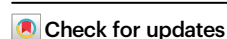


Characterizing mutation-treatment effects using clinico-genomics data of 78,287 patients with 20 types of cancers

Received: 10 July 2024

Accepted: 3 December 2024

Published online: 30 December 2024



Ruishan Liu^{1,2}, Shemra Rizzo^{3,6}, Lisa Wang^{3,6}, Nayan Chaudhary^{3,6}, Sophia Maund^{3,6}, Marius Rene Garmhausen^{4,6}, Sarah McGough³, Ryan Copping^{3,7}✉ & James Zou^{1,5,7}✉

Evaluating the effectiveness of cancer treatments in relation to specific tumor mutations is essential for improving patient outcomes and advancing the field of precision medicine. Here we represent a comprehensive analysis of 78,287 U.S. cancer patients with detailed somatic mutation profiling integrated with treatment and outcomes data extracted from electronic health records. We systematically identified 776 genomic alterations associated with survival outcomes across 20 distinct cancer types treated with specific immunotherapies, chemotherapies, or targeted therapies. Additionally, we demonstrate how mutations in particular pathways correlate with treatment response. Leveraging the large number of identified predictive mutations, we developed a machine learning model to generate a risk score for response to immunotherapy in patients with advanced non-small cell lung cancer (aNSCLC). Through rigorous computational analysis of large-scale clinico-genomic real-world data, this research provides insights and lays the groundwork for further advancements in precision oncology.

Precision medicine aims to characterize the unique responses of patients with specific genetic mutations to various treatments¹. This understanding can enhance patient outcomes by tailoring treatment recommendations to individual tumor mutation profiles. The adoption of next-generation sequencing (NGS) has revolutionized genomics profiling, positioning it as an invaluable tool in cancer care². Yet, despite the vast mutation data available, only a limited number of these mutations are associated with validated treatments³.

Recent studies have demonstrated that large-scale real-world clinico-genomics data can identify genomic biomarkers that predict patient response to cancer treatments. For example, using a cohort of 40,903 patients, researchers identified 458 statistically significant gene-treatment interactions in eight common types of cancer⁴.

In this study, we harness a substantially updated and broader dataset, with one and a half more years of data and patient follow-up (a Dec 31, 2022 data cut-off compared to a June 30, 2021 data cut-off previously), encompassing 12 additional cancer types (20 vs 8 previously), and more patients (78,287 vs 40,903 previously), to systematically identify mutations predictive of patient outcomes for specific treatments. Delving into the patients' tumor mutation profiles, treatment histories, and survival outcomes, we characterize somatic genomic alterations that notably predict the patients' survival on specific immunotherapies, chemotherapies, or targeted therapies.

Contemporary research accentuates the powerful capability of real-world data, especially those extracted from electronic health records (EHRs), in ascertaining treatment impacts, emulating control groups for singular clinical trials, and refining eligibility benchmarks

¹Department of Electrical Engineering, Stanford University, Stanford, CA, USA. ²Department of Computer Science, University of Southern California, Los Angeles, CA, USA. ³Genentech, South San Francisco, CA, USA. ⁴F. Hoffmann-La Roche Ltd., Basel, Switzerland. ⁵Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ⁶These authors contributed equally: Shemra Rizzo, Lisa Wang, Nayan Chaudhary, Sophia Maund, Marius Rene Garmhausen. ⁷These authors jointly supervised this work: Ryan Copping, James Zou. ✉e-mail: copping.ryan@gene.com; jamesz@stanford.edu

for oncology trials^{5,6}. This study demonstrates how the computational dissection of large-scale real-world data can propel precision oncology forward, offering useful insights into both gene-treatment and pathway-treatment dynamics in oncology. Moreover, we develop a machine learning model for determining the risk score for response to immunotherapy based on mutation profiles of patients with aNSCLC, enriching the precision medicine discourse.

Results

Overview of real-world cancer data

We leveraged the Flatiron Health-Foundation Medicine US-based, de-identified clinico-genomic database (FH-FMI CGDB) with follow-up until December 31, 2022. This data was sourced from ~280 US cancer clinics. The dataset includes information on patients with aNSCLC ($n=19,631$), metastatic breast cancer (mBC) ($n=13,258$), metastatic colorectal cancer (mCRC) ($n=12,705$), metastatic pancreatic cancer (mPCa) ($n=5513$), ovarian cancer (OC) ($n=5154$), metastatic prostate cancer (PC) ($n=4854$), gastric cancer (GC) ($n=4353$), advanced melanoma (aMel) ($n=2319$), advanced bladder cancer (aBCa) ($n=2266$), endometrial carcinoma (EC) ($n=2191$), metastatic renal cell carcinoma (mRCC) ($n=1500$), head and neck cancer (HNC) ($n=1222$), small-cell lung cancer (SCLC) ($n=869$), multiple myeloma (MM) ($n=525$), acute myeloid leukemia (AML) ($n=449$), hepatocellular carcinoma (HCC) ($n=450$), diffuse large B cell lymphoma (DLBCL) ($n=336$), chronic lymphocytic leukemia (CLL) ($n=308$), follicular lymphoma (FL) ($n=199$), and mantle cell lymphoma (MCL) ($n=185$). Each of these cancer types was also grouped into cancer categories (lung & upper respiratory tract; breast & ovarian, skin, gastrointestinal,

genitourinary, and hematological) based on how they may be managed in clinical practice or by types of oncology practice. The available data for each patient encompasses tumor mutations, treatment details, survival outcomes, demographics, and more. The genomic alterations were discerned through comprehensive genomic profiling of over 300 cancer-related genes using Foundation Medicine's NGS tests, and 84.2% of the tests were conducted on samples that were collected before the start of the first-line therapy (Supplementary Fig. 1). Only non-synonymous mutations are included in all of our analyses. Altogether, our dataset consists of 78,287 patients with 20 different types of cancer (summary statistics of patient characteristics are provided in Supplementary Data 1).

Association between mutations and patient survival

We examined the association of mutations within specific genes on overall survival (OS)—defined as the duration between cancer diagnosis and death—across 20 cancer types (Fig. 1). For patients with advanced or metastatic stages, the reference date was set at the time of this advanced diagnosis. To ensure sufficient statistical power, we conducted our analysis at the gene level. This means a gene was considered mutated if it contained one or more nonsynonymous genomic alterations. We employed the univariate Cox proportional hazards model to understand how certain mutations relate to OS, contrasting the survival rates between patients with and without a particular gene mutation. Adjustments were made using the Inverse Probability of Treatment Weighting (IPTW) method, accounting for potential confounding variables such as age, gender, race, performance status as per Eastern Cooperative Oncology Group (ECOG), time of diagnosis,

