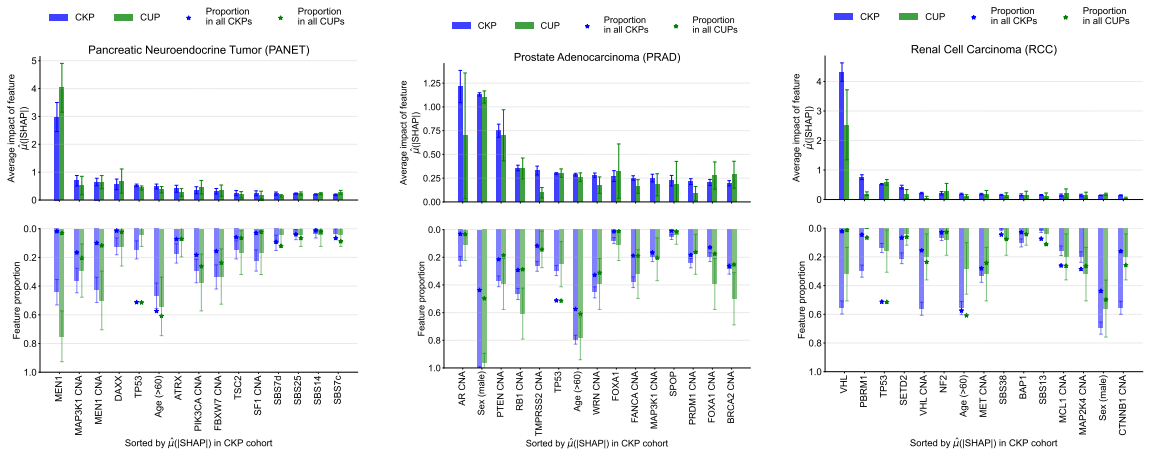

CKP
CUP
Proportion in all CKPs
Proportion in all CUPs
Bladder Urothelial Carcinoma (BLCA)
Cholangiocarcinoma (CHOL)
Colorectal Adenocarcinoma (COADREAD)
Diffuse Glioma (DIFG)
Esophagogastric Adenocarcinoma (EGC)
Gastrointestinal Neuroendocrine Tumors (GINET)
Head and Neck Squamous Cell Carcinoma (HNSCC)
Melanoma (MEL)
Ovarian Epithelial Tumor (OVT)
Average impact of feature $\hat{\mu}$(|SHAP|)
Feature proportion
Sorted by $\hat{\mu}$(|SHAP|) in CKP cohort
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

Figure S2




(j) (k) (l)

Figure S2: **Interpreting OncoNPC predictions**. Top 15 most important features, based on mean absolute SHAP values (i.e., $\hat{\mu}(|\text{SHAP}|)$ [19]), for cancer types with at least 20 CUP tumors samples were classified.

