nature medicine

Article

https://doi.org/10.1038/s41591-023-02482-6

Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary

Received: 6 January 2023

Accepted: 30 June 2023

Published online: 7 August 2023

Check for updates

Intae Moon[1,2], Jaclyn LoPiccolo[3], Sylvan C. Baca[3,4], Lynette M. Sholl[5], Kenneth L. Kehl[2], Michael J. Hassett[2], David Liu[2,3,6], Deborah Schrag[7] & Alexander Gusev[2,6,8]

Cancer of unknown primary (CUP) is a type of cancer that cannot be traced back to its primary site and accounts for 3–5% of all cancers. Established targeted therapies are lacking for CUP, leading to generally poor outcomes. We developed OncoNPC, a machine-learning classifier trained on targeted next-generation sequencing (NGS) data from 36,445 tumors across 22 cancer types from three institutions. Oncology NGS-based primary cancer-type classifier (OncoNPC) achieved a weighted F1 score of 0.942 for high confidence predictions (≥ 0.9) on held-out tumor samples, which made up 65.2% of all the held-out samples. When applied to 971 CUP tumors collected at the Dana-Farber Cancer Institute, OncoNPC predicted primary cancer types with high confidence in 41.2% of the tumors. OncoNPC also identified CUP subgroups with significantly higher polygenic germline risk for the predicted cancer types and with significantly different survival outcomes. Notably, patients with CUP who received first palliative intent treatments concordant with their OncoNPC-predicted cancers had significantly better outcomes (hazard ratio (HR) = 0.348; 95% confidence interval (CI) = 0.210–0.570; $P = 2.32 \times 10^{-5}$). Furthermore, OncoNPC enabled a 2.2-fold increase in patients with CUP who could have received genomically guided therapies. OncoNPC thus provides evidence of distinct CUP subgroups and offers the potential for clinical decision support for managing patients with CUP.

When a standardized diagnostic workup, including radiology and pathology assessments, fails to locate the primary site of a metastatic cancer, it is diagnosed as cancer of unknown primary (CUP). CUP represents about 3–5% of all cancers worldwide[1] and is characterized by aggressive progression and poor prognosis (survival of 6–16 months[2]). The hidden nature of the primary sites limits treatment options as clinical responses to some treatments are known to vary based on patients' tumor types (for example, identical BRAF V600 mutations targetable in

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA. [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. [5]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]The Broad Institute of MIT & Harvard, Cambridge, MA, USA. [7]Memorial Sloan Kettering Cancer Center, New York City, NY, USA. [8]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.
e-mail: alexander_gusev@dfci.harvard.edu

melanoma but not in colorectal cancer[3]). Emerging cancer treatments targeting actionable molecular alterations are typically developed for specific cancer types (for example, HER2 in breast cancer and EGFR mutation or ALK/ROS1 rearrangement in non-small-cell lung cancer[4]) and are thus inaccessible to patients with CUP. Accurately identifying the latent primary site for CUP tumors and demonstrating clinical benefit from site-specific therapies may thus open many existing treatment options for patients with CUP.

Pathology assessment has a key role in determining primary cancer types of malignant tumors based on immunohistochemistry (IHC) results as well as tumor morphology and clinical findings; however, pathological diagnosis can be challenging for highly metastatic or poorly differentiated tumors. For known cancer types, previous studies showed that an IHC-based diagnostic workup correctly identified 77–86% of primary tumors, which further decreased to 60–71% for metastatic tumors[5]. For patients with CUP, IHC results suggestive of a single primary diagnosis account for only 25% of tumors[2]. The subjective nature of pathological interpretation and guidelines, as well as the variability in IHC staining techniques across institutions, thus makes it challenging to establish consistent protocols for CUP diagnosis[6].

Molecular tumor profiling has been proposed as an alternative for primary site classification, potentially for CUP tumors, due to its quantitative nature and high accuracy on tumors with known cancer types[7–12]. Such tools rely on microarray DNA methylation[7], whole-genome sequencing[8,11], RNA-sequencing data[10] or gene expression profiling[12,13]. However, despite their effectiveness, these sequencing techniques have not been integrated into the standard of care and are often cost-prohibitive. In a recent study in ref. 9, it was demonstrated that accurate primary cancer-type classifications could be made from next-generation sequencing (NGS) of targeted panels which are now routinely collected at many cancer centers and are applicable to hundreds of thousands of tumors[14]. However, its clinical utility in diagnosing and aiding treatment for patients with CUP was not systematically investigated.

Several recent studies have investigated the potential clinical benefit of molecular CUP classification, in nonrandomized prospective studies[15–17] and randomized clinical trials[18]. These trials have often struggled to recruit a sufficient number of representative patients and explore the full range of available therapies. A recent randomized phase II trial[18] did not find significant improvement in 1-year survival for the treatment group receiving site-specific therapy guided by molecular profiling. However, this study was limited by a small number of patients ($n = 101$) recruited over 7 years, with few common solid tumor types and well-established therapies[19]. Assessing the clinical benefits of molecular CUP classification thus poses both an opportunity for precision medicine and a major challenge for conventional randomized studies.

Retrospective electronic health record (EHR) data, despite potential biases, can capture a larger and more heterogeneous patient population compared to prospective trials. When paired with tumor sequencing, these data can offer insights into the molecular workings of CUP tumors and how they relate to patient outcomes. As panel sequencing is often part of the standard of care, such insights also have the potential to assist diagnostic efforts and clinical management within existing molecular workflows. Here we used multicenter, NGS-targeted panel sequencing data from 36,445 tumor samples with known primary cancers to train and evaluate a machine-learning classifier predicting a primary cancer type of a given tumor sample. We applied this classifier, named OncoNPC, to 971 patients with CUP with clinical follow-up at Dana-Farber Cancer Institute (DFCI). Using the OncoNPC cancer-type predictions, we identified CUP subgroups that shared specific characteristics with their corresponding predicted primaries including significant differences in clinical outcomes and elevated germline risk. Furthermore, we showed that site-specific treatments concordant with the OncoNPC cancer-type predictions led to longer survival than those discordant with the cancer-type predictions. Finally, OncoNPC predictions yielded a 2.2-fold increase in the number of patients with CUP who could have received genomically guided therapies. Our findings suggest that many CUP tumors can be classified into meaningful subgroups with the potential to aid clinical decision-making.

Results

OncoNPC accurately classifies 22 known cancer types

We developed OncoNPC, a molecular cancer-type classifier trained on multicenter targeted panel sequencing data (Fig. 1). OncoNPC used somatic alterations, including mutations (single nucleotide variants (SNVs) and indels), mutational signatures, copy number alterations (CNAs) and patient age at the time of sequencing, and sex to jointly predict the cancer type using the XGBoost algorithm[20] (refer to Methods and Supplementary Note 1 for more details on choosing input features). OncoNPC was trained and validated on the processed data consisting of 29,176 primary and metastasis tumor samples from 22 known cancer types collected at the DFCI, Memorial Sloan Kettering (MSK) Cancer Center and Vanderbilt-Ingram Cancer Center (VICC; refer to Table 1 for details regarding patient demographics, modeled cancer types and their corresponding abbreviations). Across all 22 cancer types, OncoNPC achieved a weighted F1 score of 0.784 on the held-out test tumor samples consisting of 7,289 tumor samples (weighted precision and recall, 0.789 and 0.791, respectively). Across 13 cancer groups (grouped by sites and treatment options; Table 1), OncoNPC achieved an overall weighted F1 score of 0.806 (weighted precision and recall, 0.810 and 0.809, respectively). Despite the evident class imbalance across cancer types, OncoNPC showed well-balanced precision across the cancer types (Fig. 2a) and cancer groups (Fig. 2b; refer to Extended Data Fig. 1 for more performance details).

We evaluated the performance of OncoNPC at the following four distinct prediction confidence levels based on p_{max} (that is, the maximum predictive probability across 22 cancer types): 0.0 (encompassing all samples), 0.5, 0.7 and 0.9 (refer to Supplementary Note 2 and Supplementary Fig. 1 for an alternative approach using a cancer-type-specific threshold). Applying a threshold based on p_{max} resulted in further performance improvement—weighted F1 score of 0.830 with 91.6% remaining samples at $p_{max} \geq 0.5$ and 0.942 with 65.2% remaining samples at $p_{max} \geq 0.9$ (Fig. 2c,d). While rare cancer types had generally lower overall performance, increasing the p_{max} threshold reduced this difference between common/rare cancer types. At $p_{max} \geq 0$, common cancer types in the upper quartile in terms of the number of tumor samples (NSCLC, BRCA, COADREAD, DIFG, PRAD and PAAD) had a mean F1 score of 0.841, while rare cancer types in the lower quartile (WDTC, MNGT, GINET, PANET, AML and NHL) had a mean F1 score of 0.581, whereas at $p_{max} \geq 0.9$, common and rare cancer had mean F1 scores of 0.953 and 0.860, respectively. Furthermore, OncoNPC demonstrated robust performance against potential real-world dataset shifts due to factors including cancer center, biopsy site type, sequence panel version and patient ethnicity (Fig. 2e and Extended Data Fig. 2a; refer to Supplementary Note 3 and Supplementary Figs. 2 and 3 for more details on OncoNPC's performance regarding real-world dataset shifts and difficult-to-predict cancer types such as CHOL and HNSCC). Finally, a feature ablation study demonstrated that OncoNPC continues to achieve high performance with only the top 50% of genomic features retained (overall weighted F1 score of 0.757 versus 0.777 at p_{max} threshold of 0 and 0.950 versus 0.960 at p_{max} threshold of 0.9; Supplementary Note 4, Supplementary Table 1 and Extended Data Fig. 3).

Applying OncoNPC to CUP tumor samples

We applied OncoNPC to classify 971 CUP tumors from patients who were admitted to DFCI and sequenced as part of routine clinical care. OncoNPC classifications for CUPs had prediction probabilities lower than those of 3,690 held-out cancer with known primary (CKP) tumors at DFCI on average (0.764 versus 0.881), but comparable to those of 8,025 CKPs at DFCI, including tumors with cancer types not modeled

https://doi.org/10.1038/s41591-023-02482-6

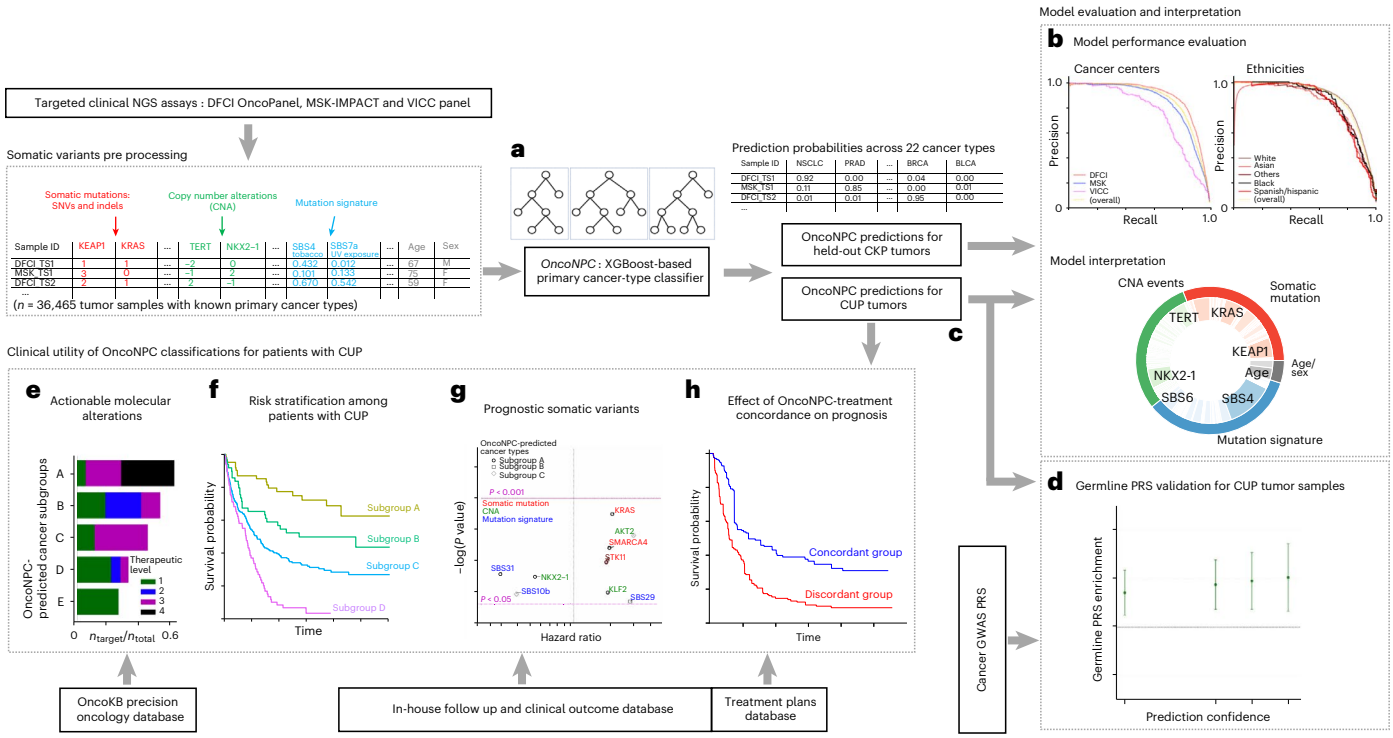


Fig. 1 | Overview of model development and analysis workflow. a, OncoNPC, an XGBoost-based classifier, was trained and evaluated using 36,465 cancer with known primary (CKP) tumor samples across 22 cancer types collected from three different cancer centers. **b**, OncoNPC performance was evaluated on the held-out tumor samples (n = 7,289). **c**, OncoNPC was applied to 971 CUP tumor samples at a single institution to predict primary cancer types. **d**–**g**, OncoNPC-predicted CUP subgroups were then investigated for association with elevated germline risk (**d**), actionable molecular alterations (**e**), overall survival (**f**) and prognostic somatic features (**g**). **h**, A subset of CUP patients with detailed treatment data was evaluated for treatment-specific outcomes.

in OncoNPC (0.769). This indicates that CUP tumors may contain other rare cancer types (Supplementary Note 5 and Extended Data Fig. 2b). Nevertheless, 41.2% of the CUP tumors (400 of 971) could still be classified with high confidence (that is, $p_{max} \geq 0.9$), and multiple classified cancer types including NSCLC, BRCA, PAAD and PRAD had distributions of prediction probabilities comparable to their corresponding CKPs (Fig. 3a). Interestingly, CUPs with predicted GINET were highly confident, despite their small number of tumor samples in the training cohort (n = 359; 0.99% of the training cohort), suggesting that some rare cancer types may nevertheless be confidently identifiable. As shown in Fig. 3b, the most common CUP cancer types were NSCLC, PAAD, BRCA, EGC and COADREAD. NSCLC, BRCA and COADREAD were also the top three most common CKP types. These rates are broadly consistent with previous findings that the most frequently revealed underlying primary cancers for CUPs by autopsy include lung, large bowel and pancreas cancers[21]. Finally, comparable rates were observed upon applying OncoNPC to 581 CUP tumors at MSK Cancer Center (Supplementary Fig. 4).