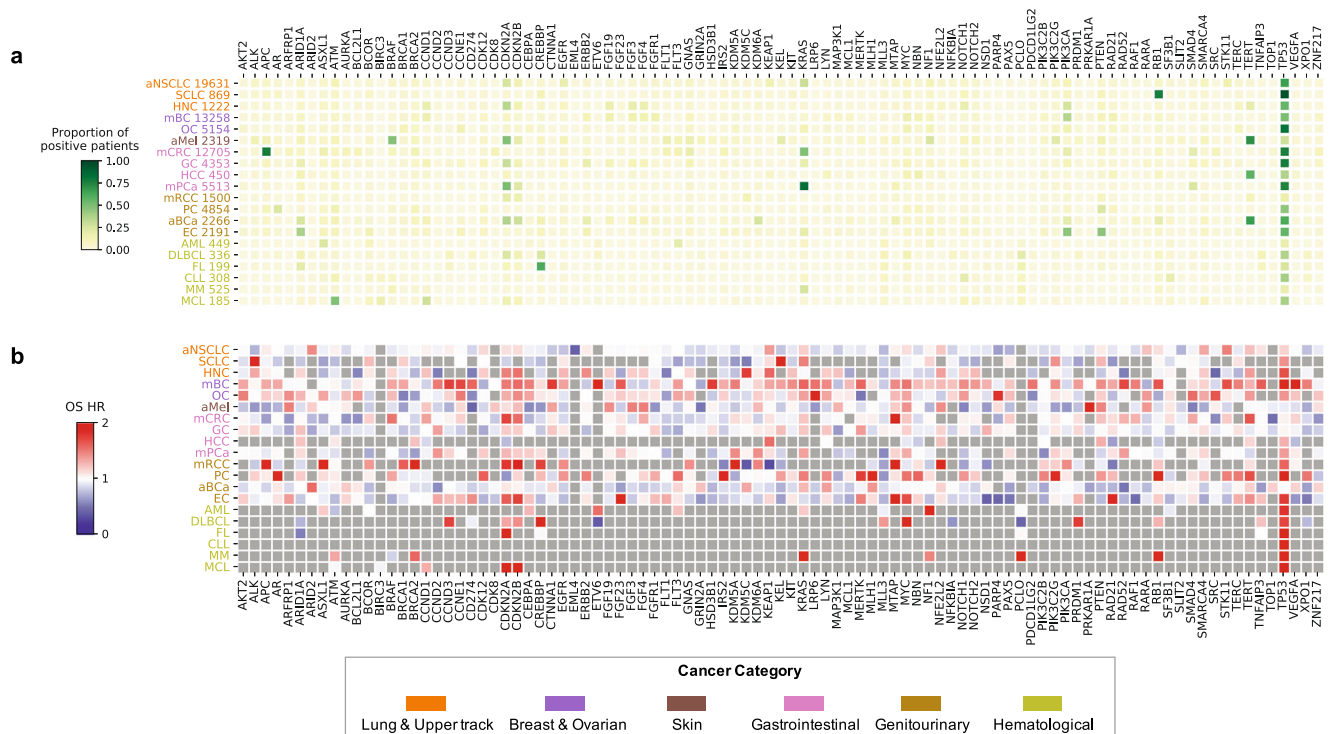


Fig. 1 | Overview of mutation statistics and mutation-survival associations.
a Proportion of patients in each cancer with a given gene mutated (normalized per row). **b** Prognostic effects of mutations in individual genes on overall survival (OS), measured by adjusted hazard ratio (HR) with two-sided Wald test. Genes that are significantly correlated with OS ($P < 0.05$) in a cancer type are colored red or blue. Red (blue) indicates that mutations in that gene have negative (positive) prognostic effects on the survival of patients. Genes that are not significantly associated with patient outcomes are shown as a gray square. Here we focus on 20 types of cancers including advanced non-small cell lung cancer (aNSCLC), metastatic breast cancer

(mBC), metastatic colorectal cancer (mCRC), metastatic pancreatic cancer (mPCa), ovarian cancer (OC), metastatic prostate cancer (PC), gastric cancer (GC), advanced melanoma (aMel), advanced bladder cancer (aBCa), endometrial carcinoma (EC), metastatic renal cell carcinoma (mRCC), head and neck cancer (HNC), small-cell lung cancer (SCLC), multiple myeloma (MM), acute myeloid leukemia (AML), hepatocellular carcinoma (HCC), diffuse large B cell lymphoma (DLBCL), chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL).

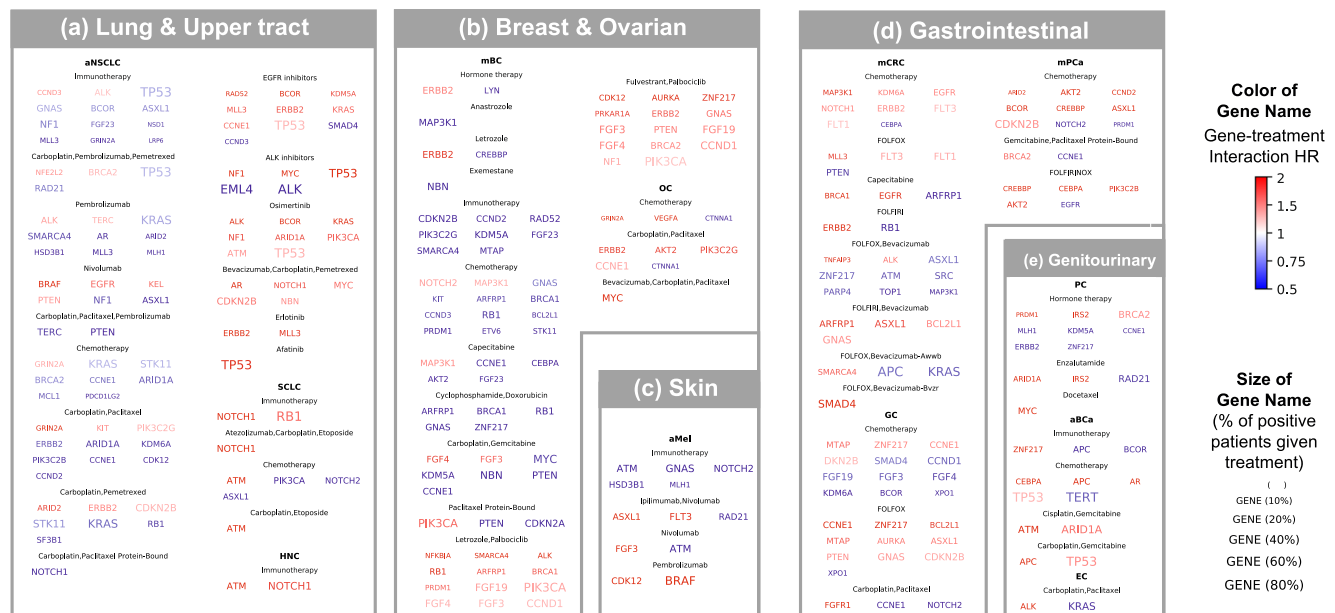


Fig. 2 | Gene-treatment interaction analysis across cancer types.