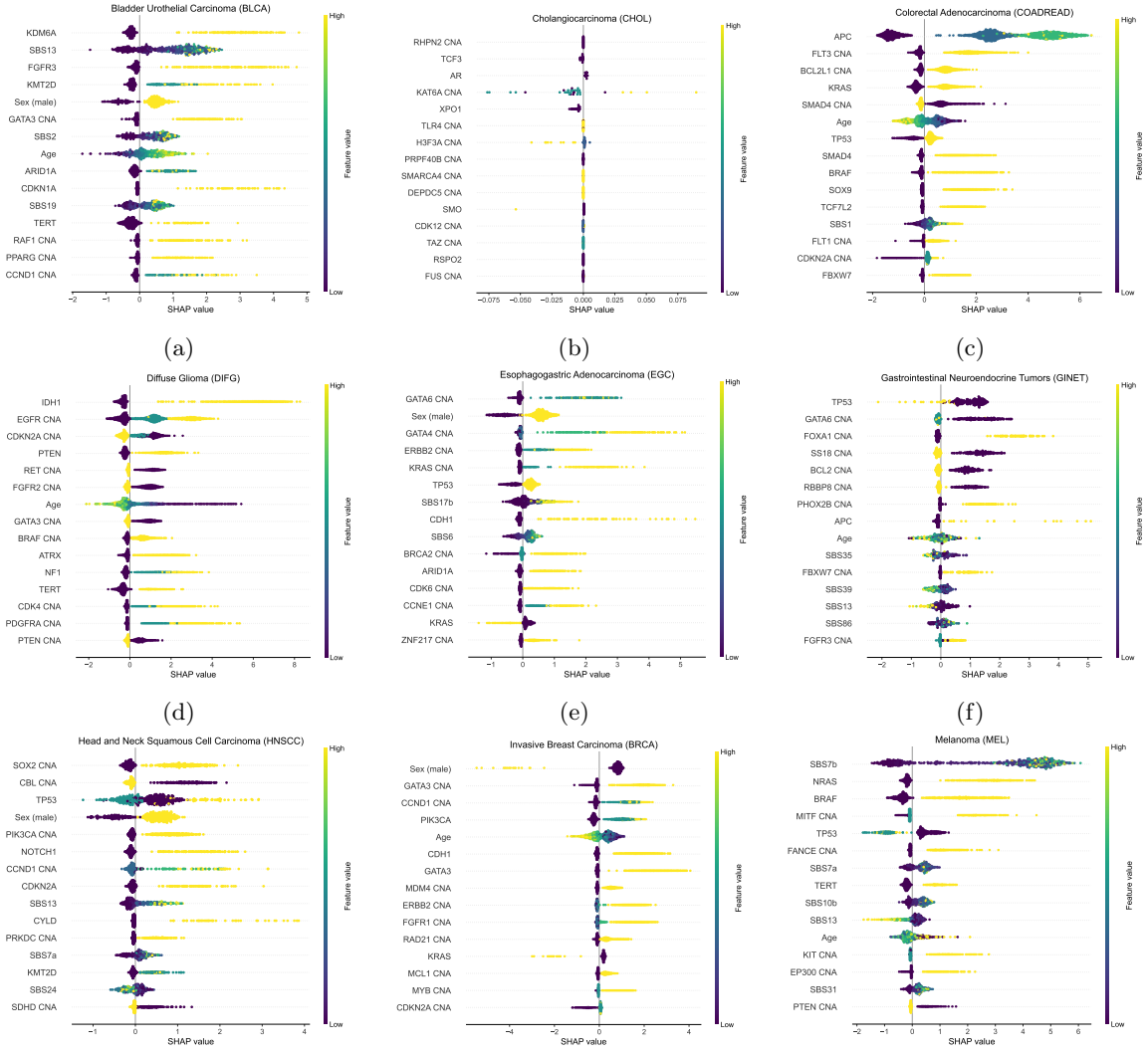

Bladder Urothelial Carcinoma (BLCA)
KDM6A
SBS13
FGFR3
KMT2D
Sex (male)
GATA3 CNA
SBS2
Age
ARID1A
CDKN1A
SBS19
TERT
RAF1 CNA
PPARG CNA
CCND1 CNA
SHAP value
Feature value
High
Low
(a)
Cholangiocarcinoma (CHOL)
RHPN2 CNA
TCF3
AR
KAT6A CNA
XPO1
TLR4 CNA
H3F3A CNA
PRPF40B CNA
SMARCA4 CNA
DEPDC5 CNA
SMO
CDK12 CNA
TAZ CNA
RSPO2
FUS CNA
SHAP value
Feature value
High
Low
(b)
Colorectal Adenocarcinoma (COADREAD)
APC
FLT3 CNA
BCL2L1 CNA
KRAS
SMAD4 CNA
Age
TP53
SMAD4
BRAF
SOX9
TCF7L2
SBS1
FLT1 CNA
CDKN2A CNA
FBXW7
SHAP value
Feature value
High
Low
(c)
Diffuse Glioma (DIFG)
IDH1
EGFR CNA
CDKN2A CNA
PTEN
RET CNA
FGFR2 CNA
Age
GATA3 CNA
BRAF CNA
ATRX
NF1
TERT
CDK4 CNA
PDGFRA CNA
PTEN CNA
SHAP value
Feature value
High
Low
(d)
Esophagogastric Adenocarcinoma (EGC)
GATA6 CNA
Sex (male)
GATA4 CNA
ERBB2 CNA
KRAS CNA
TP53
SBS17b
CDH1
SBS6
BRCA2 CNA
ARID1A
CDK6 CNA
CCNE1 CNA
KRAS
ZNF217 CNA
SHAP value
Feature value
High
Low
(e)
Gastrointestinal Neuroendocrine Tumors (GINET)
TP53
GATA6 CNA
FOXA1 CNA
SS18 CNA
BCL2 CNA
RBBP8 CNA
PHOX2B CNA
APC
Age
SBS35
FBXW7 CNA
SBS39
SBS13
SBS86
FGFR3 CNA
SHAP value
Feature value
High
Low
(f)
Head and Neck Squamous Cell Carcinoma (HNSCC)
SOX2 CNA
CBL CNA
TP53
Sex (male)
PIK3CA CNA
NOTCH1
CCND1 CNA
CDKN2A
SBS13
CYLD
PRKDC CNA
SBS7a
KMT2D
SBS24
SDHD CNA
SHAP value
Feature value
High
Low
(g)
Invasive Breast Carcinoma (BRCA)
Sex (male)
GATA3 CNA
CCND1 CNA
PIK3CA
Age
CDH1
GATA3
MDM4 CNA
ERBB2 CNA
FGFR1 CNA
RAD21 CNA
KRAS
MCL1 CNA
MYB CNA
CDKN2A CNA
SHAP value
Feature value
High
Low
(h)
Melanoma (MEL)
SBS7b
NRAS
BRAF
MITF CNA
TP53
FANCE CNA
SBS7a
TERT
SBS10b
SBS13
Age
KIT CNA
EP300 CNA
SBS31
PTEN CNA
SHAP value
Feature value
High
Low
(i)