Explaining OncoNPC cancer-type predictions

OncoNPC learns complex nonlinear relationships between input somatic variants and clinical features and provides interpretable primary cancer-type predictions, where the impact of each input feature on a prediction is quantified as a SHAP value[22]. We investigated the most impactful features in predicting each cancer type across the CKP and CUP cohorts to evaluate the face validity of OncoNPC (refer to Fig. 3d for the top three most frequently predicted cancer types in the CUP cohort as follows: NSCLC, BRCA and PAAD, and Supplementary Figs. 5 and 6 for other cancer types). For NSCLC, the most important features were EGFR mutation and SBS4, a tobacco smoking-associated mutation signature[23], for both CKP tumor samples and CUP with NSCLC-predicted tumor samples, consistent with the known etiology of lung cancer. Somatic mutation in the *EGFR* gene is frequently observed in NSCLC tumors, and the gene itself is a well-known therapeutic target for patients with NSCLC[24,25]. Carcinogens in tobacco smoke have been known to cause lung cancer[26]. For BRCA, the most important feature for both CKP and CUP tumor samples was sex, as expected, followed by somatic mutation in *PIK3CA* and CNA event in the *CCND1* gene, known drivers and prognostic indicators in breast cancer[27,28]. For PAAD, KRAS mutation was significantly more common than the population averages and by far the most important somatic feature. Mutations in the *KRAS* gene occur frequently among patients with pancreatic cancer and are known to have prognostic significance[29,30]. OncoNPC provides intuitive visualizations to explain individual-level predictions. As an example, we show how OncoNPC explained the classification of a tumor sample from a 76-year-old male patient with CUP (Extended Data Fig. 4). The feature interpretation analysis showed that OncoNPC was able to capture cancer-specific signals in somatic mutations and clinical features, both at the individual and cohort level.

Germline polygenic risk score (PRS)-based validation on CUP tumor samples

We hypothesized that, if OncoNPC was accurately identifying latent primary cancers, the classified CUP cancer types would exhibit increased germline risk for the corresponding cancers. To that end, we imputed common germline variation for each patient with CUP and quantified their PRSs across eight common cancers using external cancer genome-wide association study (GWAS) data (Methods, Supplementary Note 6 and Supplementary Fig. 7). PRSs are a continuous estimate of the underlying germline liability for a given cancer and orthogonal from the somatic data used to train OncoNPC. As hypothesized, patients with CUP had a significantly higher mean germline PRS

https://doi.org/10.1038/s41591-023-02482-6

Table 1 | Demographic information of the patients and tumor samples across DFCI, MSK and VICC

Demographics		DFCI	MSK	VICC	CUP at DFCI
Number of patients		18,106	15,151	1,310	962
Patients' age at sequence (95% CI)		60.7 (60.5–60.9)	60.2 (60.0–60.4)	58.3 (57.6–59.0)	61.9 (61.1–62.7)
Sex, male-to-female ratio		43.8–56.2	43.5–56.5	44.5–55.5	50.0
Patients' ethnicity (proportion %)					
White		16,105 (88.9%)	11,575 (76.4%)	1,089 (83.1%)	853 (88.7%)
Black		538 (3.0%)	866 (5.7%)	72 (5.5%)	38 (4.0%)
Asian		554 (3.1%)	956 (6.3%)	17 (1.3%)	34 (3.5%)
Hispanic		379 (2.1%)	744 (4.9%)	14 (1.1%)	15 (1.6%)
Others		530 (2.9%)	1010 (6.7%)	118 (9.0%)	22 (2.2%)
Sequenced tumor samples					
Total number of samples		18,816	16,294	1,355	971
Panel version (proportion %; 95% sequence date range)					
v1		OncoPanel v1	MSK-IMPACT341	VICC-01-T5A	OncoPanel v1
		1,924 (10.2%; 2013-8-20–2014-8-17)	1,803 (11.1%; not available)	307 (23.0%; not available)	47 (4.8%; 2013-9-8–2014-8-12)
v2		OncoPanel v2	MSK-IMPACT410	VICC-01-T7	OncoPanel v2
		5,304 (28.2%; 2014-9-28– 2016-10-5)	6,917 (42.5%; not available)	1,028 (77.0%; not available)	203 (20.9%; 2014-11-5– 2016-10-5)
v3		OncoPanel v3	MSK-IMPACT468		OncoPanel v3
		11,588 (61.6%; 2016-11-11– 2021-1-6)	7,574 (46.5%; not available)		701 (74.3%; 2016-12-14– 2020-12-23)
Biopsy site type					
Primary		11,662 (62.0%)	9,576 (58.8%)	622 (46.6%)	
Metastatic recurrence		5,737 (30.5%)	6,718 (41.2%)	637 (47.7%)	
Local recurrence		673 (3.6%)	Not available	64 (4.8%)	
Unspecified/others		744 (4.0%)	Not available	12 (0.9%)	
Cancer group	Cancer type	–	–	–	Predicted cancer type
Lung (thoracic)	Non-small-cell lung cancer (NSCLC)	3,489 (18.5%)	3,183 (19.5%)	137 (10.3%)	280 (28.8%)
	Pleural mesothelioma (PLMESO)	258 (1.4%)	118 (0.7%)	2 (0.1%)	9 (0.9%)
	Invasive breast carcinoma (BRCA)	2,558 (13.6%)	3,113 (19.1%)	274 (20.5%)	85 (8.8%)
	Colorectal adenocarcinoma (COADREAD)	2,525 (13.4%)	1,919 (11.8%)	232 (17.4%)	63 (6.5%)
Upper gastrointestinal	Esophagogastric adenocarcinoma (EGC)	988 (5.3%)	495 (3.0%)	59 (4.4%)	69 (7.1%)
	Pancreatic adenocarcinoma (PAAD)	772 (4.1%)	980 (6.0%)	53 (4.0%)	85 (8.8%)
	Cholangiocarcinoma (CHOL)	241 (1.3%)	338 (2.1%)	44 (3.3%)	33 (3.4%)
Neuro	Diffuse glioma (DIFG)	2,041 (10.8%)	1,069 (6.6%)	47 (3.5%)	25 (2.6%)
	Meningothelial tumor (MNGT)	179 (1.0%)	42 (0.3%)	15 (1.1%)	4 (0.4%)
Gynecologic	Ovarian epithelial tumor (OVT)	1,213 (6.4%)	525 (3.2%)	81 (6.1%)	58 (6.0%)
	Endometrial carcinoma (UCEC)	703 (3.7%)	703 (4.3%)	34 (2.5%)	18 (1.9%)
Urothelial	Renal cell carcinoma (RCC)	457 (2.4%)	497 (3.1%)	39 (2.9%)	24 (2.5%)
	Bladder urothelial carcinoma (BLCA)	550 (2.9%)	505 (3.1%)	41 (3.1%)	21 (2.2%)
	Prostate adenocarcinoma (PRAD)	601 (3.2%)	1,222 (7.5%)	27 (2.0%)	27 (2.8%)
	Melanoma (MEL)	729 (3.9%)	619 (3.8%)	187 (14.0%)	43 (4.4%)
Head and neck	Head and neck squamous cell carcinoma (HNSCC)	473 (2.5%)	285 (1.7%)	20 (1.5%)	52 (5.4%)
	Well-differentiated thyroid cancer (WDTC)	166 (0.9%)	166 (1.0%)	8 (0.6%)	1 (0.1%)
Neuroendocrine	Gastrointestinal neuroendocrine tumors (GINET)	219 (1.2%)	76 (0.5%)	18 (1.3%)	46 (4.7%)
	Pancreatic neuroendocrine tumor (PANET)	121 (0.6%)	133 (0.8%)	12 (0.9%)	23 (2.4%)
	Gastrointestinal stromal tumor (GIST)	273 (1.5%)	217 (1.3%)	5 (0.4%)	3 (0.3%)
Hematologic	Acute myeloid leukemia (AML)	150 (0.8%)	1 (0.0%)	0 (0.0%)	1 (0.1%)
	Non-Hodgkin lymphoma (NHL)	110 (0.6%)	88 (0.5%)	0 (0.0%)	1 (0.1%)

https://doi.org/10.1038/s41591-023-02482-6

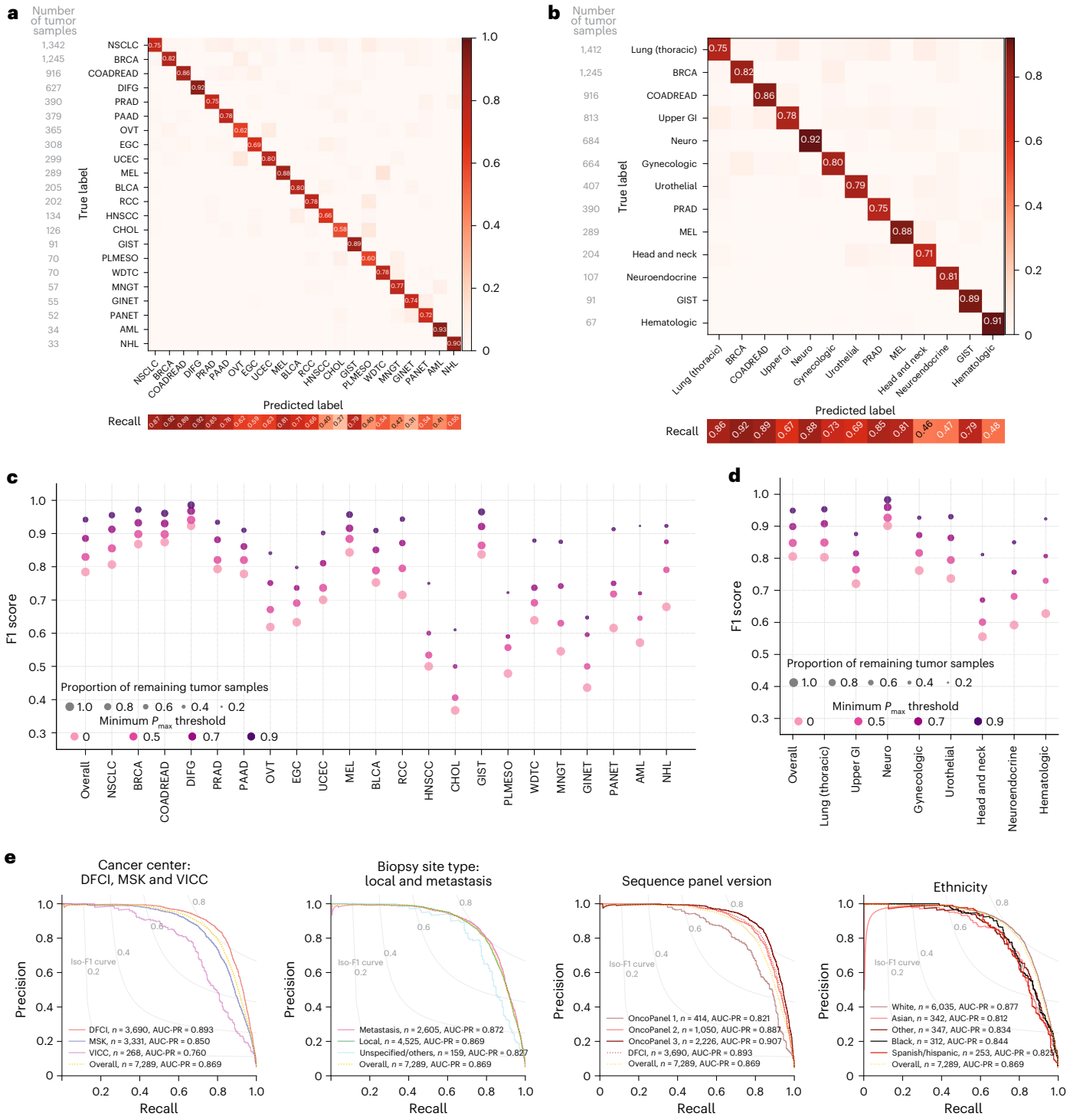


Fig. 2 | Cancer-type classification performance of OncoNPC. a,**b**, The normalized confusion matrix of OncoNPC classification performance on the held-out test set (n = 7,289) for 22 detailed cancer types (**a**) and 13 cancer groups (**b**; Table 1). Each confusion matrix displays precision for each cancer type or group on its diagonal. Below the matrix, the recall for each cancer type or group is shown, and the sample size is displayed to the left of the matrix for reference. **c**,**d**, The performance of OncoNPC in F1 score on the test set across cancer types (**c**) and groups (**d**) at four different p_{max} (that is, prediction confidence) thresholds. Each dot size is scaled by the proportion of tumor samples retained. **d**, Note that we only considered cancer groups that have more than one cancer type. Overall F1 scores were weighted according to the number of confirmed cases across cancer types and cancer groups, respectively. **e**, The precision–recall curves show OncoNPC's performance on the test set when grouped by cancer center, biopsy site type, sequence panel version and ethnicity. The yellow dotted curve represents the baseline performance across the entire test set.

for the OncoNPC-predicted cancers compared to the other cancer types (refer to Fig. 3c and Extended Data Fig. 5 for cancer-type-specific analysis). The magnitude of the difference (that is, $\hat{\Delta}_{PRS}$) increased for more confident OncoNPC predictions ($\hat{\Delta}_{PRS}$ = 0.142; 95% CI = 0.0494–0.235; two-sided Wald test, $P = 2.66 \times 10^{-3}$ at p_{max} threshold of 0.0 and $\hat{\Delta}_{PRS}$ = 0.204; 95% CI = 0.0655–0.344; two-sided Wald test,

$P = 3.98 \times 10^{-3}$ at p_{max} threshold of 0.9). As a negative control, the same analysis, conducted with randomly shuffled OncoNPC labels, showed no enrichment. As a positive control, the same analysis conducted on CKPs, with available imputed PRS ($n = 11{,}332$), also demonstrated a highly significant germline enrichment, as expected. Notably, the enrichment for CUP tumors was in between that of CKPs and tumors with randomly shuffled labels, suggesting that while OncoNPC-predicted CUP tumors are genetically correlated with their corresponding CKPs, they still exhibit additional heterogeneity.