Gene-treatment interactions are shown for **(a)** lung and upper tract cancers, **b** breast and ovarian cancers, **c** skin cancers, **d** gastrointestinal cancers, **e**, genitourinary cancers. For each cancer type and first-line treatment, genes with significant interactions with that treatment are listed (two-sided Wald test P

value < 0.05 and overall FDR < 0.05). The font size of a gene indicates the fraction of patients in that treatment group with mutations in that gene. The color indicates the interaction hazard ratio (HR) of the gene. Red (blue) indicates that mutations in the gene have negative (positive) interactions and impact the survival of patients receiving a particular treatment.

tumor stage, and histological details. All hazard ratios (HRs) were derived using left-truncated Cox models, ensuring that potential biases, like premature death preventing a Foundation Medicine test, were minimized. Furthermore, to ensure the validity of multiple tests, we regulated the false discovery rate (FDR) following the Benjamini–Hochberg procedure.

We identified 95 genes that are significantly associated with survival in at least one of the cancer types under study (P value < 0.001 , FDR < 0.05). For example, mutations in *TP53*, *CDKN2A* and *CDKN2B* are associated with worse overall survival in patients across most types of cancer (Fig. 1), consistent with previous literature⁷⁻⁹. We repeated the analysis after excluding patients who underwent only FMI liquid biopsy (Supplementary Fig. 2) and the prognostic effects of genes are consistent with Fig. 1. We also reported the OS HR only for the mutations identified by FMI as likely pathogenic, and the findings remained consistent (Supplementary Fig. 3). When these findings were compared with a previous data release from September 30, 2021 (which included patients diagnosed and followed up through June 30, 2021), all the genes with significant prognostic effects were consistent across the board. The breakdown of mutation subtypes for each significant prognostic gene is provided in Supplementary Data 2.

Gene-treatment interactions

We investigated which tumor mutations may predict a differential response to specified treatments. By applying a Cox model with an interaction term between genes and treatments, we sought to pinpoint specific gene mutations that might affect treatment outcomes. This allowed us to distinguish between gene alterations that are specifically related to treatment response (i.e., predictive) and those that generally influence patient outcomes (i.e., prognostic). After adjusting for potential confounding variables (i.e., age, gender, race, performance status as per ECOG, time of treatment, tumor stage, and histological details), we identified 776 gene-treatment interactions associated with patient survival under specific treatments (P value < 0.05 , FDR < 0.05). A summary of these interactions, especially regarding key prognostic genes, is shown in Fig. 2 and Supplementary Data 3. We repeated the

analysis considering only known and likely pathogenic mutations, with tumor mutational burden (TMB) included as an additional confounding variable (Supplementary Data 4) and the predictive effects of genes remain consistent with those reported in Fig. 2 and Supplementary Data 3. It's worth noting that the interactions can vary depending on the drug combinations used, and the varying sample sizes for these combinations in our dataset can influence the reliability of our predictions.

Upon comparing our current findings with the previous study of a smaller, earlier cohort⁴ (Supplementary Fig. 4), all significant gene-treatment overlaps were consistent in their positive or negative associations with survival (either HR > 1 or HR < 1). For example, previously we observed that aNSCLC patients with *KRAS* mutations had a worse response to EGFR inhibitors compared to patients without *KRAS* mutations. EGFR inhibitors are primarily used for patients with EGFR mutations, though some in our cohort received them without confirmed mutations due to real-world practices. Similar findings had also been reported from other small cohorts¹⁰. In this updated analysis, this finding remained consistent with a statistically significant gene-by-treatment HR of 1.63 (95% CI 1.07-2.47). Specifically in aNSCLC patients with *KRAS* mutations, treatment with EGFR inhibitors was associated with non-statistically significant worse survival compared to other therapies (HR = 1.18, 95%CI 0.81-1.72), while in patients without *KRAS* mutations, EGFR inhibitor treatment was associated with statistically significant better survival compared to other therapies (HR = 0.78, 95% CI 0.69-0.88). Additionally, *KRAS* mutant aNSCLC patients also had better survival when receiving chemotherapy (gene-by-treatment HR = 0.86, 95%CI 0.76-0.97) or the immunotherapy pembrolizumab (gene-by-treatment HR = 0.83, 95%CI 0.71-0.97).

We also identified multiple statistically significant gene-by-treatment interactions for the *NF1* gene and various treatments. *NF1* is a tumor suppressor gene encoding neurofibromin that inhibits the RAS/MAPK and PI3K-AKT-mTOR signaling pathways, which play roles in regulating cell proliferation and apoptosis. NF1 loss of function mutations cause constitutive activation of RAS/MAPK and PI3K-AKT-mTOR pathways, leading to uncontrolled cell growth, cell survival and