Figure S3

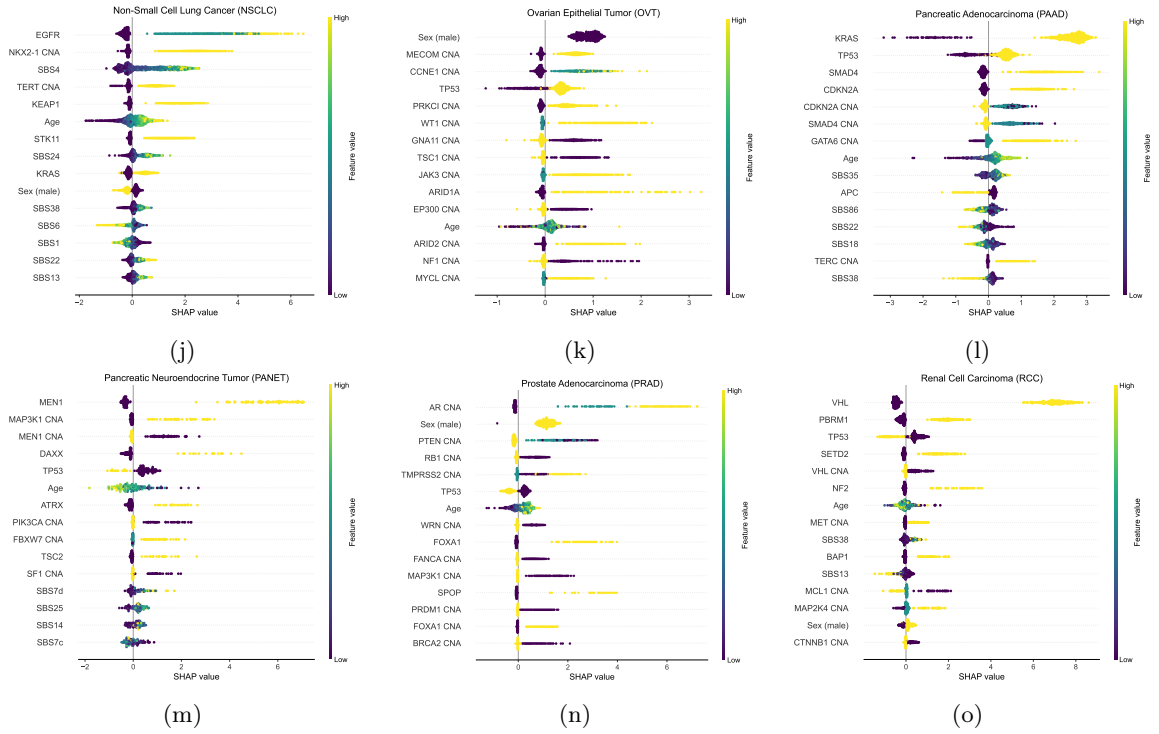



Figure S3: **SHAP summary plot** [19] for cancer types with at least 20 CUP tumors samples were classified. SAHP values (i.e., impact on OncoNPC predictions) are shown on the x-axis, while feature values are shown as a color map (from purple to yellow). In each plot, CUP and CKP tumor samples were combined into a single cohort for the corresponding cancer.

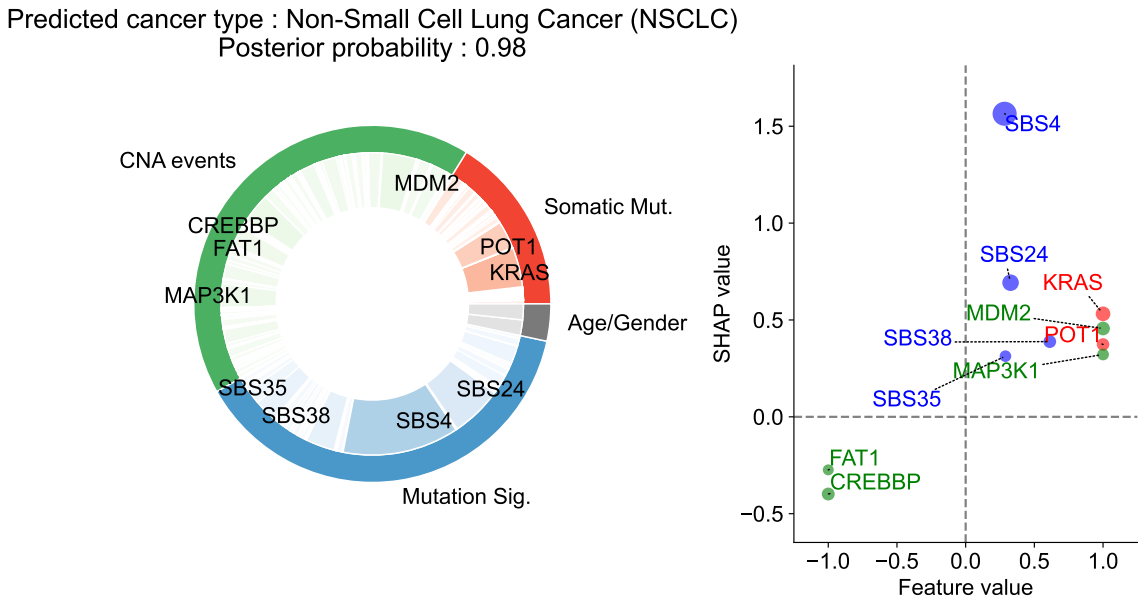
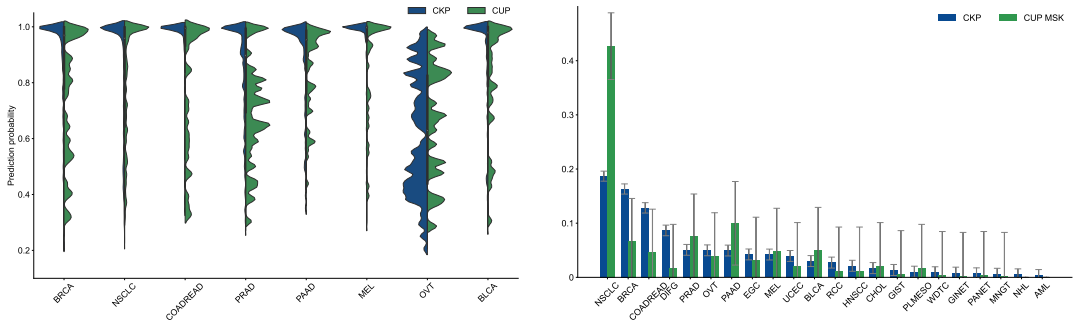



Figure S4: **Explanation of OncoNPC prediction for a patient with CUP**. The patient is a 76 year-old male, with a tumor biopsy from the liver. The pie chart on the left shows the Top 10 important features across three different feature categories (i.e., CNA events, somatic mutation, and mutation signatures), and the scatter plot on the right shows their SHAP values and feature values. The size of each dot is scaled by corresponding absolute SHAP value. From the chart review, we found that the patient reported a 60-pack year smoking history, as well as having lived near a tar and chemical factory as a child. Despite the CUP diagnosis, OncoNPC confidently classified the primary site as NSCLC with posterior probability of 0.98. SBS4, a tobacco smoking-associated mutation signature, was significantly enriched in the patient's tumor sample, which has, by far, the most impact on the prediction; followed by SBS24 mutation signature associated with known exposures to aflatoxin [20]; and KRAS mutation. Note that inhalation of aflatoxin has been linked to cause primary lung cancer [60–62], and KRAS mutation is one of the most common drivers of NSCLC [63, 64].