OncoNPC-based risk stratification among patients with CUP

To demonstrate the clinical utility of OncoNPC, we examined if OncoNPC cancer-type predictions with moderately high confidence (≥ 0.5), a threshold consistently applied in subsequent clinical analyses, can stratify overall survival among patients with CUP. We identified subgroups that had significant prognostic differences in median survival based on the OncoNPC predictions (chi-squared test, $P = 4.90 \times 10^{-14}$; Fig. 4a). Overall, the poorest prognosis was observed in patients with CUP predicted to be EGC and PAAD—median survival 8.44 months for the combined cohort (95% CI = 5.39–10.5; $n = 107$). The most favorable prognosis was observed in patients with CUP predicted to be HNSCC, GINET and PANET: median survival 48.2 months for HNSCC (95% CI = 19.6 to not estimable; $n = 41$) and not estimable median survival (that is, the estimated survival curve never reached the median) for the combined GINET and PANET cohort ($n = 57$), respectively. Our identified favorable subgroups are consistent with established favorable CUP subtypes such as poorly or well-differentiated neuroendocrine carcinomas of unknown primary and squamous cell carcinoma of nonsupraclavicular cervical lymph nodes[31]. Furthermore, median survival times were significantly correlated across cancer types between CUP-metastatic CKP pairs (Spearman's $\rho = 0.964$, $P = 4.54 \times 10^{-4}$), as detailed in Fig. 4b and Supplementary Note 7. This suggests that genetics-based OncoNPC predictions capture prognostic signals specific to each predicted cancer type. Consequently, OncoNPC subgroups can be leveraged to meaningfully stratify the survival of patients with CUP. In an exploratory analysis, we also identified prognostic somatic variants common to both predicted CUP cancer groups and their corresponding metastatic CKP groups (Supplementary Note 8 and Supplementary Fig. 8).

Survival benefit from OncoNPC-concordant treatments

We performed a retrospective survival analysis to investigate whether patients with CUP achieved clinical benefit when treated in concordance with their OncoNPC predictions. We restricted to a cohort of 158 patients with CUP, who received the first treatment at DFCI with a palliative intent (see the exclusion criteria in Extended Data Fig. 6 and demographic details in Extended Data Table 1). Each case was then manually chart-reviewed by a certified oncologist to determine whether the treatment administered was concordant with the OncoNPC prediction as per National Comprehensive Cancer Network (NCCN) guidelines or standard of care (Supplementary Note 9). We used the following two estimation strategies to minimize potential bias and estimate the impact of treatment concordance on patient survival: multivariable Cox regression and inverse probability of treatment-weighted (IPTW) Kaplan–Meier estimator, which have recently been used to emulate estimates from randomized trials[32,33]. By applying these methods, we adjusted for baseline covariates including sex, age, OncoNPC prediction uncertainty, metastasis sites and pathological histology (Methods). Notably, patients with CUP who received first palliative treatments concordant with their OncoNPC-predicted cancer types exhibited significantly better survival than those who received discordant treatments, as shown in Fig. 5a,b (multivariable Cox regression: HR = 0.348, 95% CI = 0.210–0.570, $P = 2.32 \times 10^{-5}$; proportional hazard assumption test[34]: chi-squared test with 17 degrees of freedom, $P = 0.156$; IPTW Kaplan–Meier estimator: weighted log-rank test, $P = 1.97 \times 10^{-6}$). Furthermore, after stratifying by OncoNPC-predicted cancer groups and repeating the IPTW Kaplan–Meier analysis, we found that the treatment concordant group had improved survival across the cancer groups (breast, gastrointestinal (GI) and others), with the exception of the lung cancer group (Extended Data Fig. 7). The concordant treatment group achieved better survival outcomes after restricting to a subset of patients ($n = 33$) who received their initial treatments after the OncoPanel sequencing results were available for clinical assessment (weighted log-rank test, $P = 1.50 \times 10^{-8}$; Extended Data Fig. 8 and Supplementary Table 2) and a subset of patients ($n = 133$) without a prior history of known primary cancers (weighted log-rank test, $P = 2.87 \times 10^{-5}$; see Supplementary Note 10 and Supplementary Fig. 9). Finally, the multivariable Cox regression (Fig. 5a) and the IPTW Kaplan–Meier analysis identified significant hazardous and protective associations of several baseline covariates with survival and treatment concordance, respectively (Supplementary Note 10 and Supplementary Fig. 10).

Improving access to targeted treatments in patients with CUP

Based on a comprehensive review of the medical record for 158 patients with CUP by a certified oncologist, we identified 20 patients (12.7%) who received genomically guided treatments, split evenly between concordant and discordant groups. We used the OncoKB knowledge base[35] to link actionable variants with their respective targeted treatments (Methods). Notably, we found that 24 additional patients in the cohort (representing a 2.2-fold total increase, 13 in the treatment concordant group and 11 in the discordant group) could have been eligible for genomically guided treatments based on OncoNPC predictions. Specifically, actionable somatic variants, combined with the predicted cancer types, led to 28 eligible drugs under levels 1–3, where level 1 corresponds to FDA-approved drugs, level 2 corresponds to standard care and level 3 corresponds to biological evidence[35].

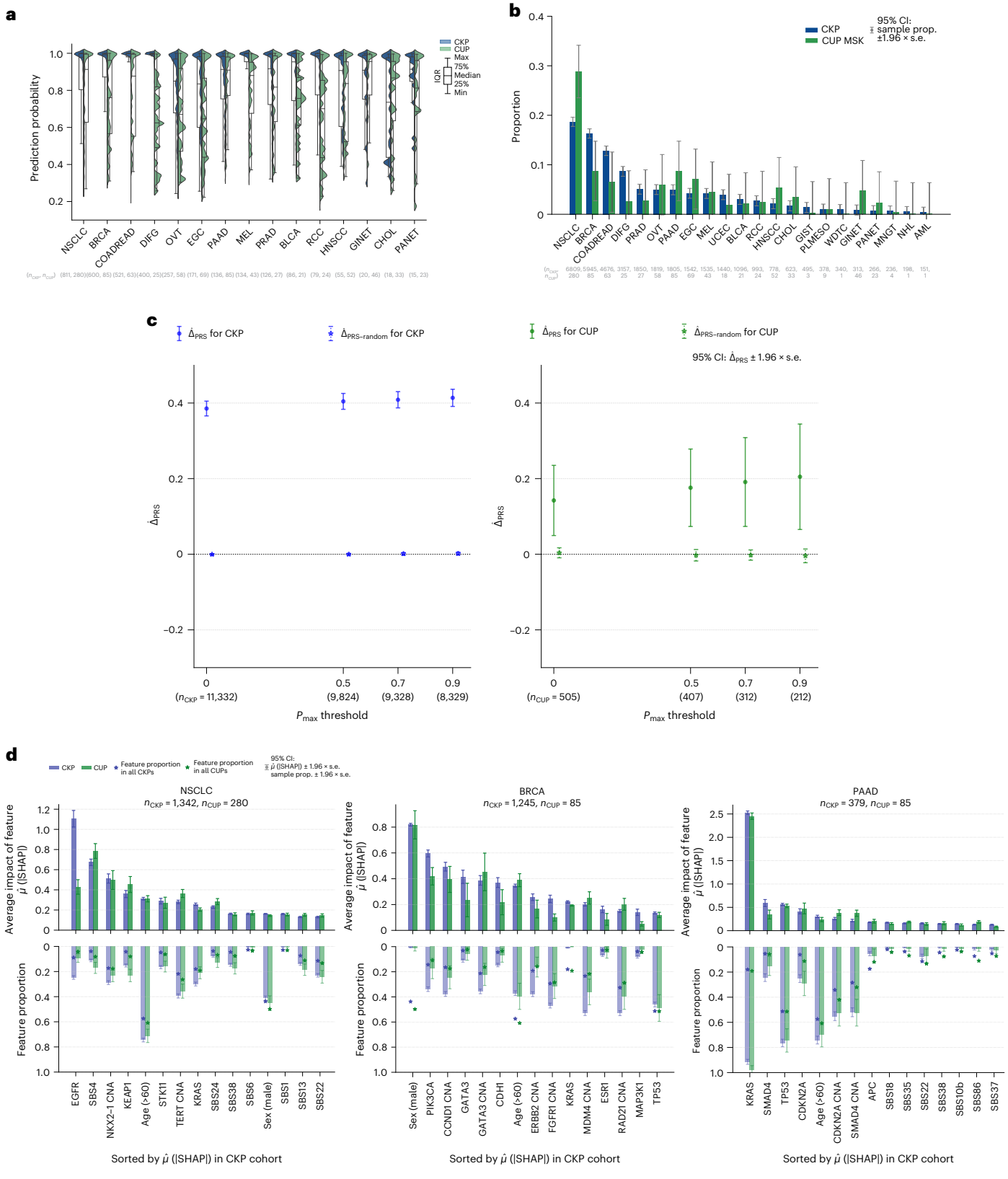
Fig. 3 | Application of OncoNPC to CUP tumors, germline PRS-based validation and interpretation of OncoNPC cancer-type predictions. **a**, Empirical distributions of prediction probabilities for correctly predicted, held-out CKP tumor samples ($n = 3{,}429$) and CUP tumor samples ($n = 934$) at DFCI across CKP cancer types (blue) and their corresponding OncoNPC-predicted cancer types for CUP tumors (green). Only OncoNPC-predicted cancer types 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 held-out CUP tumor samples ($n = 971$) are included. For both **a** and **b**, the cancer types (x axis) are ordered by the number of CKP tumor samples in each cancer type. **c**, Germline PRS enrichment of the CKP tumor samples ($n = 11{,}332$) and CUP tumor samples ($n = 505$) with available PRS data averaged across eight cancer types. The magnitude of the enrichment is quantified by $\hat{\Delta}_{PRS}$ as follows: the mean difference between the concordant (that is, OncoNPC matching) cancer-type PRS and mean of PRSs of discordant cancer types (Methods). $\hat{\Delta}_{PRS}$ is shown for CKPs in blue (for reference) and CUPs in green. As a negative control, $\hat{\Delta}_{PRS-random}$ is also shown after permuting the OncoNPC labels. **d**, Top 15 most important features based on mean absolute SHAP values (that is, $\hat{\mu}(|SHAP|)$) for the top three most frequently predicted cancer types in the CUP cohort as follows: NSCLC, BRCA and PAAD. The feature proportion (that is, carrier rate) for each feature in corresponding CKP and CUP cancer cohorts as well as the entire CKP and CUP cohorts are shown as bars going downward and star-shaped markers, respectively. For mutation signature features that have continuous values, individuals with feature values one s.d. above the mean were treated as positives and the rest as negative. For age, individuals above the population mean were treated as positives and the rest as negatives; 95% CIs were determined using the s.e. of the sample mean for $\hat{\mu}(|SHAP|)$ and the s.e. of the sample proportion for the carrier rate. These intervals are centered at the respective sample values.

https://doi.org/10.1038/s41591-023-02482-6

Figure 5c illustrates the OncoNPC-predicted cancer types, corresponding actionable variants and eligible drugs. Within a broader cohort of CUP tumors that were not chart-reviewed ($n = 794$), we similarly found that 22.8% had potentially actionable somatic variants per their respective OncoNPC cancer-type predictions (Supplementary Note 11 and Extended Data Fig. 9).

Discussion

We developed OncoNPC, a machine-learning model, for the molecular classification of tumor samples using multicenter NGS panel data. OncoNPC provided robust and interpretable predictions in held-out multicenter test data. Applied to CUP tumor samples, OncoNPC CUP subgroups showed significantly higher germline PRS risk for their



https://doi.org/10.1038/s41591-023-02482-6

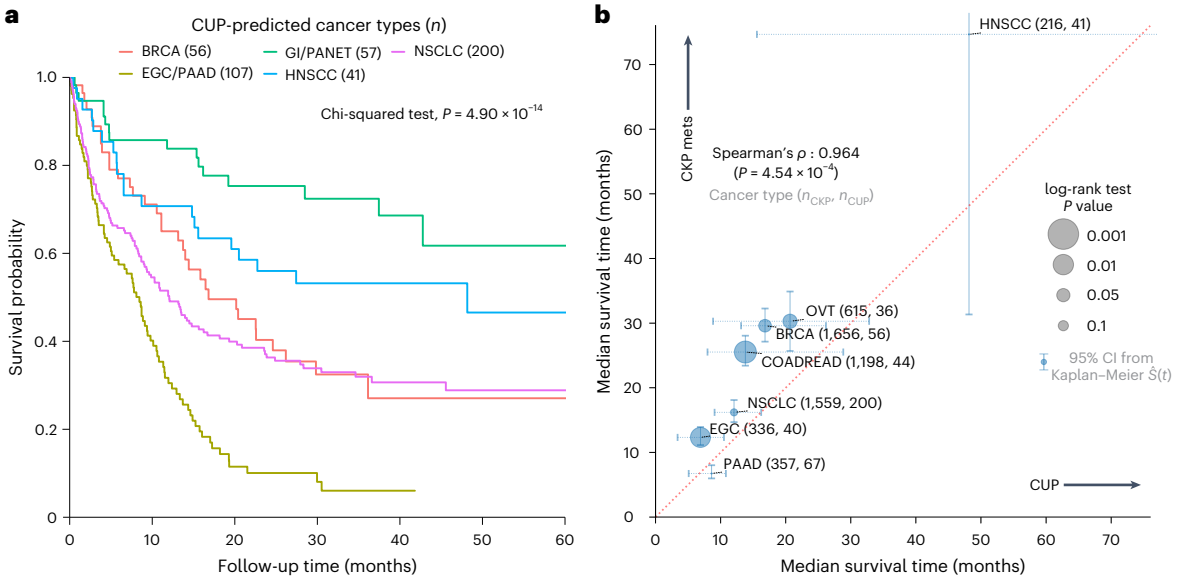


Fig. 4 | OncoNPC-based risk stratification among patients with CUP and median survival comparison between CUP and CKP metastatic cases. **a**, Survival stratification for patients with CUP based on their OncoNPC-predicted cancer types. The Kaplan–Meier estimator was used to estimate survival probability for each predicted cancer type over the follow-up time of 60 months from sequence date, with the statistical significance assessed by chi-square test. **b**, Median survival comparison between patients with CUP (across predicted cancer types in *x* axis) and patients with CKP metastatic cancer (across corresponding cancer types in *y* axis)—Spearman's $\rho = 0.964$ ($P = 4.54 \times 10^{-4}$). The size of each dot reflects the *P* value of the log-rank test for significant difference in median survival between each CUP-metastatic CKP pair. Only cancer types with at least 30 CUP tumor samples having OncoNPC prediction probabilities greater than 0.5 are shown. And, 95% CIs were obtained nonparametrically using Kaplan–Meier estimated survival function $\hat{S}(t)$.