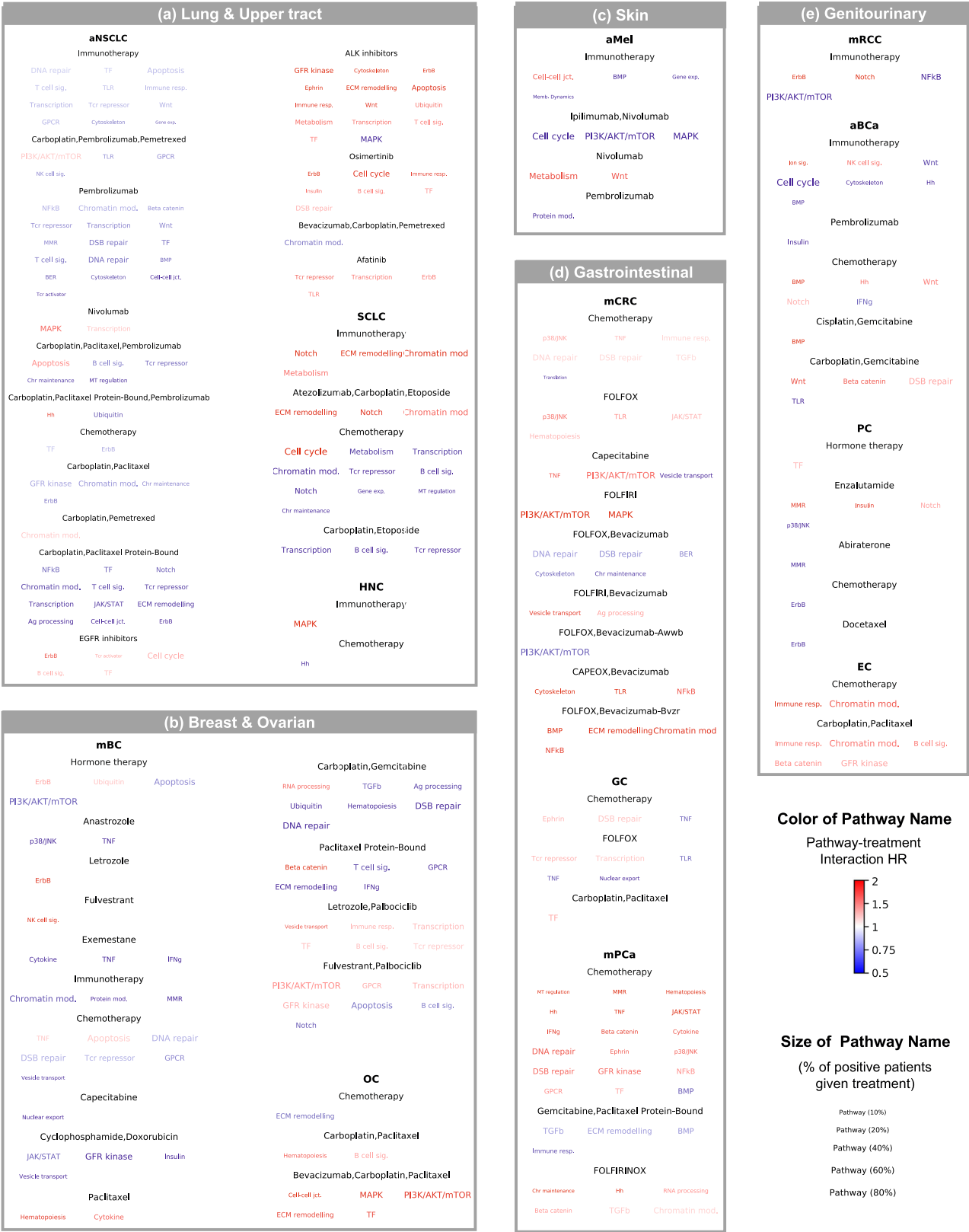


Fig. 3 | Pathway-treatment interaction analysis across cancer types. Pathway-treatment interactions are shown for **(a)** lung and upper tract cancers, **b** breast and ovarian cancers, **c** skin cancers, **(d)**. gastrointestinal cancers, **e**. genitourinary cancers. For each cancer type and first-line treatment, pathways with significant interactions with that treatment are listed (two-sided Wald test

P value < 0.05 and overall FDR < 0.05). The font size of a pathway indicates the fraction of patients in that treatment group with any gene mutation in that pathway. The color indicates the interaction hazard ratio (HR) of the pathway. Red (blue) indicates that mutations in the pathway have negative (positive) interactions and impact the survival of patients receiving a particular treatment.

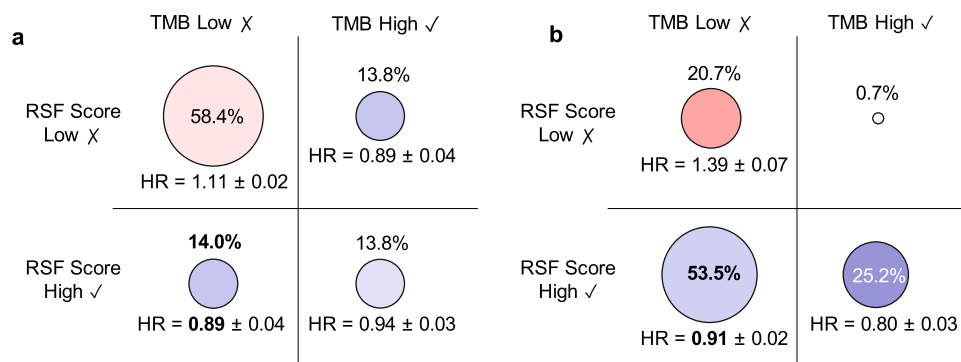


Fig. 4 | Comparison of immunotherapy and non-immunotherapy in aNSCLC patient groups. Immunotherapy versus non-immunotherapy is compared as (a) first-line and (b) second-line therapies in four aNSCLC patient groups consisting of patients with tumor mutation burden (TMB) low and Random Survival Forest (RSF) score low, patients with TMB low and RSF score high, patients with TMB high and

RSF score low, and patients with TMB high and RSF score high. The circle sizes represent the percentage of patients in each group. The circle colors represent the overall survival hazard ratio (HR) for patients who received immunotherapy vs. non-immunotherapy as a. first-line and b. second-line therapy. Cross-validation standard deviation is reported for each HR.

tumorigenesis^{11,12}. Neurofibromatosis type 1 is a hereditary disorder caused by germline alterations in *NFI*, where individuals are predisposed to developing tumors¹³. We found that in patients with aNSCLC, *NFI* mutations predict better survival with immunotherapy (gene-by-treatment HR = 0.76, 95%CI 0.65-0.88) and nivolumab treatment (gene-by-treatment HR = 0.68, 95%CI 0.47-0.97), and worse survival with targeted therapies of ALK inhibitors (gene-by-treatment HR = 3.51, 95%CI 1.80-6.85) or the EGFR inhibitor osimertinib (gene-by-treatment HR = 1.76, 95%CI 1.15-2.69). For metastatic breast cancer patients, *NFI* mutations were also associated with worse survival in patients receiving fulvestrant and palbociclib combination (ie. hormone and CDK4/6 inhibitor) based therapy (gene-by-treatment HR = 1.39, 95%CI 1.00-1.94). (Fig. 2, Supplementary Data 3)

These results are consistent with the discussion in a recent review paper of *NFI* alterations in cancer and therapies being investigated¹³. Cell line studies suggest that *NFI* alterations may be biomarkers for primary or acquired resistance to TKIs (e.g., EGFR inhibitors) and resistance to hormone therapies¹⁴⁻¹⁷. Previous work¹⁸ investigated the effectiveness of treatment with immune checkpoint inhibitors by *NFI* and TMB status in a cohort of 1661 patients with various cancer types, including 350 NSCLC patients. They observed better survival for patients with *NFI* mutations compared to those who were *NFI* wildtype.

Among mCRC patients receiving chemotherapy in our cohort, the presence of *ERBB2* (HR = 1.39, 95%CI 1.06-1.82) or *NOTCH1* (HR = 1.42, 95%CI 1.06-1.90) mutations was associated with worse survival when compared to not having the mutation. Both *ERBB2* and *NOTCH1* have, in recent years, become an increasing focus of precision medicine research in patients with mCRC¹⁹⁻²¹.

Among the new cancers studied in this paper, we found that in SCLC, patients having a *NOTCH1* mutation and receiving first-line immunotherapy were associated with worse survival based on a statistically significant gene-treatment interaction (HR = 2.3, 95%CI 1.28-4.13). Recent research suggests that NOTCH signaling may be a determinant for the benefit of checkpoint blockade immunotherapies in SCLC²². More research is needed to understand the role of NOTCH signaling and response to immunotherapy with larger sample sizes. Finally, in GC patients receiving any chemotherapy, having *CDKN2B* mutations was associated with worse survival (HR = 1.34, 95%CI 1.13-1.58) vs not having the mutation. Similarly, when receiving the chemotherapy combination FOLFOX, having the mutation was associated with worse survival (HR = 1.59, 95%CI 1.27-1.99) vs not having the mutation. Recent research demonstrates that overexpression of *CDKN2B* is correlated with poor prognosis in gastric cancer²³. All these

results support further research into targeted therapies for these genes to improve patient outcomes.