(a) (b)

Figure S5: **Applying OncoNPC to MSK CUP tumor samples. (a)** Empirical distributions of prediction probabilities for correctly predicted held-out CKP tumor samples (n = 3,429) and MSK CUP tumor samples (n = 496), broken down by CKP cancer types (blue) and their corresponding OncoNPC predicted cancer types for CUP tumors (green). Only OncoNPC classifications with at least 20 CUP tumor samples are shown. **(b)** Proportion of each CKP cancer type and the corresponding OncoNPC predicted CUP cancer type. All training CKP tumor samples (n = 36,445) and all MSK CUP tumor samples (n = 581) are shown. For both **(a)** and **(b)**, the cancer types (x-axis) are ordered by the number of CKP tumor samples in each cancer type

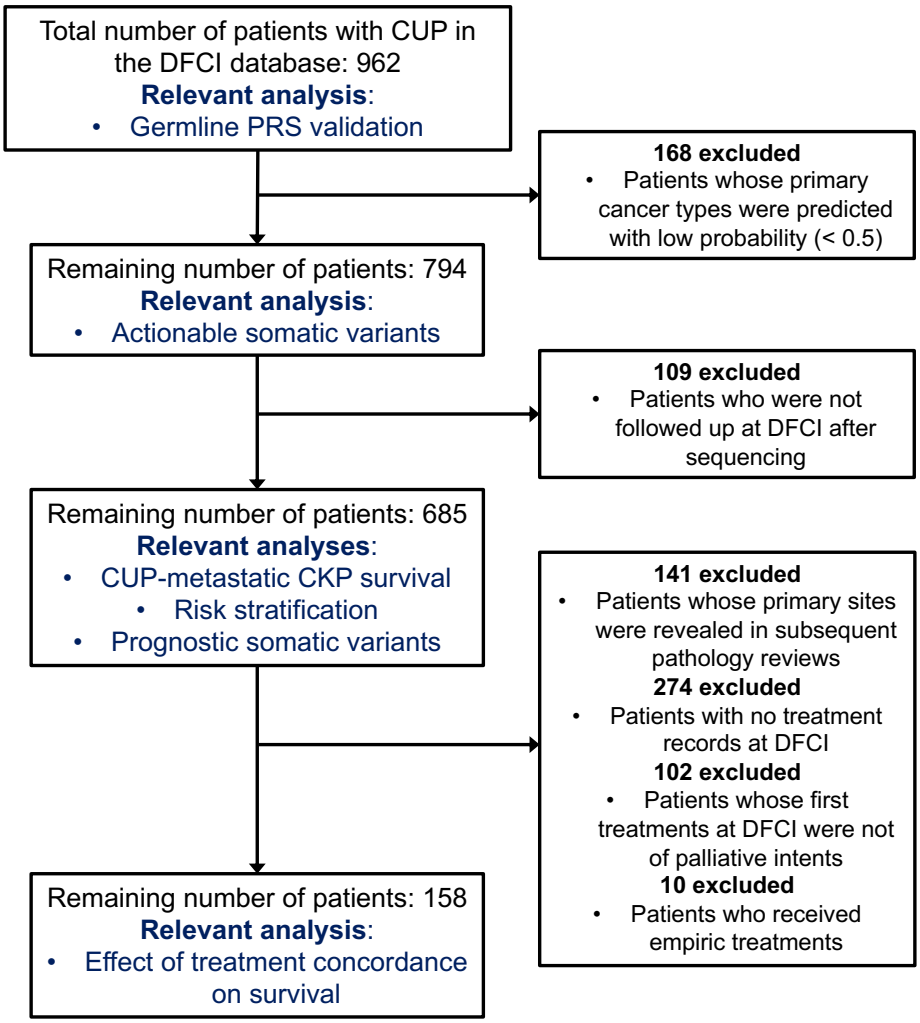
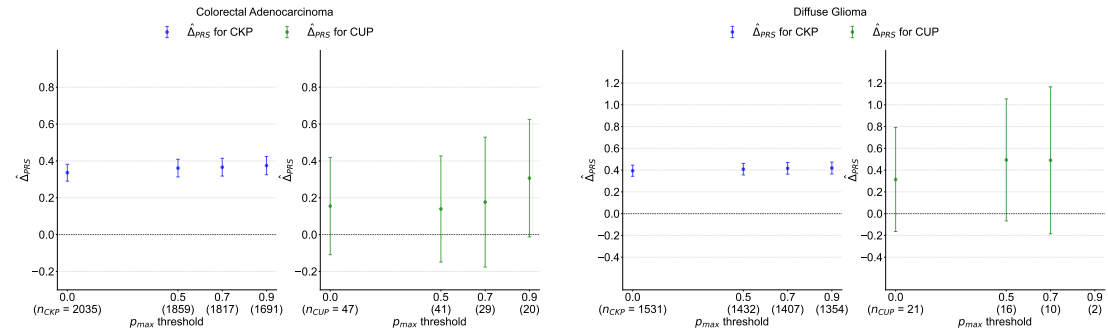


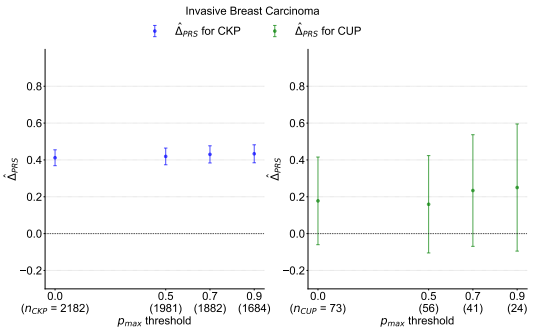

Figure S6: Exclusion criteria for downstream clinical analyses.



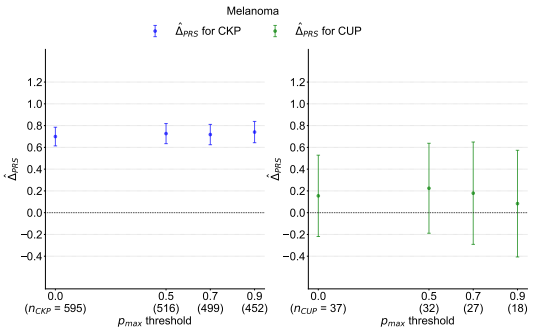

(a)



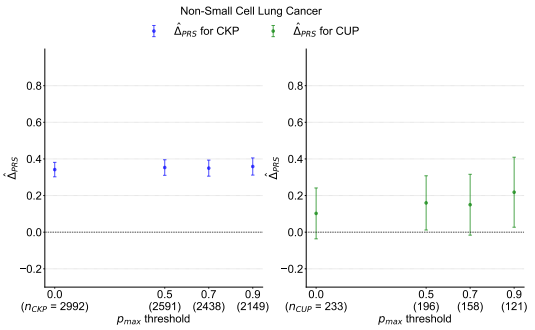

(b)