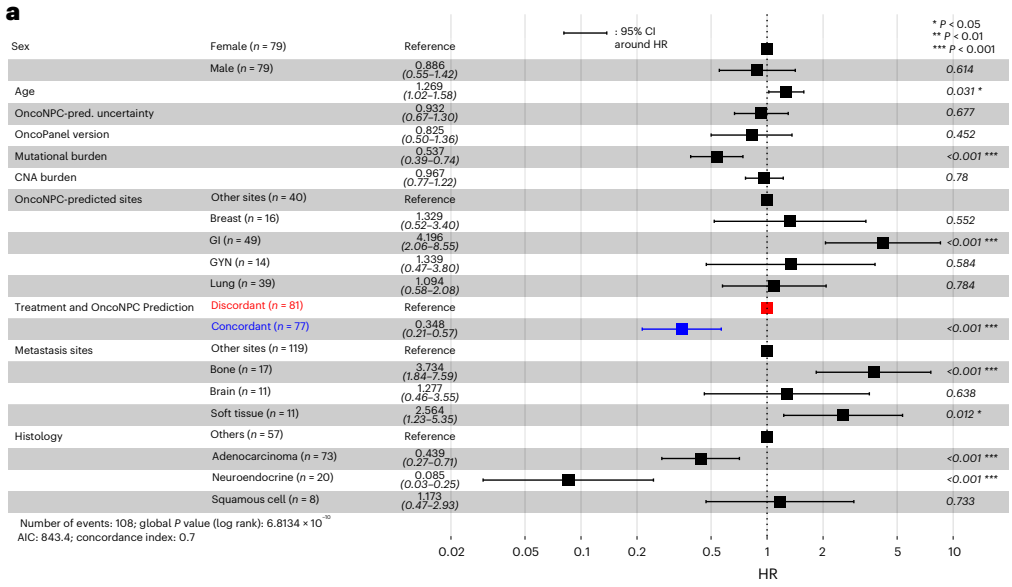
predicted cancers, the first evidence of germline genetic correlation between CUP tumors and corresponding CKP tumors to our knowledge. Furthermore, OncoNPC CUP subgroups showed significant survival differences, consistent with those observed in the corresponding CKP cancer types. In the retrospective survival analysis, patients with CUP treated in a consistent manner with their OncoNPC predictions achieved significantly longer survival than those treated in an inconsistent manner. Finally, OncoNPC predictions enabled a 2.2-fold increase in patients with CUP who could have received genomically guided therapies. Our findings suggest that CUP tumors share a genetic and prognostic architecture with known cancer types and may benefit from molecular classification.

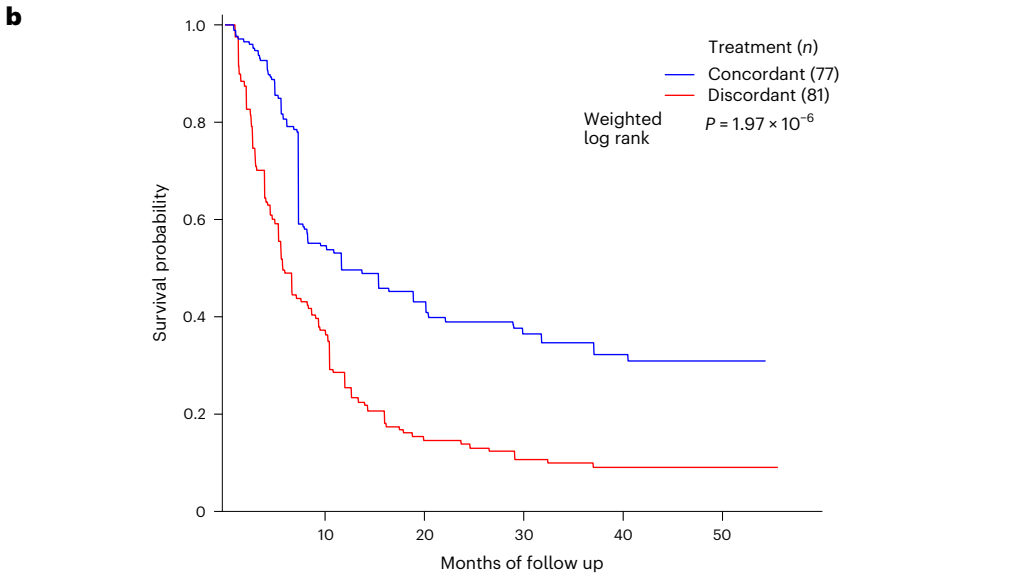
While previous studies have demonstrated accurate classification of known tumors using a variety of platforms[7–13,36,37], they typically applied algorithms to metastatic tumors of known types and did not investigate the clinical implications for CUP tumors at large scale. Notably, Moran et al[7]. observed a nominally significant difference in survival between patients with CUP who received site-specific treatments concordant with their molecular primary site predictions and those who received empiric treatments. However, this difference may be explained by systematically worse outcomes for the empirically treated group, which is typically a more challenging patient population[38]. To explicitly distinguish these scenarios, our analysis instead restricted to a CUP cohort wherein all patients received site-specific treatments as the first palliative intent therapy and estimated a significant survival benefit of concordant treatment versus discordant treatment (excluding the empirically treated group) to mimic clinical trials using real-world data[32,33]. Although we cannot rule out potential biases from unmeasured confounders, the proposed intervention (concordant treatment versus discordant treatment) is particularly challenging to ethically evaluate through RCTs, necessitating the use of retrospective causal inference.

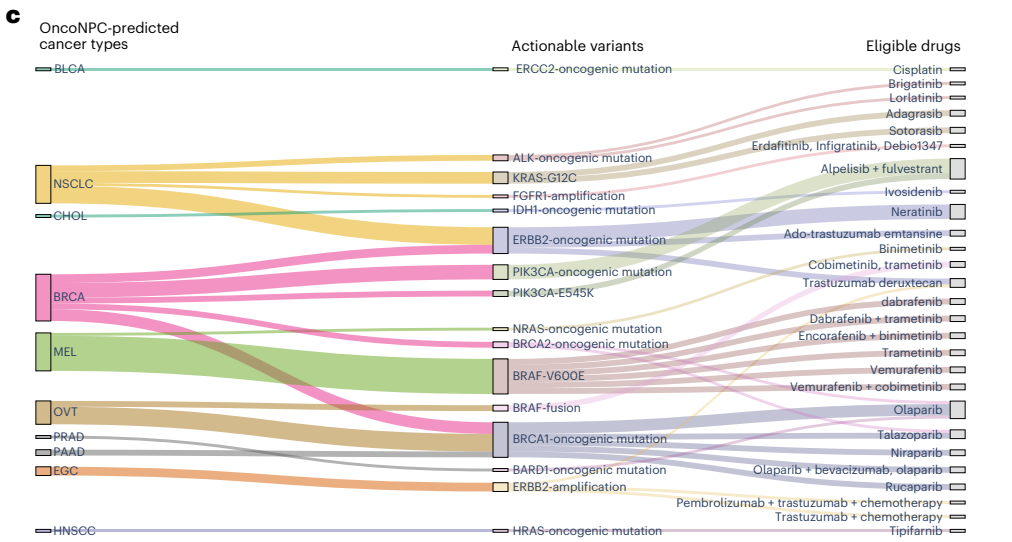
Our study has several limitations. Firstly, although we used multicenter data for training and evaluation of OncoNPC predictions, retrospective EHR data were only available from a single institution for downstream clinical analyses. Secondly, the majority of our cohort with panel sequencing data consists of white patients (83.2% in the training cohort), which may explain why OncoNPC performed better for the held-out tumors from white patients. Nevertheless, OncoNPC achieved an area under the precision–recall curve over 0.8 across all ethnicities. Thirdly, we considered only the 22 most common cancer types in the cohort as classification labels (68.1% of all tumor samples at DFCI and 69.9% across all three centers). As a result, if a CUP tumor sample harbors a distinct yet not modeled primary cancer type, then the tumor sample will likely have high uncertainty in the prediction, which we confirmed empirically (Supplementary Note 5). Nevertheless, previous work has shown that the majority of resolvable primary sites of CUP tumor samples were from common cancers (for example, lung, pancreas and GI)[21], consistent with our findings. Fourthly, our classifier and analyses relied on data from panel sequencing assays targeting 300–500 genes, which are inherently only sensitive to coding mutations and deep CNAs in the targeted genes. Other molecular features may thus improve classification performance (Supplementary Note 12). Our focus in this work was on assays that are in routine clinical use as those are linked to real-world clinical data and offer the most immediate translational potential. Notably, OncoNPC may still be effective with even more limited sequencing panels (Supplementary Note 4). Lastly, we stress that OncoNPC subgroups are still algorithmically defined and should not be considered true molecular subtypes without further molecular validation and independent replication.

Our findings suggest that routinely collected targeted tumor panel sequencing data have clinical utility in assisting diagnostic workup and prognosis and may additionally inform treatment decisions. Through our pathology-based evaluation, we discovered that 51.9% (67 of 129 cases) of CUP cases in the cohort had agreement between OncoNPC predictions and at least one pathology-based suspected primary (Supplementary Note 13). Despite being substantially higher than expected by chance (19.9%, 95% CI = 19.7–20.1%), this relatively low agreement

https://doi.org/10.1038/s41591-023-02482-6







underscores the challenge that highly metastatic or poorly differentiated tumors pose to pathological diagnosis[2,5]. In several cases, we found that OncoNPC predictions could have been helpful where multiple primaries were pathologically suspected (Supplementary Note 13). Due to the difficulty in diagnosing CUP cases, oncologists often resort to empiric treatment regimens[21,39], even when targeted therapies would

Fig. 5 | Potential clinical decision support for patients with CUP based on OncoNPC predictions of their tumors. a, Forest plot of multivariable Cox proportional hazards regression on patients in the CUP cohort with first-line palliative treatment records at DFCI ($n = 158$; refer to Extended Data Fig. 6 for the exclusion criteria). Treatment concordance (colored in blue), encoded as 1 when the first palliative treatment a patient received at DFCI is concordant with their corresponding OncoNPC prediction and 0, otherwise, was significantly associated with overall survival of patients in the cohort (HR = 0.348, 95% CI = 0.210–0.570, $P = 2.32 \times 10^{-5}$). **b**, Estimated survival curves for patients with CUP in the concordant treatment group (shown in blue) and discordant treatment group (shown in red), respectively. To estimate the survival function for each group, we used IPTW Kaplan–Meier estimator while adjusting for left truncation until the time of sequencing (Methods). Statistical significance of the survival difference between the two groups was estimated by a weighted log-rank test. **c**, Sankey diagram showing the OncoNPC-predicted cancer types, corresponding actionable variants and eligible drugs for 24 patients with CUP, which represented 15.2% of the patients in the treatment concordance analysis cohort ($n = 158$). These patients were identified as having the potential to receive genomically guided treatments based on their OncoNPC-predicted cancer types and actionable variants.

otherwise be the standard of care for a corresponding known primary. Upon retrospective chart review, we found that only 12.7% of patients with CUP (20 of 158) received genomically guided targeted treatments, which could have potentially increased to 44 (27.8%) patients based on OncoNPC predictions. In future work, we envision a multimodal foundational framework that incorporates molecular sequencing together with patient pathology images[37], longitudinal physiological data[40] and clinical notes[41] to directly predict optimal treatment regimens rather than just cancer types. We believe that our work paves a way for incorporating routine panel sequencing data into clinical decision support tools for clinically challenging cancers.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-023-02482-6.

References

1. Pavlidis, N., Khaled, H. & Gaafar, R. A mini review on cancer of unknown primary site: a clinical puzzle for the oncologists. *J. Adv. Res.* **6**, 375–382 (2015).
2. Varadhachary, G. R. & Raber, M. N. Cancer of unknown primary site. *N. Engl. J. Med.* **371**, 757–765 (2014).
3. Hyman, D. M. et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *N. Engl. J. Med.* **373**, 726–736 (2015).
4. Hainsworth, J. D. & Greco, F. A. Cancer of unknown primary site: new treatment paradigms in the era of precision medicine. *Am. Soc. Clin. Oncol. Educ. Book* **38**, 20–25 (2018).
5. Anderson, G. G. & Weiss, L. M. Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. *Appl. Immunohistochem. Mol. Morphol.* **18**, 3–8 (2010).
6. Oien, K. & Dennis, J. Diagnostic work-up of carcinoma of unknown primary: from immuno-histochemistry to molecular profiling. *Ann. Oncol.* **23**, 271–277 (2012).
7. Moran, S. et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol.* **17**, 1386–1395 (2016).
8. Jiao, W. et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728 (2020).
9. Penson, A. et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol.* **6**, 84–91 (2020).
10. He, B. et al. A neural network framework for predicting the tissue-of-origin of 15 common cancer types based on RNA-seq data. *Front. Bioeng. Biotechnol.* **8**, 737 (2020).
11. Nguyen, L., Van Hoeck, A. & Cuppen, E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat. Commun.* **13**, 4013 (2022).
12. Posner, A. et al. A comparison of DNA sequencing and gene expression profiling to assist tissue of origin diagnosis in cancer of unknown primary. *J. Pathol.* **259**, 81–92 (2023).
13. Zhao, Y. et al. CUP-AI-Dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* **61**, 103030 (2020).
14. Consortium, A. P. G. et al. AACR project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818–831 (2017).
15. Hainsworth, J. D. et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon Research Institute. *J. Clin. Oncol.* **31**, 217–223 (2013).
16. Yoon, H. et al. Gene expression profiling identifies responsive patients with cancer of unknown primary treated with carboplatin, paclitaxel, and everolimus: NCCTG N0871 (alliance). *Ann. Oncol.* **27**, 339–344 (2016).
17. Hayashi, H. et al. Site-specific and targeted therapy based on molecular profiling by next-generation sequencing for cancer of unknown primary site: a nonrandomized phase 2 clinical trial. *JAMA Oncol.* **6**, 1931–1938 (2020).
18. Hayashi, H. et al. Randomized phase II trial comparing site-specific treatment based on gene expression profiling with carboplatin and paclitaxel for patients with cancer of unknown primary site. *J. Clin. Oncol.* **37**, 570–579 (2019).
19. Conway, A.-M., Mitchell, C. & Cook, N. Challenge of the unknown: how can we improve clinical outcomes in cancer of unknown primary? *J. Clin. Oncol.* **37**, 2089–2090 (2019).
20. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 785–794 (Association for Computing Machinery, 2016).
21. Bochtler, T. & Krämer, A. Does cancer of unknown primary (CUP) truly exist as a distinct cancer entity? *Front. Oncol.* **9**, 402 (2019).
22. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
23. Tate, J. G. et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
24. da Cunha Santos, G., Shepherd, F. A. & Tsao, M. S. EGFR mutations and lung cancer. *Annu. Rev. Pathol.* **6**, 49–69 (2011).
25. Zhang, Y.-L. et al. The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis. *Oncotarget* **7**, 78985 (2016).
26. Hecht, S. S. Tobacco smoke carcinogens and lung cancer. *J. Natl Cancer Inst.* **91**, 1194–1210 (1999).
27. Dirican, E., Akkiprik, M. & Özer, A. Mutation distributions and clinical correlations of PIK3CA gene mutations in breast cancer. *Tumor Biol.* **37**, 7033–7045 (2016).
28. Elsheikh, S. et al. CCND1 amplification and cyclin D1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Res. Treat.* **109**, 325–335 (2008).