Pathway-treatment interactions

We extended our investigation to quantify how mutations in specific gene pathways affect the effectiveness of primary treatments. A mutation in a pathway is characterized by the presence of a non-synonymous alteration in any of the genes within the pathway (Supplementary Data 5). For each pathway and associated treatment, a Cox model was utilized with a corresponding interaction term; we control the false discovery rate to be less than 5%. A comprehensive overview of these interactions is provided in Fig. 3 and Supplementary Data 6. We repeated the analysis considering only pathways with known and likely pathogenic mutations, with TMB included as an additional confounding variable (Supplementary Data 7) and the predictive effects of pathways remain consistent with those reported in Fig. 3 and Supplementary Data 6. Reinforcing prior research²⁴, our findings support the association of the PI3K/AKT/mTOR pathway with worse outcomes for endocrine therapy in metastatic breast cancer (mBC), as shown by the HR value for its interaction with fulvestrant and palbociclib (pathway-by-treatment HR = 1.48, 95%CI 1.09-2.00). In addition, we found that the PI3K/AKT/mTOR pathway is a strong positive predictor of response to ipilimumab and nivolumab combination therapy for aMel (pathway-by-treatment HR = 0.41, 95%CI 0.24-0.70) and to immunotherapy for mRCC (pathway-by-treatment HR = 0.59, 95%CI 0.36-0.98), but a strong negative predictor of response to the chemotherapy regimen of FOLFIRI for mCRC (pathway-by-treatment HR = 3.08, 95%CI 1.14-8.32).

The double-strand DNA repair mechanism represents a critical component of the broader DNA repair pathway and is associated with positive response to immunotherapies in aNSCLC in our analysis (pathway-by-treatment HR = 0.90, 95%CI 0.82, 0.99). This correlation between increased DNA repair deficiency and improved survival has been previously documented²⁵. This phenomenon is often attributed to the resultant elevation in microsatellite instability (MSI), which subsequently facilitates an increased release of neoantigens. These neoantigens can be recognized by the immune system, thereby potentially contributing to the observed survival benefit.

Our analysis also highlighted the importance of the Toll-like Receptor (TLR) signaling pathway in treatment response. For example, we found that having mutations in any of the genes in the TLR pathway was associated with statistically significant better survival with immunotherapy (pathway-by-treatment HR of 0.87, 95%CI 0.77-0.99) and the specific carboplatin, pembrolizumab, pemetrexed combination

regimen (pathway-by-treatment HR = 0.79, 95%CI 0.67-0.93) in aNSCLC patients. The TLR pathway is a critical component of the innate immune system. TLRs are capable of recognizing pathogen-associated molecular patterns and damage-associated molecular patterns, leading to the activation of immune responses. Emerging evidence suggests that TLRs play a complex role in cancer, influencing inflammation, cell proliferation, survival, and the immune response to tumors. TLR signaling can shape the tumor microenvironment and potentially impact the development and progression of cancer²⁶. A number of studies are exploring the use of TLR agonists and inhibitors to coordinate anti-tumor immunity, either alone or in combination with other therapeutic modalities including immune checkpoint inhibitors²⁷.

Immunotherapy predictive modeling

The mutation-treatment analysis above identified a substantial number of mutations that are individually associated with response to immunotherapy in cancers like aNSCLC. We next investigated how well we could predict patient outcomes by combining all the mutations together using machine learning. We developed a Random Survival Forest (RSF) Score to predict the survival benefits of immunotherapy in patients with aNSCLC. The RSF model uses the patient's genomic profiles as features and we trained it to predict patient survival time under immunotherapy or other treatments (see "Methods" for details). Tumor mutation burden (TMB) is a common metric used to inform immunotherapy usage (following previous works^{28,29}, we binarized TMB using a threshold of 10 mutations/megabase). We included TMB status and mutation status for the 95 significant prognostic genes as the input features for the RSF model. In patients with low TMB and high RSF score (14% of all aNSCLC patients), immunotherapy was more effective compared to other treatment options (HR = 0.89 in first line and HR = 0.91 in second line, Fig. 4). This suggests that the RSF score could contain complementary information to TMB. To evaluate the representativeness of the learned RSF scores, we stratified our patient populations by insurance type (Supplementary Table 1) and repeated the analysis on patients with commercial health plans and those with non-commercial health plans (Supplementary Table 2). We also recomputed the RSF score incorporating patient baseline characteristics, mutation status and TMB status, and compared its performance with that of TMB alone in guiding immunotherapy usage (Supplementary Fig. 5). The findings of these analyses consistently aligned with our primary results. Survival curves for two groups of patients receiving immunotherapy—those with high RSF scores and low TMB, and those with high TMB and low RSF scores—are presented in Supplementary Fig. 6. The RSF model generates hypotheses for further investigation and should not be used to make clinical decisions.

Discussion

The integration of large-scale real-world clinico-genomics data is emerging as a powerful resource for precision medicine. Here we leverage the paired mutation profiling and follow-up information of 78,287 patients with 20 types of cancer to systematically identify biomarkers that predict patient response to specific cancer treatments.

Our research into gene-treatment interactions has revealed 776 significant interactions, thereby extending the knowledge base regarding how specific gene mutations can influence the efficacy of particular treatments. As the size of the real-world cohort increases, we are able to identify substantially more significant interactions (Supplementary Fig. 7), supporting the value of large-scale real-world clinical-genomics data. Moreover, the consistent overlap of significant gene-treatment interactions between our current and previous datasets adds robustness to these potential biomarkers, although caution is warranted due to varying sample sizes and the observational nature of the data.

Our comprehensive investigation into pathway-treatment interactions demonstrates that mutations within certain gene pathways can

substantially affect the effectiveness of cancer treatments. This underscores the multifaceted nature of cancer, where mutations in broader signaling pathways, rather than isolated genes, might play pivotal roles in determining treatment success or failure. It is essential to recognize that while specific gene mutations can provide valuable insights, understanding the broader gene pathways they belong to might offer a more holistic view of treatment dynamics. As we continue to push the boundaries of precision oncology, the assessment of pathway-treatment dynamics can serve as the tool to guide more informed therapeutic decisions and highlight potential avenues for drug development and optimization. Furthermore, we explored how combining machine learning with real-world data generates scores that can inform which patients respond to immunotherapies. Exploring such personalized prediction models is an important direction of future research.

There are several limitations to our study. Firstly, the retrospective nature of real-world data from EHRs can introduce selection biases. For example, patients who undergo genomic profiling may not be representative of the general cancer patient population. Secondly, while we accounted for known confounders, unrecorded variables in EHRs such as socio-economic status and detailed lifestyle factors could affect treatment outcomes and therefore remain a source of potential residual confounding. Furthermore, the pathway analysis, while informative, relies on the assumption that any mutation within a pathway uniformly impacts its overall function, which might oversimplify the intricate dynamics within the signaling cascades. Another potential limitation of our analyses and for our RSF score as a predictor of response to immunotherapy is that other multimodal predictors, such as features of the tumor microenvironment (TME), presence of tumor infiltrating lymphocytes (TIL), human leukocyte antigen (HLA) contexts, and transcriptome plus whole exome signatures among many others, that have been previously explored and reported^{30–33} were not available for our analyses. Lastly, the generalizability of the developed RSF score as a predictor of response to immunotherapy in aNSCLC may be limited by the heterogeneity of the clinical settings from which the data were drawn, and external validation in different populations is necessary to confirm our findings. Apart from TMB, microsatellite instability (MSI) is also a recognized biomarker of response to immunotherapy. However, our current cohort lacks sufficient MSI data to perform a separate analysis, which would be an interesting direction for future research. Additionally, while this study focuses on individual mutation-treatment interactions, future work could explore co-occurring or exclusive mutations to further contextualize these findings.