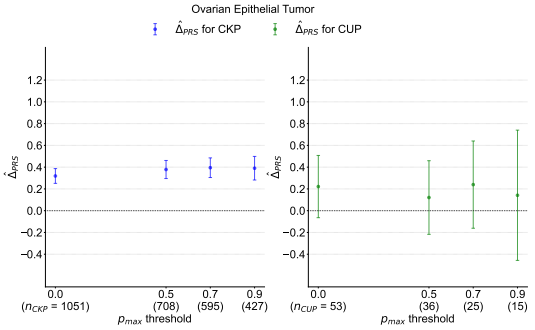

(c)



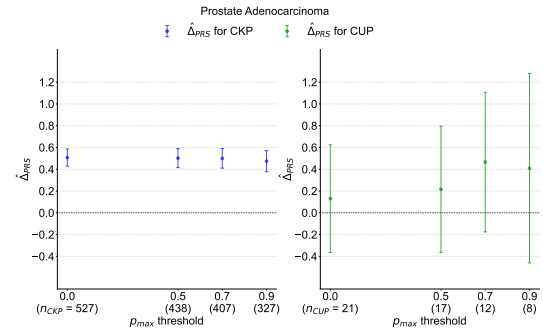

(d)



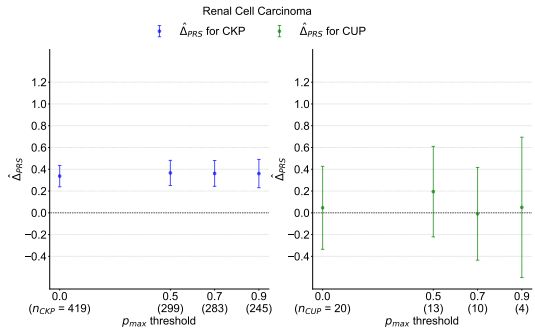

(e)




(f)



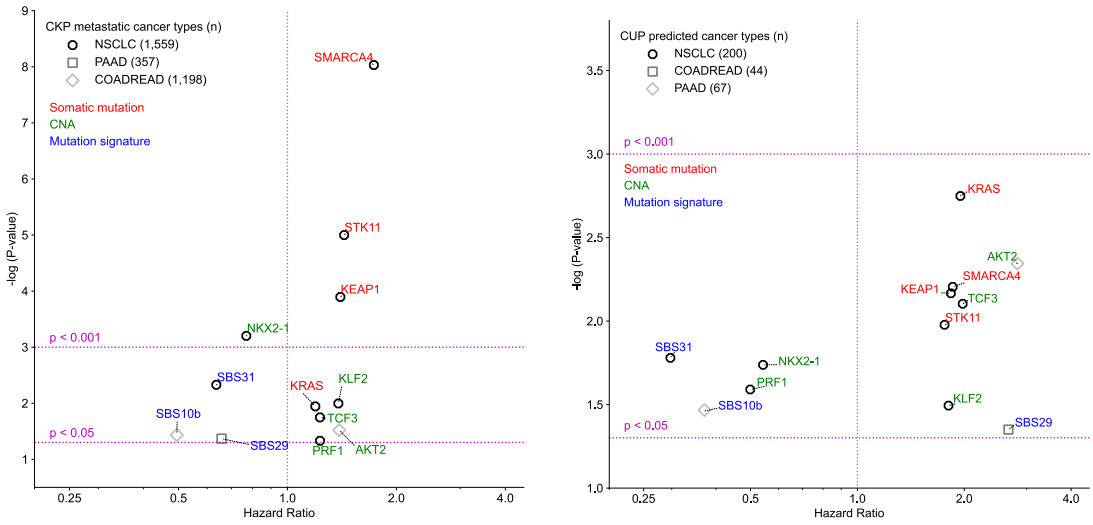

(g)




(h)