29. Kim, J. et al. Unfavourable prognosis associated with K-ras gene mutation in pancreatic cancer surgical margins. *Gut* **55**, 1598–1605 (2006).
30. Luo, J. KRAS mutation in pancreatic cancer. *Semin. Oncol.* **48**, 10–18 (2021).
31. Conway, A. M. et al. Molecular characterisation and liquid biomarkers in carcinoma of unknown primary (CUP): taking the 'U' out of 'CUP'. *Br. J. Cancer* **120**, 141–153 (2019).
32. Liu, R. et al. Systematic pan-cancer analysis of mutation–treatment interactions using large real-world clinicogenomics data. *Nat. Med.* **28**, 1656–1661 (2022).
33. Liu, R. et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).
34. Grambsch, P. M. & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994).
35. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **1**, PO.17.00011 (2017).
36. Moiso, E. et al. Developmental deconvolution for classification of cancer origin. *Cancer Discov.* **12**, 2566–2585 (2022).
37. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
38. Fizazi, K. et al. Cancers of unknown primary site: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**, v133–v138 (2015).
39. Mileshkin, L. et al. Cancer-of-unknown-primary-origin: a SEER–Medicare study of patterns of care and outcomes among elderly patients in clinical practice. *Cancers* **14**, 2905 (2022).
40. Moon, I., Groha, S., & Gusev, A. SurvLatent ODE: a neural ODE based time-to-event model with competing risks for longitudinal data improves cancer-associated venous thromboembolism (VTE) prediction. In *Proceedings of the 7th Machine Learning for Healthcare Conference*. 800– 827 (PMLR, 2022).
41. Kehl, K. L. et al. Natural language processing to ascertain cancer outcomes from medical oncologist notes. *JCO Clin. Cancer Inform.* **4**, 680–690 (2020).

Methods

Our research complies with all relevant ethical regulations. Tumor samples at DFCI were selected and sequenced from patients who were consented under institutional review board (IRB)-approved protocol 11-104 and 17-000 from the Dana-Farber/Partners Cancer Care Office for the Protection of Research Subjects. Participants in this study provided written informed consent before being included. The secondary analyses of preexisting data were conducted with approval from the Dana-Farber IRB under protocols 19-033 and 19-025. Waivers for Health Insurance Portability and Accountability Act authorization were granted for both protocols.

Patients and tumor samples

We used the NGS-targeted panel sequencing data collected at three institutions in routine clinical care as part of the AACR Project GENIE[14] as follows: DFCI (n = 18,816), MSK (n = 16,294) Cancer Center and VICC (n = 1,335). The collected tumor samples represented 22 different cancer types and included 971 total samples from CUP. National Death Index (NDI) and clinical death and last clinical appointment records were available for 20,281 DFCI patients (n = 16,376 for CKP and n = 838 for CUP). Demographic details of the patients and tumor samples can be found in Table 1.

The cancer centers, DFCI, MSK and VICC, were chosen because of similar genomic data characterization of their sequence panels in terms of coverage and alteration types[14]. DFCI samples were sequenced using a custom, hybridization-based panel called OncoPanel which targeted exons of 304–447 genes across three panel versions[14,42]. MSK samples were sequenced using a custom panel called MSK-IMPACT which targeted 341–468 genes across three panel versions[14,43]. VICC samples were sequenced using custom panels called VICC-01-T5A and VICC-01-T7, which targeted 322 and 429 genes, respectively[14]. All panels were capable of detecting SNVs, small indels, CNAs and structural variants[14]. In addition, we have provided Supplementary Data 1 which lists all the genes used to develop the OncoNPC classifier, categorized by the targeted genes across panels.

The DFCI CUP cohort consisted of 971 sequenced tumor samples (from 962 patients) with a cancer diagnosis of CUP and the following detailed cancer types: adenocarcinoma, not otherwise specified (NOS; n = 345), CUP, NOS (n = 194), squamous cell carcinoma, NOS (n = 114), poorly differentiated carcinoma, NOS (n = 118), neuroendocrine tumor/carcinoma, NOS (n = 170), small cell carcinoma of unknown primary (n = 16), undifferentiated malignant neoplasm (n = 12) and mixed cancer types (n = 2). For downstream clinical analyses, we applied additional exclusion criteria, as described in Extended Data Fig. 6.

Developing OncoNPC cancer-type classifier

We used a gradient tree boosting framework (XGBoost[20]) to develop OncoNPC for predicting cancer types from molecular features. In this framework, decision trees for the input features are sequentially added to an existing ensemble of the trees, such that the algorithm fits the new tree to the residuals from the ensembles with regularization on the tree structure. As the trees (that is, weak learners) are added, the model learns optimal weights to combine their predictions and produces the improved outcome from the combined ensemble. Owing to its high performance and scalability, the XGBoost method has been used across a wide range of applications in the healthcare space[44–46].

OncoNPC was trained and evaluated using tumors from 22 known cancer types split into 29,176 training samples and 7,289 test samples. Hyperparameter selection was conducted using random search[47] with 10-fold cross validation within the training set while using weighted F1 score as an evaluation metric. The optimal hyperparameters were then selected, and the model was evaluated on the held-out test set (n = 7,289). To predict the primary sites of CUP tumors, the model was then retrained on all CKP tumor samples and applied to the CUP tumors to estimate posterior probabilities across the 22 different cancer labels. For each tumor sample, a cancer type with the highest probability was chosen as the predicted primary site.

Feature selection and OncoNPC model interpretation

The OncoNPC model was trained on somatic variant features from tumor sequencing data and patient age at the time of sequencing and sex. To avoid bias toward known cancers or creating performance disparities across patient subgroups, OncoNPC did not consider other aspects of tumor characteristics, pathology or patient demographics (refer to Supplementary Note 1 for more details). Somatic variant features included mutations (that is, single nucleotide variants (SNV) and indels), CNA events and mutational signatures[48]. For each gene, the total count of a somatic mutation (that is, SNV and indels) was encoded as a positive integer feature. The presence of a CNA event for each gene was encoded as a categorical variable with the following five levels: −2 (deep loss), −1 (single-copy loss), 0 (no event), 1 (low-level gain) and 2 (high-level amplification). Note that CNA event data for tumor samples from MSK and VICC were encoded as −2 (deep loss), 0 (no event) and 2 (high-level amplification). Each of the 60 different mutation signatures was inferred as the dot product of the weights derived from ref. 48 and 96 single-base substitutions in a trinucleotide context. The single-base substitutions were computed using the `deconstructSigs` v1.8.0 R library[49]. Refer to Supplementary Data 2 for the full set of features.

To identify important features in OncoNPC's predictions, we used the recently proposed feature interpretation tool for tree-based models, called TreeExplainer[22] (Python `shap` v0.41.0). TreeExplainer uses an efficient polynomial time algorithm ($O(TLD^2)$, where T is number of trees, L is number of leaves and D is maximum depth) to approximate Shapley values which capture the impact of each feature on each individual model prediction. The Shapley value assigned to each feature is modeled as the average change in the model's conditional expectation function over all possible feature orderings when introducing the corresponding feature into the model. It is formulated as $\mathbb{E}_S\left[f(X)|\mathrm{do}\left(X_S = x_S\right)\right]$, where S is the set of features, X is a random variable for the feature to perturb and do[50] reflects the causal feature perturbation formulation. Refer to ref. 22 for more details on the algorithm and its properties.

Using TreeExplainer, we obtained local explanations for each OncoNPC prediction on a total of 7,289 CKP held-out and 971 CUP tumor samples. By averaging local explanations for each cancer type, we characterized the cancer type in terms of the most important or predictive features based on their mean absolute SHAP values (that is, $\hat{\mu}(|\mathrm{SHAP}|)$), which provided insights into the somatic variants and clinical features most relevant to the classification of each cancer type. Moreover, to identify the key features, we aggregated $\hat{\mu}(|\mathrm{SHAP}|)$ for each input feature by averaging them across 22 cancer types and ranked the features by their aggregated SHAP values (Supplementary Data 3). Finally, we evaluated OncoNPC's robustness by gradually reducing the input features down to the top 10% in a feature ablation study (Supplementary Note 4 and Supplementary Data 4).

Germline PRS-based validation on CUP tumor samples

To validate the OncoNPC predictions for CUP tumor samples (which do not otherwise have a ground truth), we used germline PRSs which were never available to OncoNPC for training. Germline imputation from the off-target tumor sequencing data was conducted as previously described in ref. 51. We limited our cohorts to individuals of European ancestry because the imputation model for germline variants and GWAS data for PRS was trained on a European population. Using weights from external GWAS data, we imputed PRS for NSCLC, BRCA, COADREAD, DIFG, MEL, OVT, RCC and PRAD. Pearson correlation between the PRS from off-target tumor data versus matched germline SNP array was previously shown to be higher than 0.9 without observable outliers[51]. Refer to Supplementary Note 6 for details on the accuracy of germline imputation in our cohorts.

We hypothesized that germline PRS specific to the underlying primary cancer type of a CUP tumor sample would be enriched in a manner similar to how the PRS specific to CKP tumor sample with the same primary cancer type is enriched. To that end, given the set of eight different cancer types $\mathscr{C}$ we have the imputed PRS available for, we first restricted the cohort of CUP tumor samples to those with OncoNPC predictions in $\mathscr{C}$ ($n_{\text{CUP},\mathscr{C}} = 505$). Then, we obtained standardized germline PRS values for the chosen CUP tumor samples over all the cancer types in $\mathscr{C}$. Finally, we defined $\hat{\Delta}_{\text{PRS}}$ as the estimated mean difference between the PRS specific to the predicted primary cancer type C (that is concordant PRS; PRS_C) and average of PRSs corresponding to the rest of the cancer types (that is, discordant PRS; PRS_D, where $D \in \mathscr{C} \setminus C$) as follows:

$$\hat{\Delta}_{\text{PRS}} = \hat{\mathbb{E}}[\text{PRS}_C - \hat{\mathbb{E}}_D[\text{PRS}_D|C]]$$
$$= \frac{1}{n_{\text{CUP},\mathscr{C}}} \sum_{i}^{n_{\text{CUP},\mathscr{C}}} \left(\text{PRS}_{c_i} - \frac{1}{|\mathscr{C}\setminus c_i|} \sum_{d_i \in \mathscr{C}\setminus C_i} \text{PRS}_{d_i}\right)$$

As a true positive reference, we repeated the above procedure for the CKP tumor samples. Finally, as a true negative reference, we estimated $\hat{\Delta}_{\text{PRS-random}}$, where the concordant cancer type was randomly assigned. We then repeated the random assignment 100 times to obtain estimated mean and s.e.

Survival function estimation

NDI and in-house clinical records were available for 20,281 DFCI patients (n = 16,376 for CKP and n = 838 for CUP). A patient's loss to follow-up date was determined at either the last NDI update date (12/31/2020) or their corresponding last contact date from the in-house records, whichever date is later. A patient's death date was determined from the in-house records or the NDI data if the patient was lost to follow-up.

OncoNPC-based risk stratification among patients with CUP

To identify OncoNPC CUP subgroups with significant prognostic differences, we estimated survival functions for the following seven common OncoNPC subgroups with more than 35 patients with CUP: NSCLC, PAAD, BRCA, HNSCC, EGC, GINET and PANET. Patients who were lost to follow-up at the time of sequencing were excluded, as were CUPs with an OncoNPC prediction probability lower than 0.5 (Extended Data Fig. 6). We merged subgroups with similar morphology and estimated survival functions—PAAD and EGC, and GINET and PANET. To statistically test survival differences between these five groups, we used a chi-squared test with four degrees of freedom.

Estimating impacts of treatment concordance on survival of patients with CUP

We estimated the impact of the concordance between treatment and OncoNPC CUP predictions on a mortality outcome in a retrospective survival analysis. We used the in-house patient follow-up and treatment data to identify patients with CUP who received first treatment at DFCI with a palliative intent (refer to Extended Data Fig. 6 for the exclusion criteria). Each patient was reviewed by a trained oncologist to determine whether the OncoNPC-predicted cancer type was concordant or discordant with the first line of treatment received, as per NCCN guidelines or standard of care, in most reasonable situations, and within the clinical context delineated in the medical record (Supplementary Note 9). Refer to Supplementary Data 5 for more details on clinical information of patients with CUP in the analysis, including primary cancer diagnosis, biopsy site and first chemotherapy plan at DFCI.

As we were interested in the counterfactual causal impact of the OncoNPC-treatment concordance, we used the principles of causal inference to account for potential patient heterogeneity and confounding. Specifically, we estimated the effect of treatment concordance specified by the indicator variable, A, which was 1 when the first palliative treatment for a patient with CUP was concordant with the corresponding OncoNPC prediction and 0 otherwise. Our analyses make the following identifiability assumptions:

- Conditional ignorability: $A_i \perp\!\!\!\perp T_i^{(a_i)} | X_i$, where $A_i \in 0, 1$. It means that given patient i's a set of covariates X_i, the patient's treatment concordance A_i is as good as random.
- Consistency : $T_i^{a_i} = T_i$, which means that a counterfactual outcome $T_i^{a_i}$ for patient i is the observed outcome for the patient with a treatment concordance a_i.
- Overlap : $P(0 < p(X_i) < 1) = 1$ where $p(X_i) = P(A_i = 1|X_i)$, which means all patients have a strictly positive probability of receiving concordant treatment ($A_i = 1$).

In addition to the above identifiability assumptions, we made independent censoring (that is, $C_i \perp\!\!\!\perp T_i | X_i$) and independent entry assumption given the covariates (that is, $E_i \perp\!\!\!\perp T_i | X_i$).

We adopted two different estimation strategies to obtain the impact of treatment concordance as follows: semiparametric Cox proportional hazard estimator adjusted with a set of measured confounders X (ref. 52) and nonparametric Kaplan–Meier estimator adjusted with IPTW. We formulated an IPTW, w_i, for each sample as $w_i = \frac{P(A=a_i)}{P(A_i=a_i|X_i)}$ (ref. 53) and estimated $P(A)$ nonparametrically and $P(A|X)$ using a logistic regression model[54] in a 10-fold cross-fitting. A set of measured confounders (that is, X_i) included patients' sex, age, OncoNPC prediction uncertainty (in entropy), sequencing panel (that is, OncoPanel) version, mutational burden, CNA burden, subsets of OncoNPC-predicted cancer types and metastasis sites and finally pathological histology (for example, adenocarcinoma tumor or neuroendocrine tumor). Because patients with CUP who met the treatment criteria but did not receive clinical panel sequencing (that is, entry criterion) could not be included in the analysis, we adjusted for the left truncation by defining the risk set $\mathscr{R}(t)$ at time t, which corresponds to the set of patients followed up in the analysis up to time t as follows.