Overall, our large-scale systematic analyses and results are meant to be primarily hypothesis-generating and should not be directly used to suggest treatments. Our findings highlight the value of computational analysis of large-scale real-world clinico-genomics data to generate insights for precision oncology.

Methods

This study was conducted in compliance with all relevant ethical regulations. The study protocol was approved by the Institutional Review Board (IRB) overseeing the Flatiron Health-Foundation Medicine Clinico-Genomic Database (FH-FMI CGDB), with a waiver of informed consent granted due to the de-identified nature of the data. The present research uses secondary data from Flatiron Health that is de-identified and thus is not human subjects research and does not require IRB review or approval.

Flatiron Health-Foundation Medicine Clinico-genomic Database (FH-FMI CGDB)

This study used the nationwide (US-based) multi-tumor de-identified database known as FH-FMI CGDB, comprising data from around 280 US cancer clinics (over 800 sites of care). The dataset includes

retrospective longitudinal clinical information from electronic health records (EHRs), containing both structured and unstructured patient-level data. This clinical data was linked to genomic data obtained through Foundation Medicine comprehensive genomic profiling (CGP) tests within the FH-FMI CGDB, using a secure, de-identified deterministic matching process facilitated by an independent third party³⁴. Strict privacy measures were in place to safeguard patient confidentiality, including Institutional Review Board approval and waiver of informed consent. The data was released on March 31, 2023, and encompassed patients diagnosed and followed up until Dec 31, 2022, and was subject to obligations to prevent re-identification and protect patient confidentiality.

Genomic profiling

Genomic alterations were identified through comprehensive genomic profiling (CGP) of cancer-related genes using Foundation Medicine's Next-Generation Sequencing (NGS) tests^{35–37}. Pathogenic variants, both known and likely, were identified using a tailored filtering approach³⁸. Germline variant predictions were excluded through cross-referencing with publicly available databases, including the 1000 Genome Project and the Exome Aggregation Consortium (ExAC), while retaining established cancer-driving events. Additionally, likely germline variants of uncertain significance were excluded using an algorithm distinguishing somatic from germline variants without matched normal tissue³⁹. For further validation, pathogenic variants were corroborated, and truncations and deletions in tumor suppressor genes were categorized as likely pathogenic.

Patient Selection. Patients in the FH-FMI CGDB met specific criteria: they had at least two documented clinical visits in the Flatiron Health network, underwent Foundation Medicine testing with pathologist-confirmed histology matching Flatiron Health's abstracted tumor type, and had their Foundation Medicine specimen collection date occur no earlier than 30 days after their initial diagnosis date in Flatiron Health.

Flatiron health data in FH-FMI CGDB

The FH-FMI CGDB combines data from EHRs, external commercial sources, and the US Social Security Death Index. It harmonizes structured data, like laboratory test results, across different EHR systems and maps them into standardized terminologies. Unstructured data, such as clinician notes, undergoes technology-enabled abstraction, with qualified abstractors extracting critical information aided by specialized software. Lines of therapy were rule-based and defined by expert oncology clinicians. These efforts ensure data accuracy and consistency across the roughly 280 cancer clinics represented in the database, predominantly from community oncology settings.

Participant Information. This study included 78,287 cancer patients from FH-FMI CGDB, with a diverse demographic composition across age, gender, and race as reported in Supplementary Data 1. Gender and race information was based on self-reported data and incorporated as demographic factors in the dataset. No participant compensation was involved, as this study analyzed existing de-identified, retrospective data.

Informed Consent. Informed consent was not obtained from individual participants, given the de-identified nature of the FH-FMI CGDB data. IRB approval of the study protocol was obtained prior to study conduct, which granted a waiver of informed consent based on the retrospective design and anonymized data structure.

Gender Considerations. Gender information was collected as part of the demographic data. However, the primary focus of the study was to characterize mutation-treatment interactions across cancer types, irrespective of gender, as these variables were not central to the study aims. Consequently, no disaggregated gender-based analyses were conducted. This approach aligns with the study's objective of

identifying clinically significant genomic markers across a heterogeneous patient population.

Survival analysis

Survival analysis was conducted using the Cox proportional hazards model. The index date, marking the beginning of analysis, was set as the patient's diagnosis date for overall survival (OS) analysis and the start date of the first-line treatment for treatment efficacy analysis. Patients with at least one visit in the FH database after their first CGP test report date were included for analysis. Patients were followed until death for OS analysis. Gene-treatment interactions and pathway-treatment interaction analyses were focused on OS outcomes. Censoring dates were selected as the last clinic visit for patients without death. A sensitivity analysis was also conducted that limited the survival analysis to patients whose advanced or metastatic diagnosis occurred on or prior to September 30, 2022, to allow for at least 3 months of potential follow-up for all patients (ie. up until the December 31, 2022 data-cut off date) and reduce the number of less informative patients who were diagnosed near the cut-off date with very short follow-up times (Supplementary Fig. 8). To facilitate comparisons between recently approved and older treatments, a maximum follow-up time was set, which is a standard practice in real-world data analyses. Outcomes occurring after 48 months for mBC and 27 months for other cancer types were also considered censored. Robust results were obtained with the removal of upper follow-up bounds.

Challenges in clinicogenomic data

A challenge in clinicogenomic data analysis is the limited number of patients who undergo tumor genomics tests immediately after diagnosis. Consequently, patients with shorter survival times may not have genomic data recorded, introducing a potential bias known as immortal time bias. To mitigate this bias, we adjusted the Cox proportional hazards model with left-truncation on dates corresponding to patients' genomic profiling tests.

Prognostic effect analysis

We assessed the prognostic significance of individual genes and drugs using a univariate Cox proportional hazards model. We applied Inverse Probability of Treatment Weighting (IPTW) to account for baseline factors that might confound the analysis. In this survival analysis, each patient's weight (ω_i) in the Eq. (1) is determined by their propensity score (e_i) and whether they belong to the experimental group or control group (indicated by Z_i). For instance, when studying the impact of a gene mutation on overall survival (OS), patients with the mutated gene are in the experimental group ($Z_i = 1$), while those without the mutation form the control group ($Z_i = 0$). The propensity score (e_i) is estimated using a logistic regression model in Eq. (2), where $X_{i,j}$ represents the j -th baseline patient characteristic. After adjusting by propensity score, the baseline characteristics are effectively balanced between the experimental and control groups.

$$\omega_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i} \quad (1)$$

$$e_i = \Pr(Z_i = 1 | X_i) \quad (2)$$

The baseline confounders (X) included age, gender, race, ECOG status, cancer staging, histology (aNSCLC, OC, and mRCC), smoking status (for aNSCLC), and the year of the index date. The ECOG status closest to the index date within a range of −180 to +7 days was selected. All categorical variables with at least 5% of patients were one-hot encoded. Missing data for other covariates were treated as a separate category and effectively balanced between the experimental and control groups after IPTW adjustments.