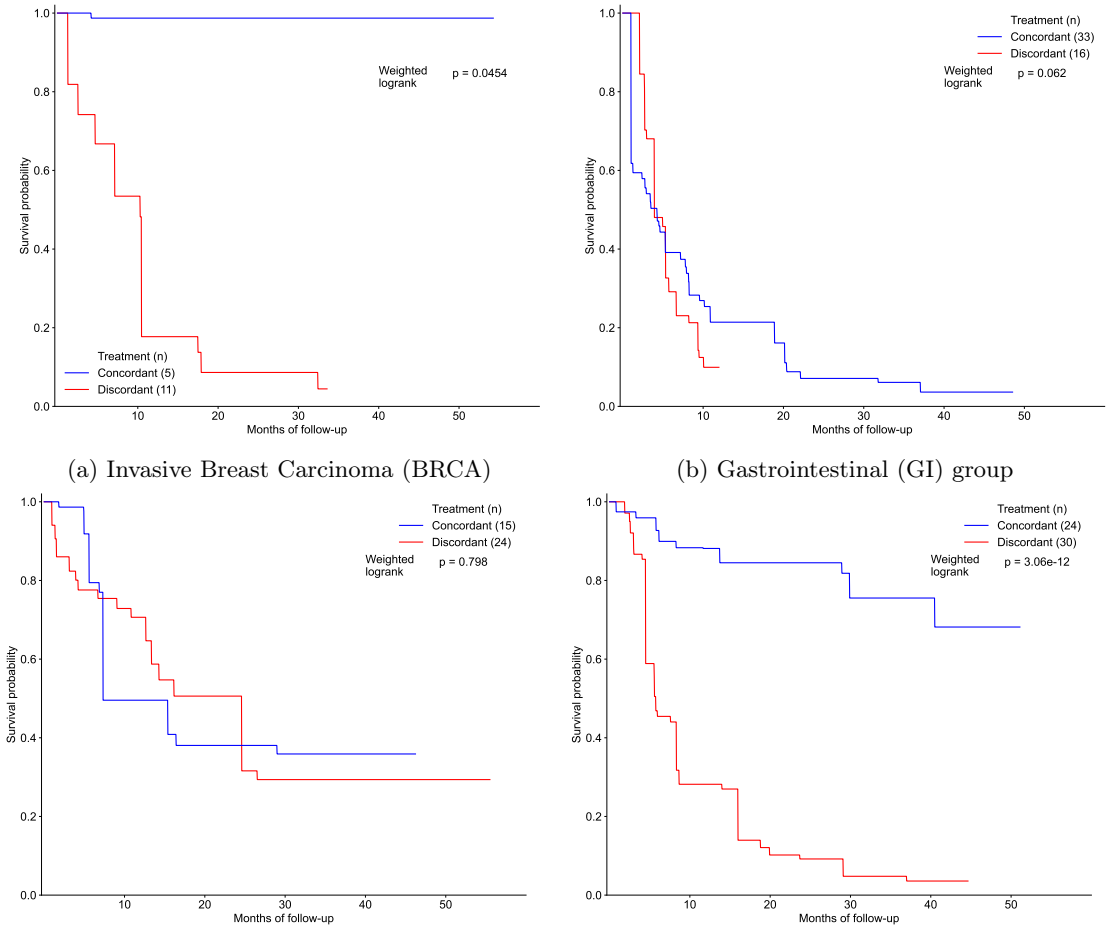
Figure S7

Figure S7: Germline Polygenic Risk Score (PRS) enrichment of CKP tumor samples and CUP tumor samples, broken down by 8 different cancer types: **(a)** Colorectal Adenocarcinoma (COADREAD), **(b)** Diffuse Glioma (DIFG), **(c)** Invasive Breast Carcinoma (BRCA), **(d)** Melanoma (MEL), **(e)** Non-Small Cell Lung Cancer (NSCLC), **(f)** Ovarian Epithelial Tumor (OVT), **(g)** Prostate Adenocarcinoma (PRAD), and **(h)** Renal Cell Carcinoma (RCC). The magnitude of the enrichment is quantified by $\hat{\Delta}_{\text{PRS}}$: the mean difference between the concordant (i.e. OncoNPC matching) cancer type PRS and mean of PRSs of discordant cancer types (see Methods). $\hat{\Delta}_{\text{PRS}}$ is shown for CKPs in blue (for reference) and CUPs in green.




(a) (b)

Figure S8: **Prognostic biomarkers between OncoNPC classifications and known cancers**. **(a), (b)** Prognostic somatic variants significantly associated with overall survival, shared between three different CUP **(a)**-metastatic CKP **(b)** pairs (NSCLC, PAAD, and COADREAD; indicated by point shape). Variant types are indicated by colors: red for somatic mutations, green for CNAs, and blue for mutation signatures.




(a) Invasive Breast Carcinoma (BRCA)

(b) Gastrointestinal (GI) group

(c) Lung

(d) Other OncoNPC cancer types

Figure S9: Estimated survival curves for patients with CUP, broken down by OncoNPC predicted cancer types: **(a)** BRCA, **(b)** Gastrointestinal (GI) group (CHOL, COADREAD, EGC, and PAAD), **(c)** Lung (NSCLC and PLMESO), and **(d)** other OncoNPC cancer types (BLCA, DIFG, GINET, HNSCC, MEL, OVT, PANET, PRAD, RCC, and UCEC). In each figure, the concordant treatment group and discordant treatment group are shown in blue and red, respectively. To estimate the survival function for each group, we utilized Inverse Probability of Treatment Weighted (IPTW) Kaplan-Meier estimator while adjusting for left truncation until time of sequencing (see Methods). Statistical significance of the survival difference between the two groups was estimated by a weighted log-rank test [59].