$\mathscr{R}(t) = \{i | E_i \le t \le T_i\}$, where E_i is the entry time of patient i. With the independent entry assumption as stated before, we obtained survival function from the Kaplan–Meier estimator as follows.

$\hat{S}(t) = \prod_{i:T_i \le t} \left(1 - \frac{\sum_{k:T_k=T_i} w_k}{\sum_{j:j\in\mathscr{R}(T_i)} w_j}\right)$. In this formulation, each individual is weighted by the corresponding IPTW, w_i, and we obtained two different survival functions for the treatment concordant and discordant groups. The adjusted Kaplan–Meier estimator provides a consistent estimate of the survival function for each group under the assumptions stated above. Once we obtained the survival estimates for the two groups, we used a weighted log-rank test to test for a significant difference in survival.

In the Cox proportional hazard regression framework, we estimated the hazard function of patient i as follows: $\lambda(t|A_i, X_i)) = \lambda_0(t) \exp(\alpha A_i + \beta^T X_i)$, where $\alpha, A_i \in \mathbb{R}$ and $\beta, X_i \in \mathbb{R}^m$ (m is the number of measured confounders). Under the above identifiability assumptions and validity of the estimation model, e^{α} is the hazard ratio capturing the causal effect of the treatment concordance A. Finally, under the assumption of no ties between event times across the patients, the parameters α and β are estimated by maximizing the following partial likelihood.[52]

$$L(\alpha, \beta) = \prod_{i:\delta_i=1} \frac{\exp(\alpha A_i + \beta X_i)}{\sum_{j:j\in\mathscr{R}(T_i)} \exp(\alpha A_j + \beta X_j)}$$

Actionable somatic variants in CUP tumors

We estimated the frequency of known, actionable somatic alterations in each OncoNPC CUP subgroup using the OncoKB knowledge base[35]. We considered three different types of somatic variants as follows: oncogenic mutations such as indels, missense mutations and splice site

mutations, amplifications such as high-level amplifications and finally fusions such as gene–gene and gene–intergenic fusions as specified in OncoKB. For each actionable somatic variant, we assigned one of the four therapeutic levels as follows: level 1 for FDA-approved drugs, level 2 for standard care drugs, level 3 for drugs supported by clinical evidence and level 4 for drugs supported by biological evidence. Refer to Supplementary Data 5 for more details on actionable variants and corresponding genomically guided treatments.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The multicenter NGS tumor panel sequencing data is available upon request at the AACR Project GENIE website: https://www.aacr.org/professionals/research/aacr-project-genie/. The fully trained OncoNPC model, processed somatic variants data from Profile DFCI and deidentified clinical data used in the treatment concordance analysis are available in https://github.com/itmoon7/onconpc.

Code availability

We used the R (v4.0.2) and Python (v3.9.13) programming languages for OncoNPC feature processing (R `deconstructSigs` v1.8.0), OncoNPC model development and interpretation (Python `xgboost` v1.2.0, `shap` v0.41.0) and survival analysis (R `survival` v3.2.7, `stats` v4.0.2, Python `lifelines` v0.27.4, `scipy` v1.7.1). Please see https://github.com/itmoon7/onconpc for the preprocessing script, the fully trained OncoNPC model, a notebook demonstration on how to use OncoNPC and other reference materials.

References

42. Garcia, E. P. et al. Validation of oncopanel: a targeted next-generation sequencing assay for the detection of somatic variants in cancer. *Arch. Pathol. Lab. Med.* **141**, 751–758 (2017).
43. Cheng, D. T. et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
44. Chen, Y. et al. Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost. *Physiol. Meas.* **39**, 104006 (2018).
45. Hatton, C. M. et al. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. *J. Affect. Disord.* **246**, 857–860 (2019).
46. Ogunleye, A. & Wang, Q.-G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 2131–2140 (2019).
47. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
48. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
49. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
50. Janzing, D., Minorics, L. & Blöbaum, P. Feature relevance quantification in explainable AI: a causal problem. In *Proceedings of International Conference on Artificial Intelligence and Statistics* 2907–2916 (PMLR, 2020).
51. Gusev, A., Groha, S., Taraszka, K., Semenov, Y. R. & Zaitlen, N. Constructing germline research cohorts from the discarded reads of clinical tumor sequences. *Genome Med.* **13**, 179 (2021).
52. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–202 (1972).
53. Xie, J. & Liu, C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat. Med.* **24**, 3089–3110 (2005).
54. Marschner, I. glm2: Fitting generalized linear models with convergence problems. *The R Journal* **3**, 12–15 (2011).

Acknowledgements

The participation of patients and the efforts of an institutional data collection system made this study possible, and we are grateful for their contributions. We would also like to express our appreciation to the DFCI Oncology Data Retrieval System (OncDRS) and AACR Project GENIE team for their role in aggregating, managing and delivering the data used in this project.

I.M. and A.G. were supported by R01 CA227237, R01 CA244569 and grants from The Louis B. Mayer Foundation, The Doris Duke Charitable Foundation, The Phi Beta Psi Sorority and The Emerson Collective. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

I.M. and A.G. conceived and designed the study. I.M. curated the data, developed and evaluated the model and performed analyses. J.L. and L.S. performed clinical chart reviews. I.M. wrote the first manuscript. I.M., J.L. and G.S. revised the manuscript. All the authors took part in interpreting the findings and reviewing the manuscript.

Competing interests

The authors declare no conflicts of interest.

Additional information

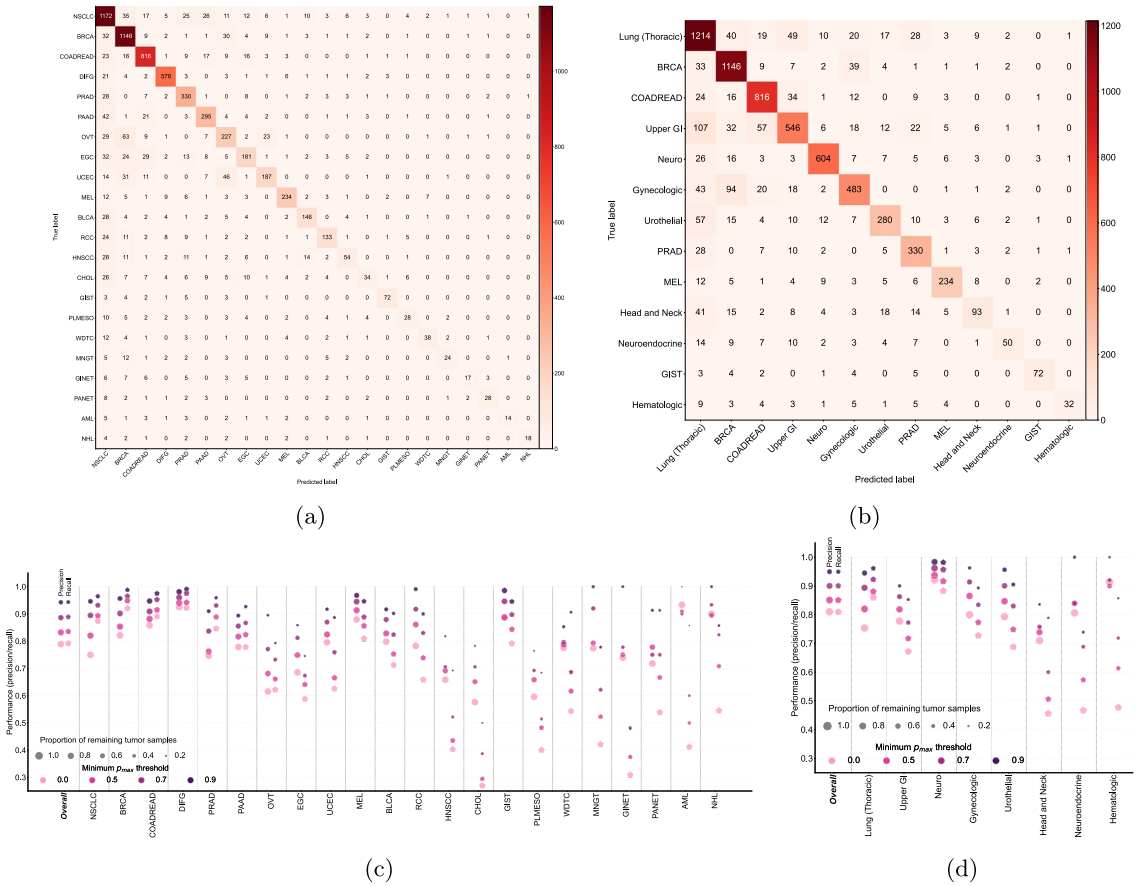
Extended data is available for this paper at https://doi.org/10.1038/s41591-023-02482-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-023-02482-6.

Correspondence and requests for materials should be addressed to Alexander Gusev.

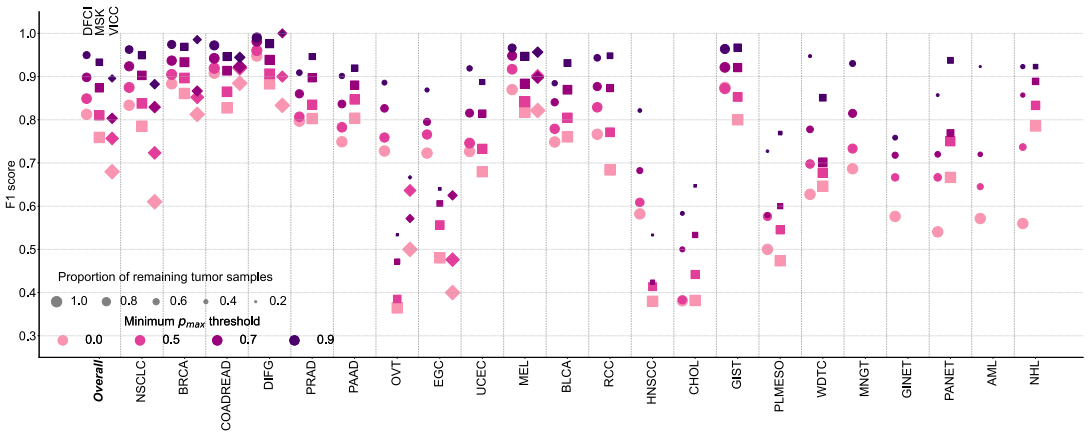
Peer review information *Nature Medicine* thanks Lincoln Stein, Linda Mileshkin and E. Cuppen for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-023-02482-6

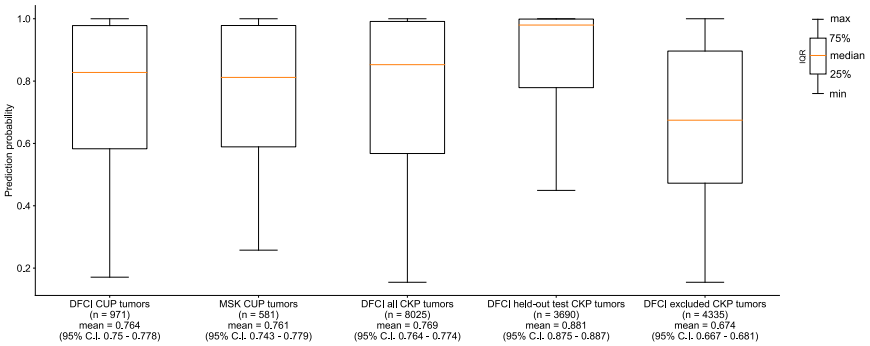


Extended Data Fig. 1 | OncoNPC classification performance: confusion matrix, and precision and recall. Confusion matrices on the held-out test set (n = 7,289) for **(a)** 22 detailed cancer types and **(b)** 13 cancer groups (see Table 1). **(c)**,**(d)** OncoNPC performance in precision and recall on the test set across **(c)** cancer types and **(d)** cancer groups at 4 different prediction confidences using p_{max} as a threshold. Each dot size is scaled by the proportion of tumor samples retained. In (d), we only considered cancer groups that have more than one cancer type. Overall F1 scores were weighted according to the number of confirmed cases across cancer types and cancer groups, respectively.

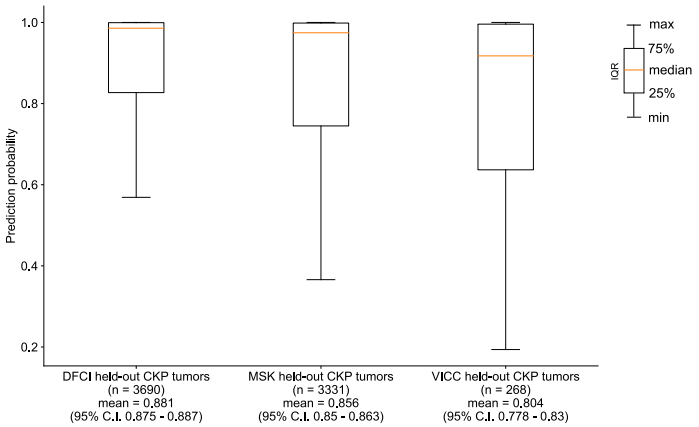
https://doi.org/10.1038/s41591-023-02482-6



(a)