Gene-treatment interaction analysis

We determined predictive associations between gene mutations and treatments if a significant gene-by-treatment interaction ($P < 0.05$) was observed. The interactive Cox proportional hazards model for a single gene mutation and treatment included gene mutation status (G), treatment status (T), a gene-by-treatment interaction term ($G \times T$), and the baseline confounders (X) as defined in Eq. (3):

$$h(t) = h_0(t) \exp(\beta_G G + \beta_T T + \beta_{G \times T} G \times T + \sum_{i=1}^{nc} \beta_{X_i} X_i) \quad (3)$$

In this equation, $h(t)$ represents the hazard function at time t , and $h_0(t)$ is the baseline hazard. The HRs (Hazard Ratios) for gene mutation ($\exp(\beta_G G)$), treatment ($\exp(\beta_T T)$), and their interaction ($\exp(\beta_{G \times T} G \times T)$) are obtained from the antilog of the respective regression coefficients. A gene-by-treatment interaction HR ($\exp(\beta_{G \times T})$) greater than 1 indicates that the treatment's effect is more adverse in patients with tumors harboring a mutation in that gene compared to patients without that genetic mutation. The same set of confounders (X) used in the prognostic effect analysis was applied here. Additionally, we explored interactions between genes and categories of treatment, such as immunotherapies.

In our study, the term 'chemotherapy' refers exclusively to chemotherapy-only treatments. 'Immunotherapy', 'hormone therapy', and 'targeted therapy' refer to treatments that include immunotherapy agents, hormone agents, and targeted agents, respectively.

Pathway annotation/classification

Pathway annotations for genes baited by past and current tissue- and liquid-based Foundation Medicine assays (FoundationOne®, FoundationOne®CDx, FoundationOne®Heme, FoundationACT®, FoundationOne®Liquid, and FoundationOne®Liquid CDx) were curated based on data from GeneCards® and the literature⁴⁰ (Supplementary Data 5). Pathways most relevant to cancer-related biological processes were selected. One gene may be assigned to multiple pathways. Genes for which pathways were missing or for which there was little consensus in GeneCards® were subject to a literature search to identify the most relevant pathways in cancer.

Pathway-treatment interaction analysis

We next examined predictive associations between gene mutations and treatments at the pathway level. A pathway was considered mutated if any gene within that pathway was mutated. We determined predictive associations if a significant ($P < 0.05$) pathway-by-treatment interaction was observed. For the pathway-treatment interaction analysis, the same Cox proportional hazards model and analytical method described for the gene-treatment interaction analysis was applied, except that the gene mutation status (G) variable was replaced by the pathway mutation status (P) variable in the model

$$h(t) = h_0(t) \exp\left(\beta_P P + \beta_T T + \beta_{P \times T} P \times T + \sum_{i=1}^{nc} \beta_{X_i} X_i\right) \quad (4)$$

RSF score for immunotherapy

We propose the RSF score as a quantitative indicator of the potential impact of immunotherapy on patient survival based on the genomic profile. We developed the Random Survival Forest (RSF) algorithm to predict patient survival time from the patient genomics profile. The input features for the RSF model are the mutation status for the 95 significant prognostic genes and TMB status. The model has 20 trees, 10 minimum number of samples required to split an internal node, and 15 minimum number of samples required to be at a leaf node. We implement cross-validation by dividing the data randomly

into a training set comprising 70% of the data and a test set with the remaining 30%. This process is repeated for 10 iterations, and the outcomes are averaged to ensure robustness. Within the RSF framework, we made two predictions: 1) RSF was used to predict patient survival under immunotherapy, shedding light on the potential benefits of this treatment, and 2) RSF predicted patient survival under an alternative scenario, treatment without immunotherapy, providing a baseline against which to compare immunotherapy. The RSF score is 1 (i.e., high) if the predicted survival time with immunotherapy is longer than the predicted survival time without immunotherapy for each patient, and 0 (i.e., low) if vice versa.

To comprehensively assess the predictive power of the RSF score, we compared it with an established biomarker: Tumor Mutation Burden (TMB). TMB, measuring genetic mutations in tumors, was evaluated for its ability to predict patient responses to immunotherapy. Following thresholds used in previous works^{28,29}, we binarized TMB using a cutoff of 10 mutations/megabase for analysis. This comparative analysis aimed to identify the most robust predictive factors, offering realistic insights into the potential benefits of immunotherapy for individual patients. We estimated survival curves using the Kaplan-Meier method, adjusted for baseline patient characteristics. We aggregated the test data from the 10 iterations. IPTW is used to adjust for the same set of baseline confounders as in the prognostic analysis. Our comparison specifically involves two patient groups on immunotherapy: those with high RSF scores and low TMB versus those with high TMB and low RSF scores.

To evaluate the combined impact of clinical and genomic data in predicting immunotherapy outcomes, we recompute the RSF score by incorporating patient baseline characteristics, the mutation status for the 95 significant prognostic genes and TMB status as input feature for the RSF model. We use the same set of baseline patient characteristics as used in our prognostic effect analysis. We then repeat the comparative analysis of this updated RSF score against TMB alone.

To assess the impact of different features on the predictive power of RSF, we compute the permutation feature importance for the RSF in predicting immunotherapy outcomes in a single iteration. The importance of the top 30 features is shown in Supplementary Fig. 9. To calculate the permutation feature importance, we first determine a baseline using the concordance index evaluated on the dataset. We then permute a feature and re-evaluate the concordance index. The permutation importance of a feature is defined as the difference between the baseline concordance index and the index after permuting the feature column. Each feature is permuted 1000 times to ensure statistical reliability.

Statistics & Reproducibility. No statistical method was used to predetermine sample size; sample size was determined by the availability of data within the FH-FMI CGDB. Analyses were conducted using established statistical methods, including Cox proportional hazards models for survival analysis and IPTW to adjust for confounding variables such as age, gender, race, and cancer stage. No additional data were excluded from the analyses. The experiments were not randomized, as this was an observational study using retrospective, real-world data. The investigators were not blinded to allocation during experiments or outcome assessment, as the study design did not involve intervention allocation but rather analysis of existing clinical-genomic data. The reproducibility of findings was supported by the use of comprehensive, de-identified real-world data and rigorous validation steps, including comparisons with prior datasets to confirm the consistency of gene-treatment interaction effects.

Reporting summary

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

Data availability

The data that support the findings of this study originated by Flatiron Health, Inc. and Foundation Medicine, Inc. The Flatiron Health dataset is available under restricted access due to patient privacy. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to PublicationsDataAccess@flatiron.com and cgdb-fmi@flatiron.com.

Code availability

The code used in this paper is available on Github at <https://github.com/RuishanLiu/precision-cancer2023> and has been archived on Zenodo with <https://doi.org/10.5281/zenodo.14015946>⁴¹.

References

- Hodson, R. Precision medicine. *Nature* **537**, S49 (2016).
- Morash, M., Mitchell, H., Beltran, H., Elemento, O. & Pathak, J. The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *J. Pers. Med.* **8**, E30 (2018).
- Garraway, L. A., Verweij, J. & Ballman, K. V. Precision oncology: an overview. *J. Clin. Oncol.* **31**, 1803–1805 (2013).
- Liu, R. et al. Systematic pan-cancer analysis of mutation–treatment interactions using large real-world clinico-genomics data. *Nat. Med.* **28**, 1656–1661 (2022).
- Liu, R. et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).
- Booth, C. M., Karim, S. & Mackillop, W. J. Real-world data: towards achieving the achievable in cancer care. *Nat. Rev. Clin. Oncol.* **16**, 312–325 (2019).
- Petitjean, A., Achatz, M. I. W., Borresen-Dale, A. L., Hainaut, P. & Olivier, M. TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* **26**, 2157–2165 (2007).
- Zhao, R., Choi, B. Y., Lee, M.-H., Bode, A. M. & Dong, Z. Implications of genetic and epigenetic alterations of CDKN2A (p16(INK4a)) in cancer. *EBioMedicine* **8**, 30–39 (2016).
- Bonelli, P., Tuccillo, F. M., Borrelli, A., Schiattarella, A. & Buonaguro, F. M. CDK/CCN and CDKI alterations for cancer prognosis and therapeutic predictivity. *BioMed. Res. Int.* **2014**, 361020 (2014).
- Raponi, M., Winkler, H. & Dracopoli, N. C. KRAS mutations predict response to EGFR inhibitors. *Curr. Opin. Pharmacol.* **8**, 413–418 (2008).
- Philpott, C., Tovell, H., Frayling, I. M., Cooper, D. N. & Upadhyaya, M. The NF1 somatic mutational landscape in sporadic human cancers. *Hum. Genomics* **11**, 13 (2017).
- McCubrey, J. A. et al. Mutations and deregulation of Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascades which alter therapy response. *Oncotarget* **3**, 954–987 (2012).
- Giraud, J.-S., Bièche, I., Pasmant, É. & Tlemsani, C. NF1 alterations in cancers: therapeutic implications in precision medicine. *Expert Opin. Investig. Drugs* **32**, 941–957 (2023).
- de Bruin, E. C. et al. Reduced NF1 expression confers resistance to EGFR inhibition in lung cancer. *Cancer Discov.* **4**, 606–619 (2014).
- Tao, J., Sun, D., Dong, L., Zhu, H. & Hou, H. Advancement in research and therapy of NF1 mutant malignant tumors. *Cancer Cell Int.* **20**, 492 (2020).
- Pearson, A. et al. Inactivating NF1 mutations are enriched in advanced breast cancer and contribute to endocrine therapy resistance. *Clin. Cancer Res.* **26**, 608–622 (2020).
- Sokol, E. S. et al. Loss of function of NF1 is a mechanism of acquired resistance to endocrine therapy in lobular breast cancer. *Ann. Oncol.* **30**, 115–123 (2019).
- Li, X., Sun, J. & Wang, L. NF1-mutant cancer and immune checkpoint inhibitors: a large database analysis. *Clin. Lung Cancer* **22**, 480–481 (2021).
- Torres-Jiménez, J., Esteban-Villarrubia, J. & Ferreiro-Monteagudo, R. Precision medicine in metastatic colorectal cancer: targeting ERBB2 (HER-2) oncogene. *Cancers* **14**, 3718 (2022).
- Tyagi, A., Sharma, A. K. & Damodaran, C. A review on notch signaling and colorectal cancer. *Cells* **9**, 1549 (2020).
- Yu, W., Wang, Y. & Guo, P. Notch signaling pathway dampens tumor-infiltrating CD8+ T cells activity in patients with colorectal carcinoma. *Biomed. Pharmacother.* **97**, 535–542 (2018).
- Roper, N. et al. Notch signaling and efficacy of PD-1/PD-L1 blockade in relapsed small cell lung cancer. *Nat. Commun.* **12**, 3880 (2021).
- Chen, X., Yu, X. & Shen, E. Overexpression of CDKN2B is involved in poor gastric cancer prognosis. *J. Cell. Biochem.* **120**, 19825–19831 (2019).
- Araki, K. & Miyoshi, Y. Mechanism of resistance to endocrine therapy in breast cancer: the important role of PI3K/Akt/mTOR in estrogen receptor-positive, HER2-negative breast cancer. *Breast Cancer Tokyo Jpn.* **25**, 392–401 (2018).
- Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
- Huang, B., Zhao, J., Unkeless, J. C., Feng, Z. H. & Xiong, H. TLR signaling by tumor and immune cells: a double-edged sword. *Oncogene* **27**, 218–224 (2008).
- Zheng, R. & Ma, J. Immunotherapeutic implications of toll-like receptors activation in tumor microenvironment. *Pharmaceutics* **14**, 2285 (2022).
- Hellmann, M. D. et al. Nivolumab plus Ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
- Allgäuer, M. et al. Implementing tumor mutational burden (TMB) analysis in routine diagnostics—a primer for molecular pathologists and clinicians. *Transl. Lung Cancer Res.* **7**, 703–715 (2018).
- Bai, R., Lv, Z., Xu, D. & Cui, J. Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. *Biomark. Res.* **8**, 34 (2020).
- Roelofsen, L. M., Kaptein, P. & Thommen, D. S. Multimodal predictors for precision immunotherapy. *Immuno-Oncol. Technol.* **14**, 100071 (2022).
- Bagae, A. et al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell* **39**, 845–865.e7 (2021).
- Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).
- Singal, G. et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinico-genomic database. *JAMA* **321**, 1391–1399 (2019).
- Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
- He, J. et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood* **127**, 3004–3014 (2016).
- Woodhouse, R. et al. Clinical and analytical validation of FoundationOne Liquid CDx, a novel 324-Gene cfDNA-based comprehensive genomic profiling assay for cancers of solid tumor origin. *PLoS One* **15**, e0237802 (2020).
- Hartmaier, R. J. et al. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. *Cancer Res.* **77**, 2464–2475 (2017).
- Sun, J. X. et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Computational Biol.* **14**, e1005965 (2018).

40. Stelzer, G. et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1–1.30.33 (2016).
41. Liu, R. et al. Characterizing mutation-treatment effects using clinico-genomics data of 78,287 patients with 20 types of cancers. GitHub repository. <https://doi.org/10.5281/zenodo.14015946> (2024).

Acknowledgements

We thank Craig Cummings, Stephanie Hilz, Matthew Wongchenko, Minu Srivastava, Barzin Nabet for providing valuable comments, discussion, and feedback at various stages of the manuscript.

Author contributions

R.L., S.R., L.W., N.C., S.M., M.R.G., S.M., R.C., and J.Z. designed the project. R.L. developed methodology and performed the analysis with support from all the authors. S.R., L.W., N.C., S.M., M.R.G., S.M. provided clinical interpretations. All the authors contributed to writing the paper. R.C. and J.Z. supervised the project.

Competing interests

S.R., L.W., N.C., S.M., M.R.G., S.M., and R.C. are employees of F. Hoffmann-La Roche Ltd or of Genentech, Inc., and are shareholders of Roche. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-55251-5>.

Correspondence and requests for materials should be addressed to Ryan Copping or James Zou.

Peer review information *Nature Communications* thanks the other anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024