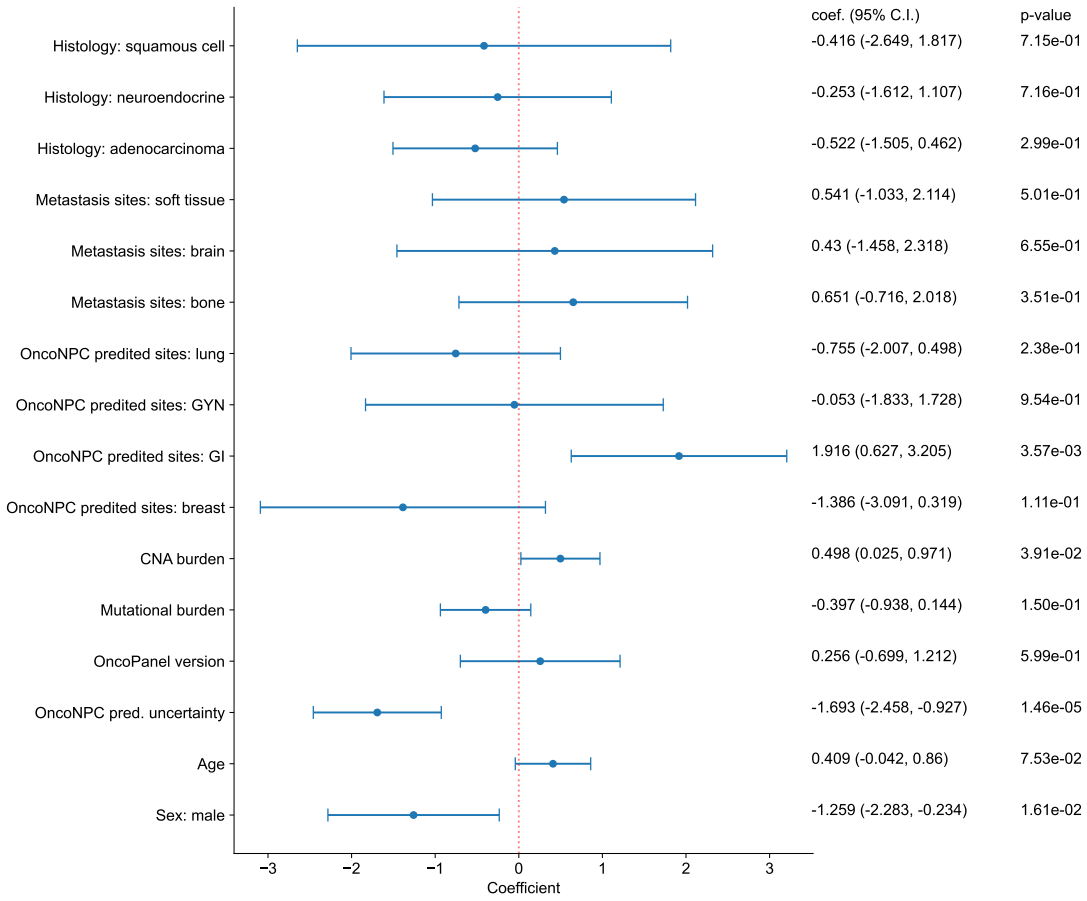



Figure S10: **Summary of coefficients for estimating treatment-OncoNPC concordance.** Formally, we estimated out-of-sample $P(A|X)$, where $A$ corresponds to the treatment-OncoNPC concordance, using a logistic regression model in a 10-fold cross-fitting. The coefficients were obtained from the first fold. See Methods.

Table S1: Center-specific number of held-out CKP tumor samples, broken down by cancer types and prediction confidence (i.e., $p_{\max}$) thresholds.

| | | Minimum $p_{\max}$ threshold | | | |
|---|---|---|---|---|---|
| | | 0.0 | 0.5 | 0.7 | 0.9 |
| Overall | DFCI | 3690 | 3438 | 3047 | 2502 |
| | MSK | 3331 | 3012 | 2608 | 2112 |
| | VICC | 268 | 230 | 192 | 136 |
| Non-Small Cell Lung Cancer (NSCLC) | DFCI | 811 | 735 | 644 | 533 |
| | MSK | 717 | 618 | 520 | 430 |
| | VICC | 36 | 27 | 23 | 19 |
| Invasive Breast Carcinoma (BRCA) | DFCI | 600 | 572 | 514 | 433 |
| | MSK | 727 | 675 | 598 | 474 |
| | VICC | 68 | 62 | 48 | 35 |
| Colorectal Adenocarcinoma (COADREAD) | DFCI | 521 | 502 | 479 | 436 |
| | MSK | 375 | 358 | 330 | 303 |
| | VICC | 55 | 52 | 48 | 37 |
| Diffuse Glioma (DIFG) | DFCI | 400 | 390 | 383 | 361 |
| | MSK | 214 | 204 | 187 | 168 |
| | VICC | 11 | 10 | 8 | 4 |
| Prostate Adenocarcinoma (PRAD) | DFCI | 126 | 118 | 98 | 67 |
| | MSK | 300 | 280 | 233 | 163 |
| | VICC | 16 | 10 | 6 | 3 |
| Pancreatic Adenocarcinoma (PAAD) | DFCI | 136 | 125 | 104 | 71 |
| | MSK | 233 | 216 | 187 | 154 |
| | VICC | 10 | 8 | 6 | 1 |
| Ovarian Epithelial Tumor (OVT) | DFCI | 257 | 229 | 184 | 112 |
| | MSK | 100 | 60 | 38 | 10 |
| | VICC | 12 | 9 | 5 | 2 |
| Esophagogastric Adenocarcinoma (EGC) | DFCI | 171 | 153 | 114 | 66 |
| | MSK | 82 | 70 | 44 | 24 |
| | VICC | 11 | 8 | 7 | 2 |
| Endometrial Carcinoma (UCEC) | DFCI | 123 | 116 | 95 | 73 |
| | MSK | 105 | 100 | 91 | 70 |
| | VICC | 7 | 6 | 6 | 2 |
| Melanoma (MEL) | DFCI | 134 | 127 | 115 | 103 |
| | MSK | 108 | 103 | 98 | 92 |
| | VICC | 24 | 24 | 23 | 23 |
| Bladder Urothelial Carcinoma (BLCA) | DFCI | 86 | 81 | 67 | 52 |
| | MSK | 93 | 84 | 78 | 65 |
| | VICC | 4 | 4 | 4 | 3 |

| | | Minimum $p_{\max}$ threshold | | | |
|---|---|---|---|---|---|
| | | 0.0 | 0.5 | 0.7 | 0.9 |
| Renal Cell Carcinoma (RCC) | DFCI | 79 | 71 | 61 | 50 |
| | MSK | 85 | 75 | 68 | 56 |
| | VICC | 6 | 5 | 4 | 3 |
| Head and Neck Squamous Cell Carcinoma (HNSCC) | DFCI | 55 | 50 | 39 | 28 |
| | MSK | 27 | 18 | 12 | 5 |
| | VICC | . | . | . | . |
| Cholangiocarcinoma (CHOL) | DFCI | 18 | 12 | 10 | 7 |
| | MSK | 40 | 31 | 24 | 16 |
| | VICC | 1 | . | . | . |
| Gastrointestinal Stromal Tumor (GIST) | DFCI | 47 | 46 | 43 | 40 |
| | MSK | 34 | 33 | 31 | 30 |
| | VICC | . | . | . | . |
| Well-Differentiated Thyroid Cancer (WDTC) | DFCI | 17 | 15 | 14 | 9 |
| | MSK | 31 | 31 | 29 | 25 |
| | VICC | 1 | 1 | 1 | . |
| Pleural Mesothelioma (PLMESO) | DFCI | 24 | 21 | 14 | 10 |
| | MSK | 18 | 17 | 10 | 6 |
| | VICC | 5 | 3 | 2 | 1 |
| Meningothelial Tumor (MNGT) | DFCI | 27 | 25 | 23 | 20 |
| | MSK | 3 | 3 | 1 | . |
| | VICC | 1 | 1 | 1 | 1 |
| Gastrointestinal Neuroendocrine Tumors (GINET) | DFCI | 20 | 17 | 16 | 11 |
| | MSK | 3 | 3 | 2 | . |
| | VICC | . | . | . | . |
| Pancreatic Neuroendocrine Tumor (PANET) | DFCI | 15 | 14 | 13 | 8 |
| | MSK | 24 | 22 | 19 | 15 |
| | VICC | . | . | . | . |
| Acute Myeloid Leukemia (AML) | DFCI | 15 | 11 | 10 | 6 |
| | MSK | . | . | . | . |
| | VICC | . | . | . | . |
| Non-Hodgkin Lymphoma (NHL) | DFCI | 8 | 8 | 7 | 6 |
| | MSK | 12 | 11 | 8 | 6 |
| | VICC | . | . | . | . |

# Supplementary Notes

## Identifying prognostic somatic variants shared in CUP-metastatic CKP pairs

To identify prognostic somatic variants shared between CUP/metastatic-CKP pairs, we again restricted to the 7 common OncoNPC subtypes with at least 35 CUP patients: NSCLC, PAAD, BRCA, COADREAD, HNSCC, EGC, GINET, and OVT. For somatic variants, we utilized the same processed features utilized in the OncoNPC model training (see Methods: Feature selection and OncoNPC model interpretation). To ensure sufficient statistical power, we restricted to candidate somatic variants (i.e., mutated genes and CNA genes) present in at least 15 samples in a given OncoNPC subtype and corresponding metastatic CKP cohort, as well as all 96 mutational signatures.

After selecting the cancer types to consider in the CUP-metastatic CKP pairs and candidate somatic variants for each pair, we iteratively tested each feature for association with survival in each OncoNPC subtype and in each corresponding metastatic CKP cohort. A multivariable Cox Proportional Hazard regression [32] model was used with time-to-death from sequencing as the outcome. To adjust for baseline effects, we included age at sequencing, sex, tumor sequencing panel version, mutational burden (i.e., sum of total somatic mutations in each tumor sample), and CNA burden (i.e., sum of total CNA events in each tumor sample) as covariates. Finally, to identify shared prognostic somatic variants for each CUP-metastatic CKP pair, we retained somatic variants which passed Schoenfield residuals-based proportional hazard tests (Python `lifelines` v0.27.4 [65]: p-value threshold: 0.05) and were nominally significant ($p < 0.05$) for both CUP and CKP cancer types in each pair.

Three out of 14 tested CUP-metastatic CKP pairs (NSCLC, PAAD, and COADREAD) exhibited shared prognostic somatic variants significantly associated with overall survival with nominal p-value cut-off at 0.05 (Fig. S8a and S8b). In patients with known or classified NSCLC, three somatic mutations were associated with poor survival in both groups: SMARCA4 (CUP: H.R. 1.86, 95% C.I. 1.19 - 2.89, p-value $6.23 \times 10^{-3}$, CKP mets: H.R. 1.73, 95% C.I. 1.44 - 2.09, p-value $9.30 \times 10^{-9}$), STK11 (CUP: H.R. 1.76, 95% C.I. 1.14 - 2.71, p-value $1.05 \times 10^{-2}$, CKP mets: H.R. 1.43, 95% C.I. 1.22 - 1.68, p-value $1.00 \times 10^{-5}$), and KEAP1 (CUP: H.R. 1.83, 95% C.I. 1.18 - 2.85, p-value $6.82 \times 10^{-3}$, CKP mets: H.R. 1.40, 95% C.I. 1.18 - 1.66, p-value $1.27 \times 10^{-4}$). These associations of somatic mutations in SMARCA4, STK11, and KEAP1 genes with overall survival are well established for NSCLC [66–68]. Interestingly, a CNA event in NKX2-1 was associated with improved survival in the patients from the NSCLC pair (CUP: H.R. 0.542, 95% C.I. 0.326 - 0.901, p-value $1.83 \times 10^{-2}$, CKP mets: H.R. 0.770, 95% C.I. 0.662 - 0.894, p-value $6.28 \times 10^{-4}$), consistent with prior meta-analyses [69]. In patients with known or classified COADREAD tumors, SBS10b mutation signature, linked to polymerase epsilon exonuclease domain mutations [20], was associated with longer overall survival (CUP: H.R. 0.371, 95% C.I. 0.148 - 0.928, p-value $3.41 \times 10^{-2}$, CKP mets: H.R. 0.495, 95% C.I. 0.255 - 0.958, p-value $3.68 \times 10^{-2}$). Finally, in patients with known or classified PAAD tumors, the SBS29 mutation signature (commonly found in tumor samples from individuals with a tobacco chewing habit [20]) was associated with poor survival in CUPs but nominally protective in metastatic

CKPs (CUP: H.R. 2.66, 95% C.I. 1.02 - 6.93, p-value $4.46 \times 10^{-2}$, CKP mets: H.R. 0.657, 95% C.I. 0.438 - 0.986, p-value $4.28 \times 10^{-2}$). Although these somatic associations remain to be validated in independent cohorts, by categorizing patients with CUP based on their OncoNPC predictions, we were able to identify prognostic somatic variants, consistent with recent research findings.

## Determining treatment-OncoNPC concordance

Concordance of OncoNPC predicted cancer type with a first palliative treatment assignments at DFCI was classified in one of five categories: 1) "TRUE": the OncoNPC cancer type matched the clinically proven/suspected tumor type and the predicted treatment matched the treatment received, which was dictated by NCCN guidelines and/or standard of care, within the clinical context provided by the medical record; 2) "FALSE": the OncoNPC cancer type did not match the clinically proven/suspected cancer type and the predicted treatment was not appropriate per NCCN guidelines or standard of care, in most reasonable situations, and within the context of the medical record; 3) "SOFT FALSE": the OncoNPC cancer type did not match the clinically proven/suspected cancer type, but the treatment received was not chosen based on NCCN guidelines or standard of care, owing to the unique clinical context provided by the medical record, 4) "EMPIRIC": treatment received was empiric treatment for cancer of unknown primary (e.g., carboplatin/taxol or gemcitabine/cisplatin) with the corresponding clinical rationale; in cases where patients received these regimens but not with the clinical intent of empiric CUP treatment (i.e., as regimens intended for treating other tumor types), the predicted treatment was not labeled as "EMPIRIC" and the case was instead evaluated in context of the proven/suspected tumor type. In our analysis, we considered the TRUE group as the concordant group, and FALSE and SOFT FALSE groups as the discordant group. We did not include the EMPIRIC group, which is typically a more challenging patient population with systematically worse outcomes [40].