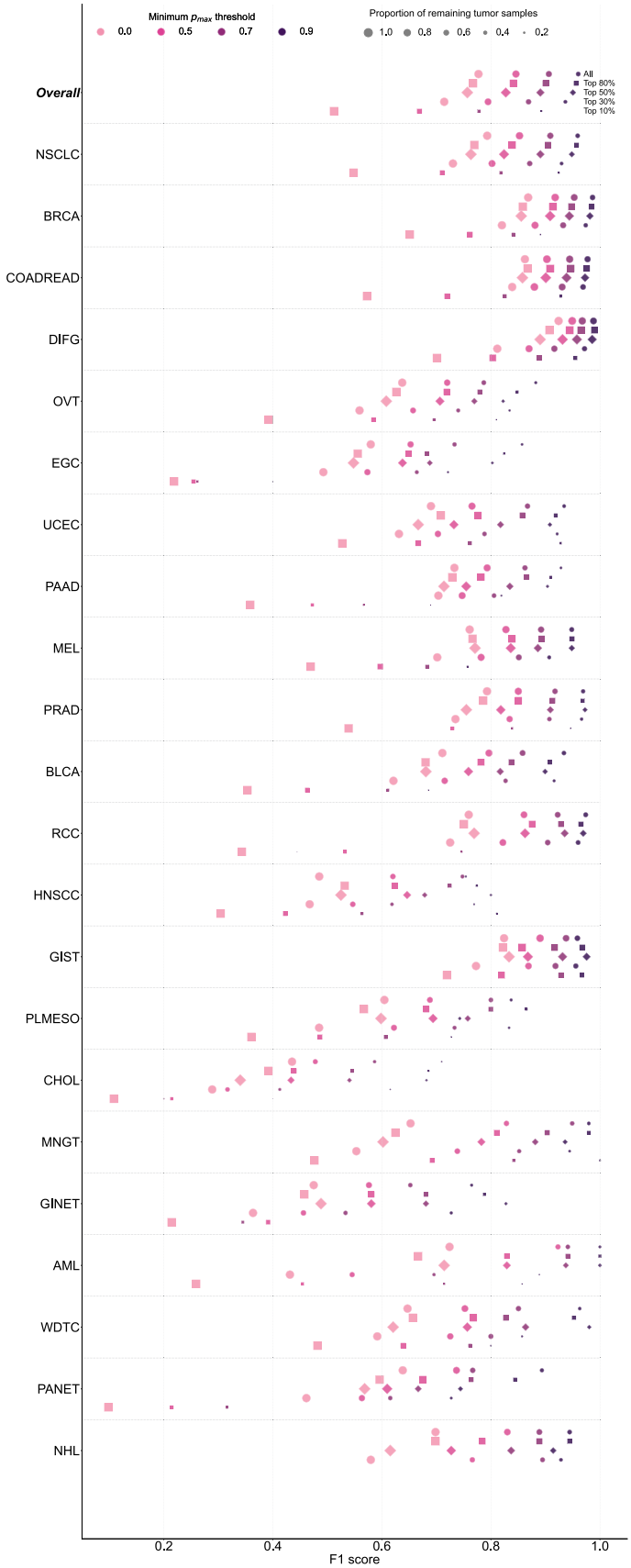
(b)



(c)

Extended Data Fig. 2 | OncoNPC prediction performance and prediction confidence levels (that is, p_{max}) across different cohorts and centers. **(a)**, Center-specific OncoNPC performance (in 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 (that is, minimum p_{max} thresholds). Each dot size is scaled by the proportion of tumor samples retained. See Supplementary Table 3 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 confidence (p_{max}) across **(b)** DFCI CUP tumors, MSK CUP tumors, all DFCI CKP tumors (including those with cancer types not modeled in OncoNPC), DFCI held-out CKP tumors, and DFCI excluded CKP tumors (specifically those with cancer types not modeled in OncoNPC), and **(c)** DFCI held-out CKP tumors, MSK held-out CKP tumors, and VICC held-out CKP tumors. Note that DFCI excluded CKP tumors refers to the cohort of the rare CKP tumors whose cancer types were not considered during the development of OncoNPC. All cohorts in the analysis for **(b)** and **(c)** were not seen by OncoNPC during the model training.

https://doi.org/10.1038/s41591-023-02482-6

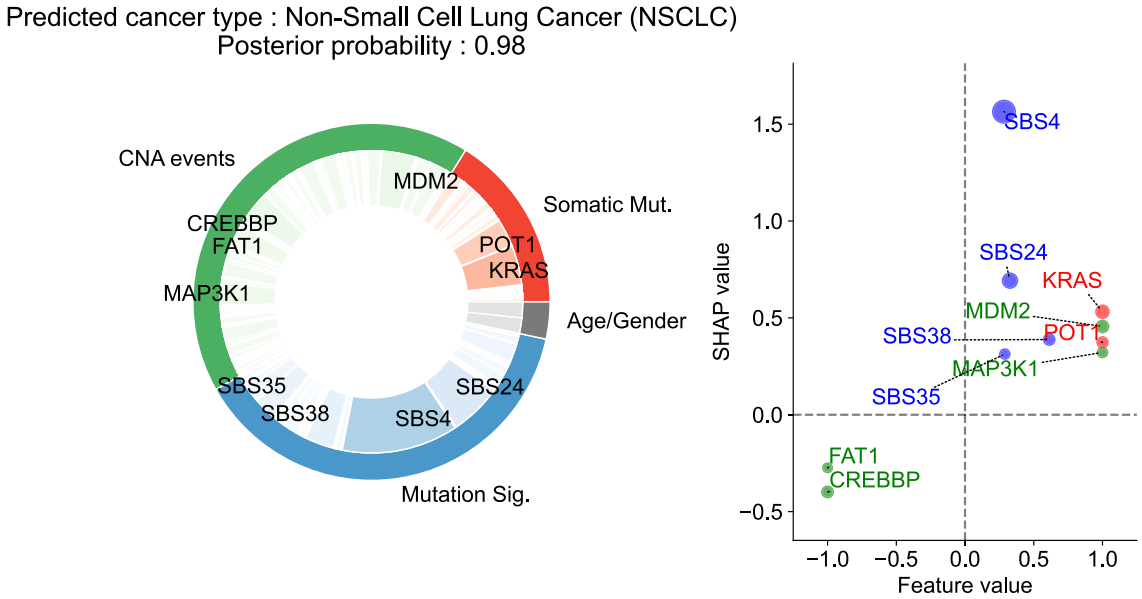


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-023-02482-6

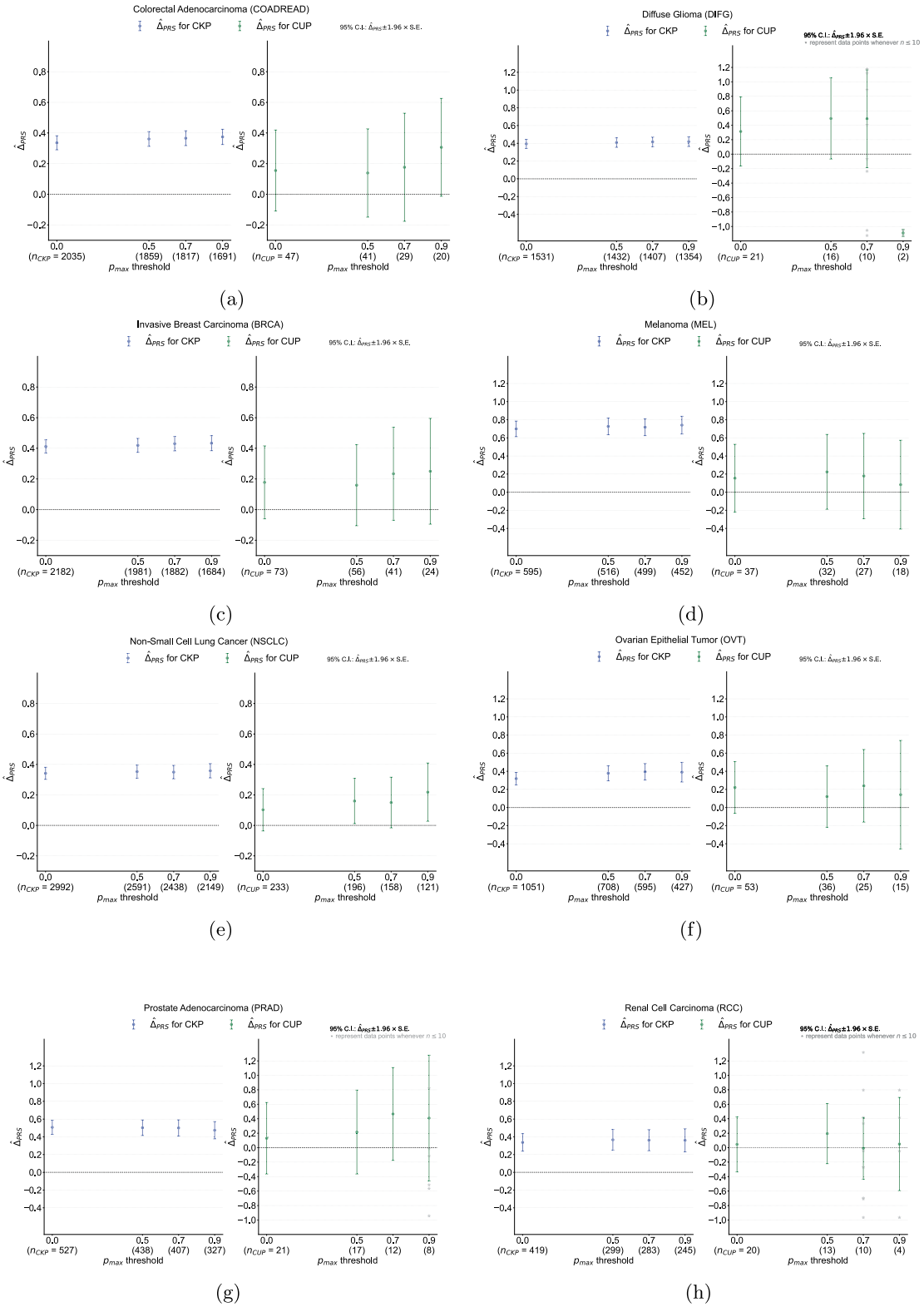
Extended Data Fig. 3 | Robustness of OncoNPC performance with respect to input genomics features. The figure shows the breakdown of OncoNPC performance in F1 score by 22 cancer types across increasing prediction confidence. The cancer types on the y-axis are sorted in a decreasing order of the number of tumor samples. In order to investigate the impact of input genomics features on OncoNPC's robustness, we performed a feature ablation study, where we chose the most important genes based on their aggregated SHAP values and gradually reduced them from all 846 features associated with those genes, as well as age and sex, to only the top 10% (that is, top 29 features). In each feature configuration, we re-trained the model with the same set of hyperparameters and evaluated its performance on the held-out CKP tumor samples (n = 7,289), which were utilized throughout this work. Supplementary Data 4 provides a list of input features that correspond to the selected genes in each configuration.

https://doi.org/10.1038/s41591-023-02482-6



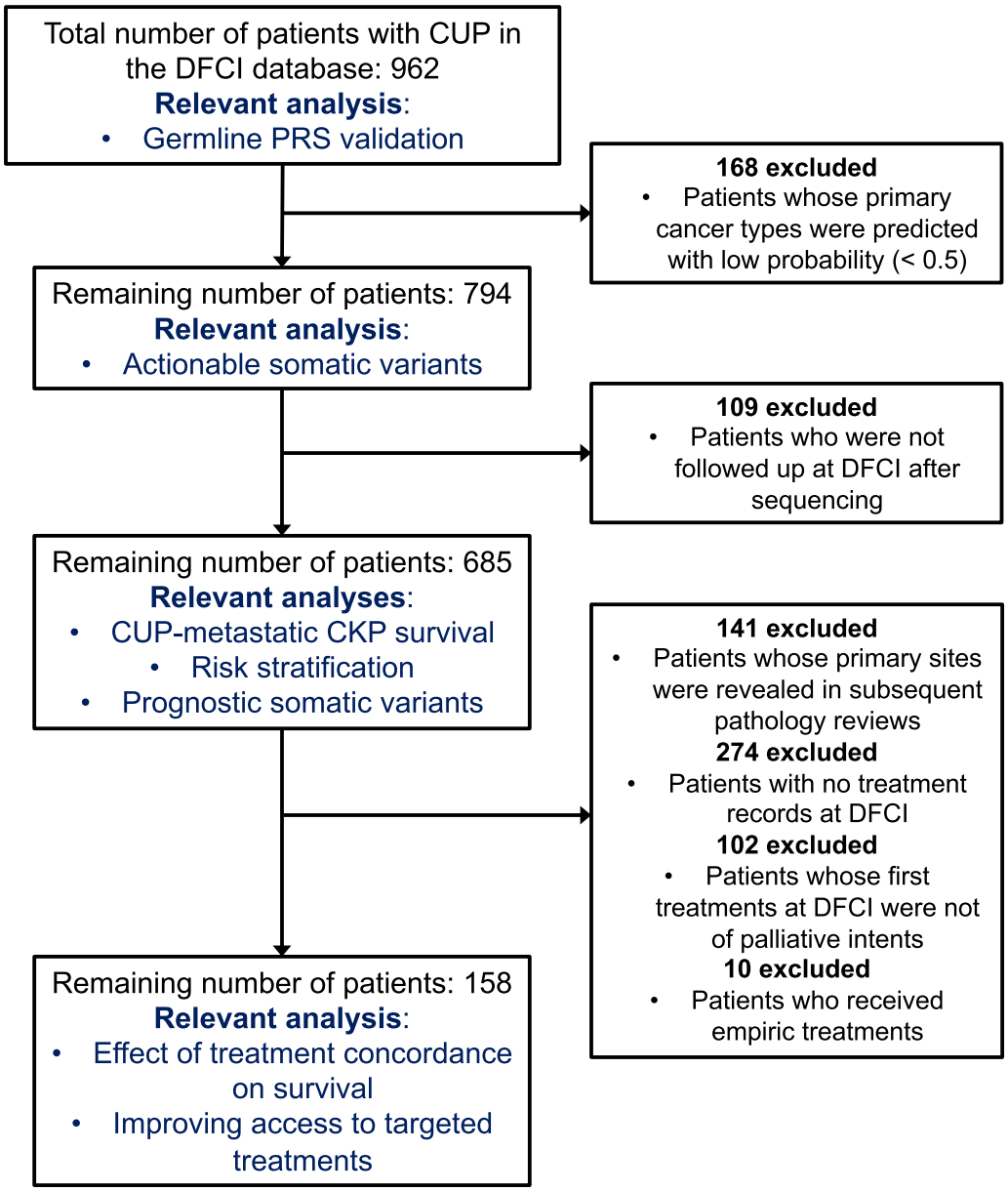
Extended Data Fig. 4 | Explanation of OncoNPC prediction for a patient with CUP. The patient is a 76-year-old male with a tumor biopsy from the liver. The pie chart on the left shows the top 10 important features across three different feature categories (that is, CNA events, somatic mutation and mutation 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, and KRAS mutation.

https://doi.org/10.1038/s41591-023-02482-6



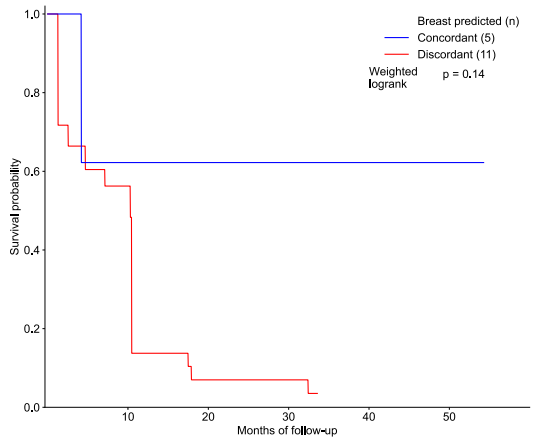
Extended Data Fig. 5 | Germline polygenic risk score (PRS) enrichment of CKP tumor samples and CUP tumor samples, broken down by 8 different cancer types. (a), Colorectal adenocarcinoma (COADREAD), **(b)** diffuse glioma (DIFG), **(c)** invasive breast carcinoma (BRCA), **(d)** melanoma (MEL), **(e)** 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 $\hat{\Delta}_{PRS}$: the mean difference between the concordant (that is OncoNPC matching) cancer type PRS and mean of PRSs of discordant cancer types (see Methods). $\hat{\Delta}_{PRS}$ is shown for CKPs in blue (for reference) and CUPs in green.

https://doi.org/10.1038/s41591-023-02482-6

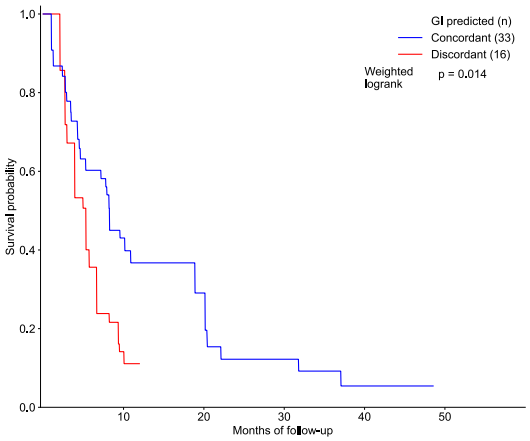


Extended Data Fig. 6 | Exclusion criteria for downstream clinical analyses. The boxes on the left show the number of the remaining patients in the cohort and relevant analyses, while the boxes on the right illustrate the exclusion criteria and the number of patients who were consequently removed.

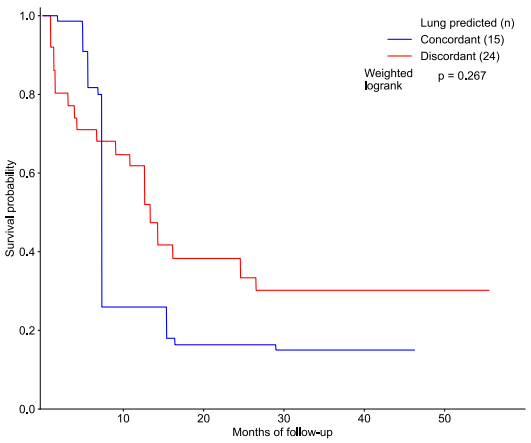
https://doi.org/10.1038/s41591-023-02482-6



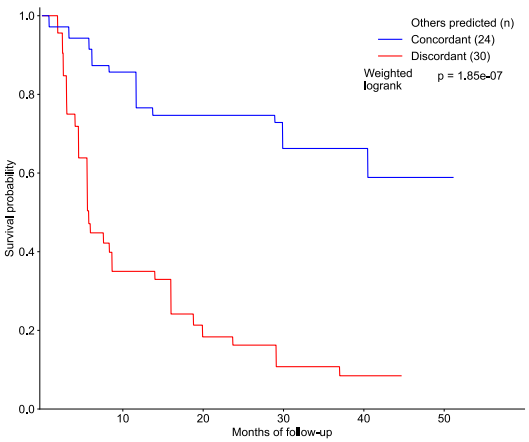
(a) Invasive Breast Carcinoma (BRCA)



(b) Gastrointestinal (GI) group



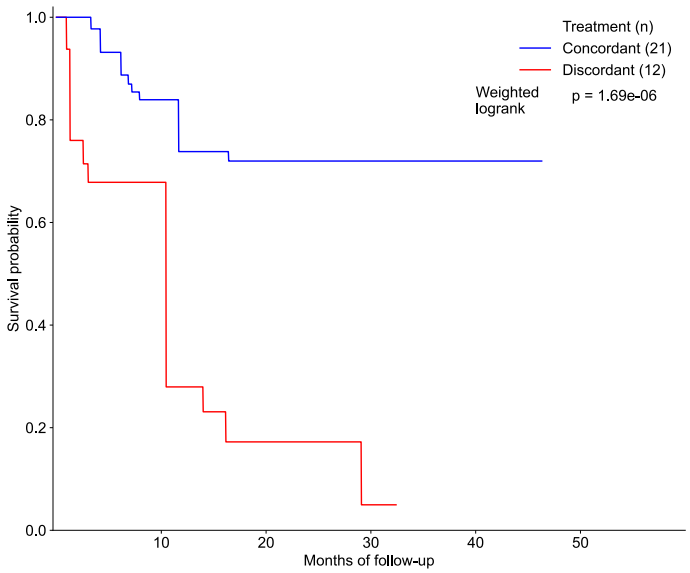
(c) Lung



(d) Other OncoNPC cancer types

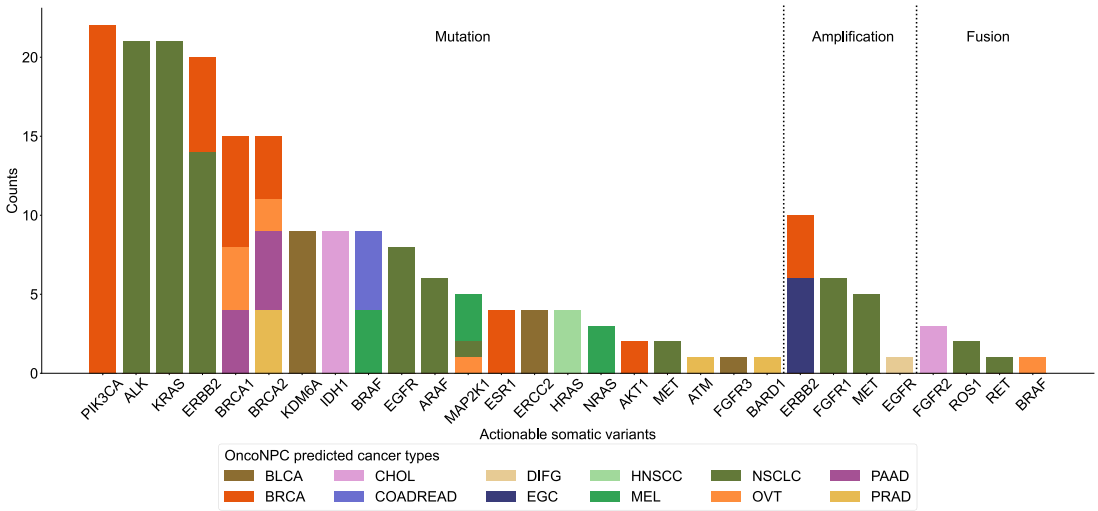
Extended Data Fig. 7 | Estimated survival curves for the concordant and discordant treatment groups among patients with CUP, broken down by OncoNPC predicted cancer types. a, BRCA, (**b**) gastrointestinal (GI) group (CHOL, COADREAD, EGC and PAAD), (**c**) lung (NSCLC and PLMESO) and (**d**) other OncoNPC 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 each survival curve, we utilized inverse probability of treatment weighted (IPTW) Kaplan-Meier estimator while adjusting for patient covariates and 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.

https://doi.org/10.1038/s41591-023-02482-6

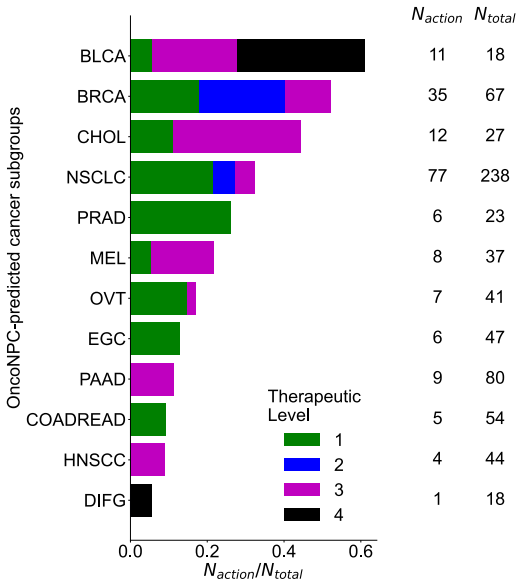


Extended Data Fig. 8 | Estimated survival curves for the concordant and discordant treatment groups among patients with CUP who received their initial treatments after the results of the OncoPanel sequencing were available to clinicians. Similarly, we utilized inverse probability of treatment weighted (IPTW) Kaplan-Meier estimator for each survival curve while adjusting for patient covariates and 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. Refer to Supplementary Table 2 for demographic information on the cohort.

https://doi.org/10.1038/s41591-023-02482-6



(a)



(b)

Extended Data Fig. 9 | OncoNPC-guided actionable variants in patients with CUP. (a), The number of CUP tumors with actionable targets, based on OncoKB (Methods), across actionable somatic variants (mutations, amplifications and fusions). Each bar corresponds to the total number of CUP tumors associated with each actionable target. The bars are color-coded by predicted cancer types. Note that each tumor may contain more than one actionable somatic variant. **(b)**, Proportions of CUP tumor samples with actionable somatic variants (N_{action}) to the total number of patients (N_{total}) across OncoNPC predicted cancer types. Proportions for four different therapeutic levels based on OncoKB are shown in each bar: level 1—FDA-approved drugs, level 2—standard of care drugs, level 3—drugs supported by clinical evidence and level 4—drugs supported by biological evidence.

https://doi.org/10.1038/s41591-023-02482-6

Extended Data Table 1 | Demographic details of patients with CUP in the concordant and discordant treatment groups

	Concordant treatment group (n = 77)	Discordant treatment group (n = 81)
Sex; male-female ratio	0.442-0.558	0.556-0.444
Age at seqeuncing (95% C.I.)	64 (61.6 - 66.4)	62 (59.4 - 64.6)
OncoNPC prediction uncertainty (in entropy; 95% C.I.)	0.550 (0.426 - 0.675)	0.988 (0.850 - 1.127)
OncoPanel version (proportion in %)		
v1	1 (1.30%)	1 (1.24%)
v2	9 (11.7%)	15 (18.5%)
v3	67 (87.0%)	65 (80.2%)
Mutational burden (95% C.I.)	0.027 (0.021 - 0.033)	0.033 (0.027 - 0.040)
CNA burden (95% C.I.)	0.201 (0.166 - 0.236)	0.186 (0.155 - 0.217)
OncoNPC predicted cancer groups (proportion in %)		
Lung	15 (19.5%)	24 (29.6%)
Breast	5 (6.50%)	11 (13.6%)
GI	33 (42.9%)	16 (19.8%)
Gyn	9 (11.7%)	5 (6.17%)
Others	15 (19.5%)	25 (31.0%)
Metastatic sites (proportion in %)		
Brain	4 (5.20%)	8 (9.88%)
Bone	7 (9.10%)	10 (12.3%)
Soft tissue	6 (7.79%)	5 (6.17%)
Others	60 (77.9%)	58 (71.6%)
Histology (proportion in %)		
Adenocarcinoma	41 (53.2%)	32 (39.5%)
Neuroendocrine	9 (11.7%)	11 (13.6%)
Squamous cell	2 (2.60%)	6 (7.41%)
Others	25 (32.5%)	32 (39.5%)
Treatment start date (95% C.I.)	2018-4-30 (2017-12-24 - 2018-9-3)	2018-3-1 (2017-10-28 - 2018-7-3)

The OncoNPC-predicted cancer groups, except for the GI group, match the cancer groups defined in Table 1. The GI group in this analysis consists of the upper GI group, including cholangiocarcinoma (CHOL), esophagogastric adenocarcinoma (EGC), pancreatic adenocarcinoma (PAAD), as well as colorectal adenocarcinoma (COADREAD).

nature portfolio

Corresponding author(s): Alexander Gusev

Last updated by author(s): Jun 26, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The multicenter NGS tumor panel sequencing data was collected as part of the AACR Project GENIE initiative, and clinical data from the DFCI was collected and curated by the OncDRS team at DFCI. We obtained access to the DFCI data through SQL requests. No additional software was utilized on the already-curated data.
Data analysis	We utilized the R (v4.0.2) and Python (v3.9.13) programming languages for OncoNPC feature processing (R deconstructSigs vl.8.0), OncoNPC model development and interpretation (Python xgboost vl.2.0, shap v0.41.0), and survival analysis (R survival v3.2.7, stats v4.0.2, Python lifelines v0.27.4, scipy vl.7.1). Finally, analysis scripts can be found at https://github.com/itmoon7/onconpc.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The multicenter NGS tumor panel sequencing data is available upon request at https://www.aacr.org/professionals/research/aacr-project-genie/. The fully trained OncoNPC model, processed somatic variants data from Profile DFCI, and de-identified clinical data used in the treatment concordance analysis are available in https://github.com/itmoon7/onconpc. OncoKB Precision Oncology Knowledge Base can be found in https://www.oncokb.org/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	In our study, we used self-reported biological sex assigned at birth as one of the input features for OncoNPC and as a covariate in our survival analyses. We included information about the patient demographics, including sex, in Tables 1.
Population characteristics	The OncoNPC model was developed and evaluated using NGS tumor sequencing data from 34,567 patients across three institutions: DFCI, MSK, and VICC. The main discovery cohort for this study consisted of 962 patients with CUP treated at DFCI. Table 1 includes demographic information for these patients, including their age, biological sex, ethnicity, panel version, and biopsy site type
Recruitment	The OncoNPC model was developed and evaluated using NGS tumor sequencing data from routine medical practice at three institutions. For our clinical analyses, we retrospectively analyzed data from patients who received routine treatment at DFCI and had undergone targeted panel sequencing. These patients were selected without regard to their survival outcomes.
Ethics oversight	DFCI PROFILE samples were selected and sequenced from patients who were consented under institutional review board (IRB)-approved protocol 11-104 and 17-000 from the Dana-Farber/Partners Cancer Care Office for the Protection of Research Subjects. Participants in this study provided written informed consent before being included. The secondary analyses of preexisting data were conducted with approval from the Dana-Farber IRB under protocols 19-033 and 19-025. Waivers for Health Insurance Portability and Accountability Act (HIPAA) authorization were granted for both protocols.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size for this study was determined by the number of patients who consented to participate and had data available at the time of data collection.
Data exclusions	For our clinical analyses, we applied exclusion criteria mainly based on diagnosed cancer types, primary site prediction confidence and patient follow-up status. Pleaser refer to Extended Data Figure 6 for more details.
Replication	We identified the most common types of cancer predicted in CUP tumor samples and found that these rates were similar to previous research showing that the most common underlying primary cancers for CUPs, as identified by autopsy, are lung, large bowel, and pancreas cancers. To replicate our findings, we applied the OncoNPC model to 581 CUP tumors at MSK and obtained comparable results. In the future, we plan to validate the relationship between treatment concordance with OncoNPC and survival benefit in the CUP cohort at MSK.
Randomization	The cohort in this study was not randomized and is observational in nature. For our clinical analyses, we took into account potential confounding factors such as age, sex, panel version, histological characteristics, and other technical covariates.
Blinding	This is an observational study, rather than a clinical trial, so the patients in the study did not receive treatment as part of a pre-specified research protocol. Instead, they received treatment based on clinical decisions made by their healthcare providers.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging