nature communications

Article

https://doi.org/10.1038/s41467-025-59397-8

DiffInvex identifies evolutionary shifts in driver gene repertoires during tumorigenesis and chemotherapy

Received: 7 August 2024

Accepted: 22 April 2025

Published online: 13 May 2025

Check for updates

Ahmed Khalil[1] & Fran Supek[1,2,3]

Somatic cells can transform into tumors due to mutations, and the tumors further evolve towards increased aggressiveness and therapy resistance. We develop DiffInvex, a framework for identifying changes in selection acting on individual genes in somatic genomes, drawing on an empirical mutation rate baseline derived from non-coding DNA that accounts for shifts in neutral mutagenesis during cancer evolution. We apply DiffInvex to >11,000 somatic whole-genome sequences from ~30 cancer types or healthy tissues, identifying genes where point mutations are under conditional positive or negative selection during exposure to specific chemotherapeutics, suggesting drug resistance mechanisms occurring via point mutation. DiffInvex identifies 11 genes exhibiting treatment-associated selection for different classes of chemotherapies, linking selected mutations in *PIK3CA*, *APC*, *MAP2K4, SMAD4*, *STK11* and *MAP3K1* with drug exposure. Various gene-chemotherapy associations are further supported by differential functional impact of mutations pre- versus post-therapy, and are also replicated in independent studies. In addition to nominating drug resistance genes, we contrast the genomes of healthy versus cancerous cells of matched human tissues. We identify noncancerous expansion-specific drivers, including *NOTCH1* and *ARID1A*. DiffInvex can also be applied to diverse analyses in cancer evolution to identify changes in driver gene repertoires across time or space.

The somatic genome accumulates driver mutations, evolving from healthy cells to early-stage tumors, and further towards late-stage malignancies, which progress towards higher aggressiveness. One important aspect of the latter is gaining the ability to resist chemotherapy, a standard treatment for cancer, where multiple drugs are typically used in combination, supplemented with immunotherapy or radiotherapy to kill rapidly dividing cancer cells. However, the surviving cells typically give rise to resistant tumors rather than being an exception. This poses a significant challenge in cancer treatment: tumors often show good initial response to chemotherapy, but later relapse[1,2], indicating that some genetic or epigenetic features have evolved to grant tumor cells resistance. Identifying these changes and understanding the underlying mechanisms is the first step toward anticipating and countering resistance to chemotherapy.

A genomic analysis comparing mutations present in pre-treated tumors with those in treatment-naive tumor genomes should identify genes associated with the emergence of resistance. Over the last decade, large-scale tumor profiling studies, such as Hartwig Medical Foundation (henceforth: Hartwig)[3], PCAWG[4], POG[5], DECIDER[6], GENIE[7], and FH-FMI CGDB[8], have provided genomic mutational landscapes of thousands of patients before and/or after treatment

[1] Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain. [2] Biotech Research and Innovation Centre (BRIC), Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark. [3] Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain. e-mail: fran.supek@bric.ku.dk

with various combinations of drugs. These genomic profiles, along with the clinical information, enable studies of tumor evolution during treatment and the identification of genetic elements under increased selection upon treatment, that presumably underlie patients' response to different chemotherapy drugs. Recently, evolutionary approaches that can assess conditional (differential) selection in cancer genomes have been developed and applied to the study of epistasis between driver mutations[9,10], as well as to assess how selection varies between tissues, and which mutational process generates particular driver mutations[11,12].

A major challenge in such comparative analysis is that the background mutation rates and signatures are altered due to chemotherapy exposure itself, as well as other biological and technical factors that may differ between the treated and untreated cohorts. This complicates identifying causally selected mutations that drive chemotherapy resistance, against the shifting backdrop of passenger changes resulting from changing mutagenic exposures during tumor evolution. Given the increasing availability of WGS data, mutations in introns and other non-coding regions can serve for tracking local neutral mutation rate in gene coding regions[13]. This provides an empirical baseline that obviates the need to infer mutation risk from covariate information, such as replication timing or gene expression, often used as proxies for mutation rates[14,15]. The empirical baseline allows for the simultaneous assessment of the contribution of known and unknown factors to the mutation rates and spectra during the course of tumor evolution. For example, the individual treatments in combination therapy can have different, in principle separable effects on the background mutagenic processes and also on the selective pressures acting on individual genes.

Motivated by the above, in this study, we have developed a statistical framework, named DiffInvex, to quantify conditional selection in cancer by comparing selective pressures upon mutations in genes between two or more conditions. While DiffInvex is broadly applicable to diverse comparative analyses of cancer genomes, here we apply it to identify genes that increase or decrease selection pressures during transformation from healthy to cancerous cells, and further during chemotherapy treatment. The DiffInvex empirical local mutation rate baseline is more accurate than a covariate-based approach, and effectively adjusts for confounding by changes in mutation burdens and spectra during tumor evolution. We apply DiffInvex to a large set of over 11,000 somatic whole-genomes of ~30 tissue types, assessing changes in selective pressures on point mutations. DiffInvex mapped the cancer type distribution of the known drug-resistance mutations, and we further identify 11 drug-gene associations with altered positive selection upon exposure to specific classes of therapeutics. The hits are validated using a test for differential mutation functional impact, and some are also replicated in independent cohorts. These genes are a subset of common driver genes, suggesting tumors typically gain drug resistance via general increases in cell fitness arising from additional driver changes, rather than via specialized drug resistance mechanisms. As another application of DiffInvex, we model early tumor evolution, assessing changes in selection upon driver genes between healthy cells versus tumors of matched tissues. We identify genes that are under positive selection in normal cells similarly as in cancer cells, suggesting their status as cancer driver genes merits revisiting.

Results

Comparing 8591 whole-genome sequenced tumors from pre-treated versus treatment-naive patients

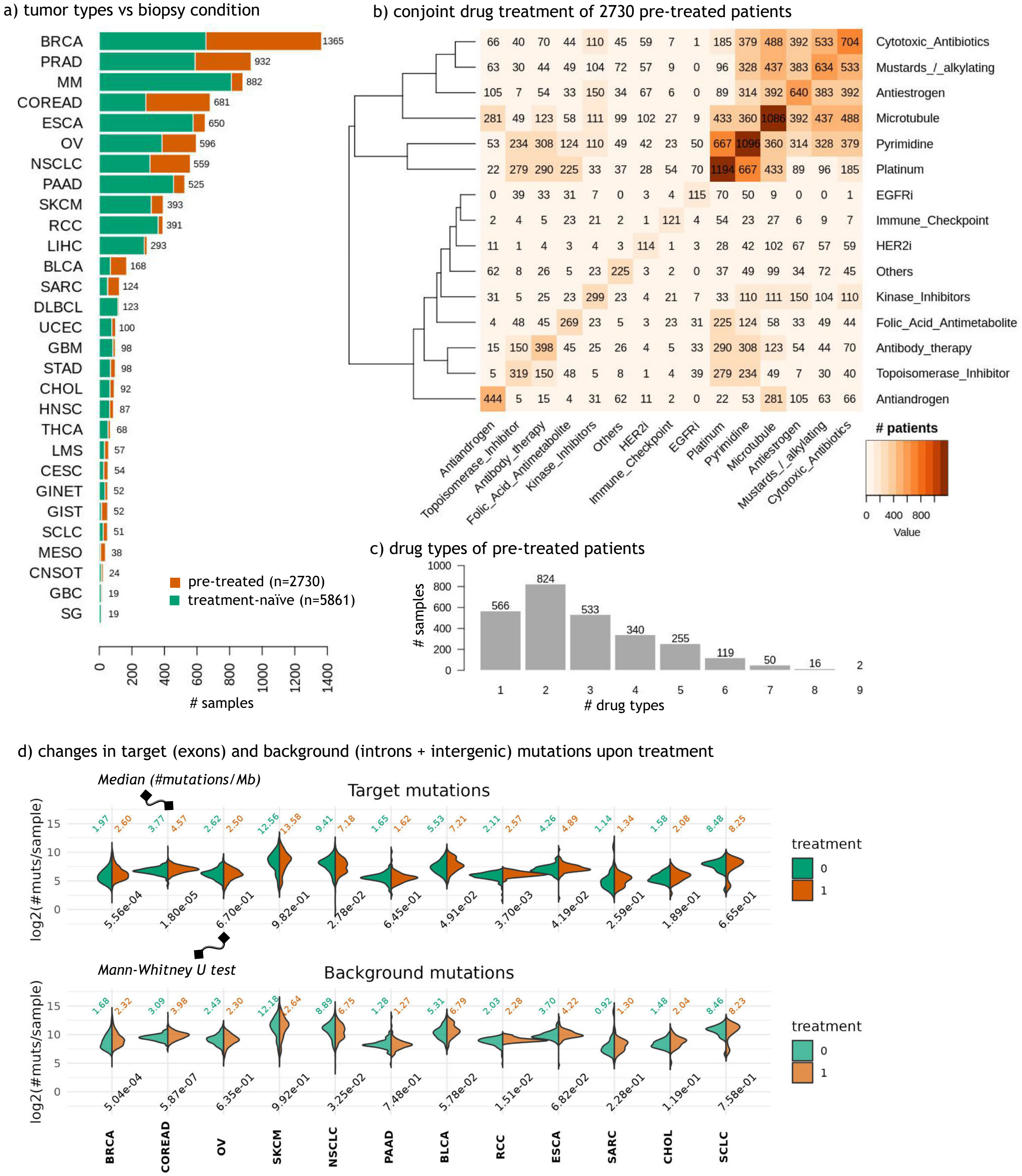
To identify the genes under differential positive selection associated with chemotherapy treatment, we collected the somatic single-nucleotide variant (SNV) mutations from WGS of 9953 tumor samples. This data set was collated from the cohorts with pre-treated samples of Hartwig Medical Foundation (henceforth "Hartwig"), POG570, DECIDER, and MMRF-COMMPASS (henceforth: MMRF), and the largely treatment-naive WGS from the PCAWG, CPTAC-3, Mutographs, ICGC (additional samples other than PCAWG), and OVCARE cohorts. After excluding samples with low-quality data (e.g. tumor purity <20%, PCAWG-blacklisted) and MSI samples, or with missing treatment information, and tumor types with less than 5 treated or less than 5 treatment-naive patients, a total of 8591 tumor samples (3360 Hartwig, 1865 PCAWG, and 880 MMRF, 778 ICGC, 541 Mutographs, 510 POG570, 322 CPTAC-3, 202 DECIDER, 133 OVCARE) from 29 tumor types were used in the analyses (see Methods, Supplementary Fig. 1, Supplementary Data 1). This includes 4138 (48%) primary tumor samples and 4453 (52%) metastatic tumor samples (Supplementary Fig. 1b). Among them, 2730 (32%) samples were obtained from biopsy pre-treated with chemotherapy drug(s) and/or other drug types, and 5861 (68%) samples were treatment-naive (Supplementary Fig. 1a). Pre-treated samples are from Hartwig (2167 samples), POG570 (424 samples), DECIDER (70 samples) and MMRF (69 samples). Eight out of 29 tumor types have more than 500 samples each (Fig. 1a): breast cancer (BRCA, 1365), prostate cancer (PRAD, 932), multiple myeloma (MM, 882 samples), colorectal cancer (COREAD, 681), esophageal cancer (ESCA, 650), ovarian cancer (OV, 596), non-small cell lung cancer (NSCLC, 559) and pancreatic cancer (PAAD, 525). The Hartwig, PCAWG, and POG570 studies included samples from diverse tumor types (>= 19), while other studies focused on few tumor types (Supplementary Fig. 1c), four tumor types from CPTAC-3, two tumor types from MUTOGRAPHS and ICGC each, and one tumor type from each of DECIDER, OVCARE, and MMRF.

Statistics of drug treatments across cancer types show common overlaps in drug regimens

We then examined and classified the drugs that were given to the 2730 patients prior to the biopsy, based on their drugs' mechanism-of-action category (https://go.drugbank.com/), dividing them into 15 drug type groups (Fig. 1b, Supplementary Data 2). These included 6 DNA damaging chemotherapy groups (platinum-based alkylating agents, pyrimidine analogs, cytotoxic antibiotics, nitrogen mustards and other alkylating, topoisomerase inhibitor, and folic acid antimetabolites) and 9 other chemotherapy drug groups (microtubule agents, antiestrogens, antiandrogens, antibody therapy, kinase inhibitors, immune checkpoint inhibitors, EGFRi, HER2i, and other drugs). The "other drugs" is a mixed group of drugs that we could not classify to one of the other 14 drug types.

We observed a wide diversity in the drugs administered to patients in the analyzed data, covering 138 different drugs in total (Supplementary Fig. 1d, e), with the most common being platinum-based alkylating agents, pyrimidine analogs, and microtubule drugs (> 1000 patients each). Most drug types were used across many tumor types: 12 drug types were given to ≥ 5 patients from more than 5 tumor types, while only 3 drug types (antiandrogens, HER2i and EGFRi) were mainly given to one or two tumor types (Supplementary Fig. 1d). This suggests that performing a pan-cancer study would help to boost the power for statistical tests of association between drug exposure and driver mutation.

Most patients were treated with drug combinations of different drug types, while only 548/2730 patients were treated with a single drug type (Fig. 1c). For example, antiestrogen drugs are jointly used with diverse chemotherapy drug groups such as pyrimidine analogs, microtubule agents and cytotoxic antibiotic drugs, while antiandrogen drugs were mainly used with microtubule drugs. These combination therapy regimens may confound association studies between individual drugs and the occurrence of mutations in their putative resistance genes, and should thus be rigorously controlled for.



	Antiandrogen	Topoisomerase_Inhibitor	Antibody_therapy	Folic_Acid_Antimetabolite	Kinase_Inhibitors	Others	HER2i	Immune_Checkpoint	EGFRi	Platinum	Pyrimidine	Microtubule	Antiestrogen	Mustards_/_alkylating	Cytotoxic_Antibiotics
Cytotoxic_Antibiotics	66	40	70	44	110	45	59	7	1	185	379	488	392	533	704
Mustards_/_alkylating	63	30	44	49	104	72	57	9	0	96	328	437	383	634	533
Antiestrogen	105	7	54	33	150	34	67	6	0	89	314	392	640	383	392
Microtubule	281	49	123	58	111	99	102	27	9	433	360	1086	392	437	488
Pyrimidine	53	234	308	124	110	49	42	23	50	667	1096	360	314	328	379
Platinum	22	279	290	225	33	37	28	54	70	1194	667	433	89	96	185
EGFRi	0	39	33	31	7	0	3	4	115	70	50	9	0	0	1
Immune_Checkpoint	2	4	5	23	21	2	1	121	4	54	23	27	6	9	7
HER2i	11	1	4	3	4	3	114	1	3	28	42	102	67	57	59
Others	62	8	26	5	23	225	3	2	0	37	49	99	34	72	45
Kinase_Inhibitors	31	5	25	23	299	23	4	21	7	33	110	111	150	104	110
Folic_Acid_Antimetabolite	4	48	45	269	23	5	3	23	31	225	124	58	33	49	44
Antibody_therapy	15	150	398	45	25	26	4	5	33	290	308	123	54	44	70
Topoisomerase_Inhibitor	5	319	150	48	5	8	1	4	39	279	234	49	7	30	40
Antiandrogen	444	5	15	4	31	62	11	2	0	22	53	281	105	63	66

Fig. 1 | Overview of a dataset of pre-treated tumor genomes showing drug combinations and differences in tumor mutation burden. a Number of treatment-naive and pre-treated patients from the 29 cancer types included in this study. **b** Number of patients pre-treated with different drug combinations. The elements on the diagonal represent the total number of patients pre-treated with the drug type on the x-axis, with or without other drug types. The color intensity reflects patient counts, with darker shades indicating higher frequencies as encoded by the heatmap gradient. **c** Number of drug types that were administered to patients prior to biopsy. **d** Differences in the tumor mutation burden (here, number of point mutations per sample) between the treatment-naive (in green) and pre-treated (in red) metastatic tumors in different cancer types. The top and bottom panels represent the changes in the exonic (DiffInvex target) regions and intronic (DiffInvex background) regions per sample, respectively. The median mutation burdens per megabase were shown on top of each panel for treatment-naive tumors (in green) and the pre-treated (in red). The numbers on the bottom of each panel are the *p*-values of the Mann–Whitney U test, two-tailed, comparing the tumor mutation burden between the treatment-naive and pre-treated patients. Source data are provided as a Source Data file.

Controlling for tumor mutation burden changes associated with treatment in tests for selection

Tumor cell mutation burden increases upon treatment with some chemotherapies, and the trinucleotide mutational spectrum typically changes[16]; mutagenicity is clear for the commonly-applied platinum and 5-FU therapies[17,18]. To confirm these trends in our data, we compared the tumor mutation burden (TMB) of point mutations between metastatic tumors in pre-treated versus treatment-naive patients. As expected, we found that exonic mutation burden per sample was positively associated with treatment for many tumor types (*p*-values of 1.8e-5, 5.6e-4, 3.7e-3, 2.8e-2, 4.2e-2 and 4.9e-2 for COREAD, BRCA, RCC, NSCLC, ESCA and BLCA tumors, respectively, using Mann−Whitney U test) (Fig. 1d). We also observed similar differences in metastatic tumors treated with one drug type or other drug types (Supplementary Fig. 2). For example, comparing tumors treated with platinum drugs versus other drugs showed that the TMBs were significantly different for BRCA, COREAD and NSCLC tumors with p = 3.0e-2, 6.4e-3, and 6.7e-6, respectively. Similarly, the TMBs were significantly altered between patients treated with pyrimidine analogs (this group includes 5-FU and its prodrug capecitabine) and patients treated with other drug types (p = 3.4e-2 and 3.3e-2 for NSCLC and OV tumors, respectively, using Mann−Whitney U test). These associations of treatment exposure with exonic mutation burdens (which largely consist of passenger mutations) have the potential to significantly confound the identification of drug resistance driver mutations.

An opportunity presents itself in the rapidly increasing availability of WGS data. This allows to utilize the mutations in the intronic and, optionally, UTR (untranslated regions) and neighboring intergenic regions as a background to model the shifts of exonic mutation burdens. This is the rationale underlying the InVEx (introns-versus-exons) method for identifying positive somatic selection in WGS[13], and here we propose to use the intronic mutations as baseline in a test for differential somatic selection, i.e., identifying conditional driver genes. To validate this concept of intronic rate baselines, we repeated the TMB comparisons using the background intronic/UTR/neighboring intergenic mutations (see "**Methods**"). Expectedly, the treatment-associated changes in the background mutation rates per sample indeed mirrored the pattern of coding exonic mutation changes per sample for most tumor types (Fig. 1d and Supplementary Fig. 2). In particular, the background mutation burden was increased significantly for COREAD, BRCA, RCC and NSCLC metastatic tumors upon treatment (all p < 0.05, Mann−Whitney U test) (Fig. 1d). Similarly, the increase in background mutation burden was significant for platinum drugs versus other drugs for BRCA, COREAD, NSCLC and SARC tumors, and for pyrimidine drugs versus other drugs for NSCLC and OV tumors (p < 0.05, Mann−Whitney U test) (Supplementary Fig. 2).

Next, we further examined the differences between drug treatments on the target and background TMB. We observed strong variation of the target (exonic) mutation rate per sample between different drug combinations, when normalized by the median mutation rate (Supplementary Fig. 3a). However, these variations diminished when target mutation rate per sample was normalized by the background intronic mutation rate per sample (Supplementary Fig. 3b). Taken together, these results show it is essential to account for deviation in passenger mutation rate upon treatment with different drugs while identifying drug-resistance driver gene associations (i.e. conditional positive selection associated with drug treatment), and intronic and UTR/neighboring intergenic mutations can compensate for these deviations.

DiffInvex framework for quantifying conditional selection in cancer

Motivated by our search for genes under stronger positive selection after chemotherapy exposure, we have developed DiffInvex (Differential Introns Versus Exons), a statistical method to identify conditional selection on point mutation in cancer (Fig. 2a). Inspired by the InVEx' approach[13], DiffInvex first estimates the local baseline mutation rate (LBMR) for each gene based on the intronic mutations in addition to UTRs and flanking intergenic mutations to increase statistical power. This empirical LBMR estimate controls for the heterogeneous mutational landscape across the genome, without the need for inferring LBMR from covariate information such as gene expression, replication timing, and histone marks (see Methods). Further, to control for within-gene variation in mutation risk, DiffInvex employs a locus sampling approach, matching the trinucleotide and pentanucleotide composition to control for mutational signature confounding, and also matches the DNA methylation status, a further determinant of sub-gene scale local mutation rates[19], between target (exonic) and background (intronic and intergenic) regions. Additionally, our DiffInvex implementation filters out problematic genomic regions that could distort mutation rate analysis, such as low-mappability regions and hypermutated sites (Methods). Finally, we identified and excluded highly recurrently mutated intronic hotspots (Methods) as they are unlikely to have resulted from selection.

To evaluate if the DiffInvex empirical LBMR could accurately estimate the neutral variation in the mutational landscape between genes, we benchmarked it against modeling of LBMR using epigenomic covariates from dNdScv[14], a state-of-the-art method for identifying driver cancer genes. We showed that our DiffInvex empirical LBMR achieved higher R^2 in explaining the variation in exonic coding mutation rate, compared to dNdScv covariates in the pan-cancer analysis (DiffInvex using pentanucleotide-matching $R^2 = 0.76$, dNdScv covariates $R^2 = 0.62$) as well as most tumor types (Fig. 2b top-left panel, Supplementary Fig. 4a, Supplementary Data 3); we tested only passenger genes here, so as to minimize effects of selection, while assessing the accuracy of the BMR estimates in predicting neural rates. The LBMR estimates of DiffInvex and dNdScv covariates, while largely consistent, were discordant for a small number of genes (Supplementary Fig. 4b; in this particular test, we do not limit to passenger genes). Across cancer types, the DiffInvex R^2 as well as the difference between DiffInvex R^2 and dNdScv R^2 were positively correlated with the mutation burden (Fig. 2b bottom-left panel), probably reflecting noise resulting from a low number of mutations. This suggests that the benefits of an empirical LBMR estimate such as DiffInvex, will grow with the inevitable increase of cancer WGS data. To further support the utility of the intronic LBMR estimate, we also applied DiffInvex and the dNdScv tools to the task of identifying cancer drivers in genomes of different tumor types (see Methods). DiffInvex achieved comparable accuracy to dNdScv in detecting known cancer genes in the Cancer Gene Census database (Supplementary Fig. 4c). Finally, we evaluated the impact of the width of the DiffInvex background window as well as the contribution of DiffInvex modules such as methods for filtering problematic genomic regions, background region width, and use of DNA methylation-based matching, towards accuracy of modeling the neutral LBMR (see Methods, Supplementary Fig. 5). In the pan-cancer analysis, we found DiffInvex filters and DNA methylation-based matching improved the R^2 by 2.5% and 1%, respectively (Supplementary Fig. 5a). We also observed that different background window widths (10 kb, 20 kb and 50 kb) achieved similar R^2 in the pan-cancer analysis, however the 50 kb achieved higher R^2 in individual tumor types (NSCLC, COREAD, ESCA, BRCA and OV), likely because it captures a higher number of background (intronic and intergenic) mutations (Supplementary Fig. 5b). DiffInvex determines how the excess of point mutations in target regions (here, coding exons) over the baseline regions (here, introns and UTRs/gene flanks) differs between two conditions or time points (e.g. pre-treated vs treatment-naive). To that end, DiffInvex utilizes a Poisson regression model for count data, further regularized by a weakly-informative prior to stabilize estimates from sparse data (Methods). Most importantly, while leveraging the empirical intronic-based baseline, the regression model can control for

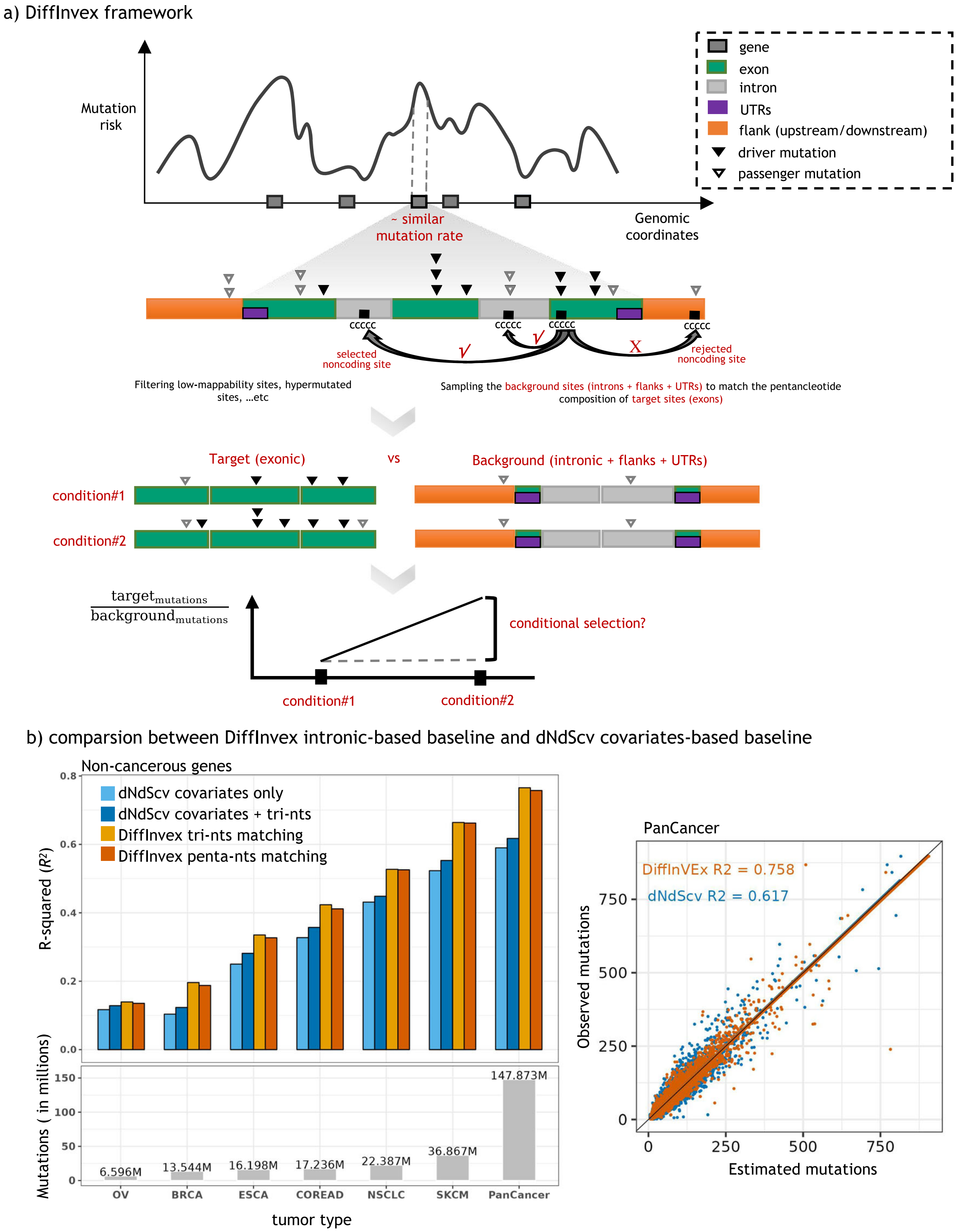


Fig. 2 | DiffInvex method for quantifying conditional selection in cancer whole genomes. a A schematic of the DiffInvex framework that utilizes the intronic, UTRs (untranslated regions) and flanking intergenic regions of each gene as a mutation rate baseline to infer somatic selection and conditional selection. Under no selection, exonic and intronic/UTR regions of the gene have similar mutation rates, after filtering genomic problematic regions and matching their pentanucleotide compositions. DiffInvex normalized the frequency of target (exonic) mutations to background (intronic, UTRs, optionally gene flanks) mutation frequency, comparing two or more conditions to evaluate the conditional selection acting upon that gene. **b** Accuracy of DiffInvex empirical mutation rate baseline compared to the baseline inferred from dNdScv epigenomic covariates in explaining the variation (R^2) in exonic coding mutation rate across genes. Only passenger genes are used in this test, thus removing the effects of selection, and focusing on assessing the accuracy of the neutral mutation rate baselines. The top-left panel shows the R^2 of four different models to estimate the exonic mutations: 20 dNdScv covariates only, 20 dNdScv covariates and exonic-trinucleotide compositions, DiffInvex intronic mutation rate after trinucleotide matching between exonic and intronic regions, and DiffInvex intronic mutation rate after pentanucleotide matching between exonic and intronic regions. The bottom-left panel shows the total number of mutations in different tumor types, demonstrating that R^2 is correlated with the number of mutations. The right panel shows a correlation between the estimated and observed mutation counts in passenger genes (dots) in the pan-cancer cohort, using the DiffInvex rates with pentanucleotide matching (in red) and the modeling based on dNdScv covariates together with exonic trinucleotide composition (in green). Source data are provided as a Source Data file.

the interactions between various tested conditions, as well as the confounding factors, such as tumor types, which may be differently abundant between pre-treated and treatment-naive samples (Fig. 1a and Supplementary Fig. 1d), and it can control for technical variation between data sourced from different cohorts (Supplementary Fig. 1c). In the current study, the DiffInvex methodology allows us to control for combined drug treatments where one drug treatment confounds gene association testing with the other frequently co-administered drugs.

Identifying chemotherapy-associated drivers via DiffInvex analysis of cancer genomes

We utilized DiffInvex on a large-scale dataset of 8591 cancer WGS and 147.87 million somatic mutations to identify putative drug resistance genes associated with each drug type, genes that gain positive selection through treatment; additionally, putative drug sensitivity genes that gain negative selection (in relative terms) are identified. First, in a pan-cancer analysis, we applied DiffInvex to call these drug type-specific resistance genes by comparing the mutation profiles of patients exposed to a specific group of drugs (e.g. platinum-based drugs) and treatment-naive patients, controlling for confounders including tumor type, cohort, tumor stage (primary or metastatic) and patient sex. For each gene, 15 DiffInvex regression tests were performed to assess the interaction with each drug type separately, in the first instance (Methods).

As a positive control, we asked whether DiffInvex can identify previously-reported genes in which occurrence of point mutations can grant resistance to targeted therapy drugs. Indeed, we identified association of mutations in the *ESR1* gene with prior exposure to antiestrogen drugs, mutations in *AR* gene with antiandrogen drugs, *EGFR* gene mutations with EGFRi drugs, and *KIT* mutations with kinase inhibitor drugs (all effect sizes > 2.0, all FDRs <1e-8; the “effect size” here refers to natural log fold-enrichment of coding exonic mutation rates in pre-treated samples compared to treatment-naive, normalizing for the change in intronic mutation burdens) (Supplementary Fig. 6a). However, in the initial analysis that considered every drug type separately, we observed that some genes were strongly associated with resistance to many drug types (Supplementary Fig. 6b), which is implausible. For example, *ESR1* gene mutations were also strongly associated with 7 drug types including nitrogen mustards and other alkylating drugs, cytotoxic antibodies, topoisomerase inhibitors, kinase inhibitors, microtubule agents and pyrimidines (all FDR <$1e^{-10}$), which might be because most BRCA patients were treated with antiestrogen drugs in combination with other drugs (Fig. 1c). Similarly, *APC* and *SMAD4* tumor suppressor genes and *PIK3CA* and *BRAF* oncogenes were significantly (all FDR < 25%) associated with 9, 4, 6, and 3 drug types respectively. These patterns indicated that conjoint drug treatment (only 548/2730 patients were pre-treated with only a single drug type) likely confounded our initial drug-gene target association analysis.

To overcome this challenge, we applied DiffInvex regression to quantify conditional selection associated with each drug type by a joint analysis comparing the mutation profiles of patients exposed to different combinations of the 15 drug types (single and multiple drug types) with treatment-naive patients. For each gene, DiffInvex assesses the contribution of each drug type towards the excess of selected mutations upon treatment simultaneously in a single regression test (Methods).

This allowed DiffInvex to control for the confounding introduced by the conjoint drug treatments as well as tumor type and tumor stage (primary or metastatic). The positive control associations remained (Fig. 3a, Supplementary Data 4) including *ESR1* mutations associated with antiestrogen drugs (effect size = 3.1, FDR = 5.8e-29), *EGFR* mutations with EGFR inhibitor drugs (effect size = 2.71, FDR = 1.0e-25), *KIT* mutations with kinase inhibitor drugs (effect size = 1.99, FDR = 1.5e-9) and *AR* mutations with antiandrogen drugs (effect size = 1.83, FDR = 3.6e-7). In the revised analysis, we no longer see *ESR1*, *APC*, and *SMAD4* mutations significantly associating with multiple drug types (1, 2, and 1 associations at FDR < 25%, respectively), suggesting that the joint drug treatments are not strongly confounding (Supplementary Fig. 6b). Further, a quantile-quantile plot of the *p*-values from the association analysis suggested that the *p*-values were conservative (inflation factor λ = 0.54, Supplementary Fig. 6c), meaning that false positive rates are stringently controlled, albeit at the expense of some false negatives i.e., missed associations.

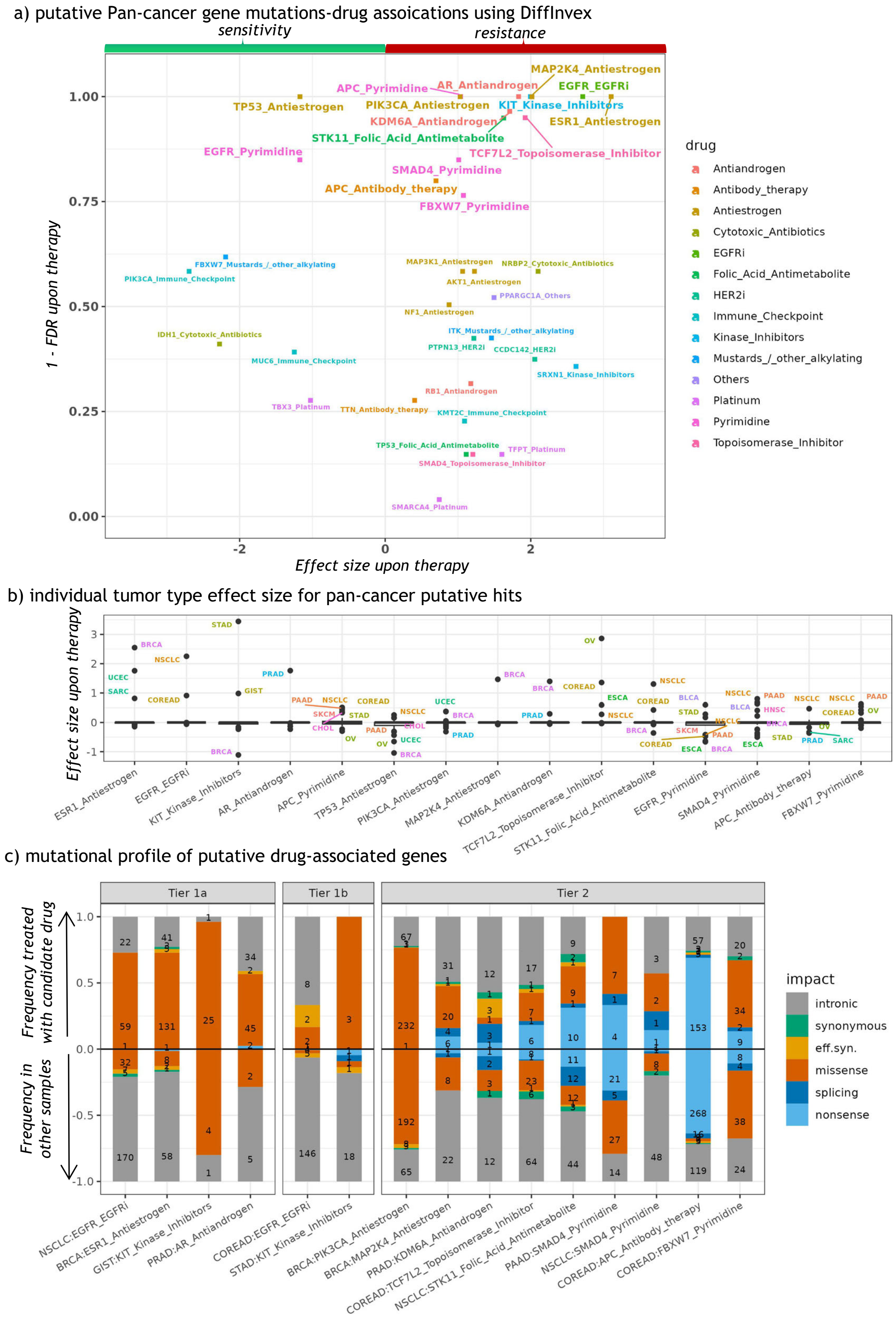
In addition to these known positive control genes, DiffInvex identified additional treatment-associated genes with different drug types (11 gene-drug associations with FDR < 25%, of which 9 with resistance and 2 with sensitivity; Fig. 3a). Two general properties of these drug resistance genes stand out. Firstly, the hits are all known oncogenes and tumor suppressor genes, meaning they are under positive selection even without therapy exposure. (Of note, our methodology is in principle able to identify genes which have selected mutations only after therapy but are not otherwise mutational cancer drivers, such as the known example of *AR* gene. Indeed, the lower-confidence tier of hits with FDR = 25-80% does contain 2 such genes, *NRBP2* and *PPARGC1A*). Secondly, the conditional selection effect sizes of newly-identified hits are lower (natural log mutation rate enrichments all <=2) than those of the well-known common drug resistance mutations, the *ESR1* in relation to antiestrogen drugs and *EGFR* mutations in relation to EGFRi drugs (natural log mutation rate enrichments ~ 2.5).

The specific associations we identified at FDR < 25% include those of mutations in *PIK3CA* and *MAP2K4* with resistance to antiestrogen drugs, *APC*, *SMAD4*, and *FBXW7* mutations with resistance to pyrimidine drugs, *TCF7L2* mutations with resistance to topoisomerase inhibitor drugs, *KDM6A* mutations with resistance to antiandrogen drugs, and *STK11* mutations with resistance to folic acid antimetabolites. Overall, this suggests that resistance to various anticancer drugs is commonly mediated not through mutations in drug-specific resistance or target genes, but by accumulating additional general driver mutations with moderate effect sizes.

Robustness of the methodology and comparison to other methods

We evaluated DiffInvex against Coselens[9] and cancereffectsizeR[10] methods for quantifying conditional selection in cancer (Supplementary Figs. 7–10). Both methods estimate the baseline mutation rate using the covariate-based dNdScv approach[14]. Moreover, they assess the association with different conditions-of-interest individually, being naïve to the correlations between the conditions (here, co-administered drug treatments). In brief, these methods identified implausibly many drug associations with various genes (Supplementary Fig. 7), including known, control drug resistance genes (Supplementary Figs. 8, 10). While they confirmed many DiffInvex-identified associations (Supplementary Fig. 9), they linked many other drug types with those genes, sometimes with unusually large effect sizes (Supplementary Fig. 10). These differences, as well as example genes affected, are discussed in more detail in Supplementary Note 1. Overall, we suggest DiffInvex reports a restricted yet high-confidence set of associations with positive selection, which is beneficial for prioritizing associations for follow-up studies on drug resistance.

We next assessed the importance of our empirical, intron-based mutation rate baseline. We implemented the DiffInvex regression model using only the exonic mutations for identifying drug-associated genes (see Methods; this exon-only test is not able to control for differences in mutation rates and spectra associated with treatment). While the *AR*-antiandrogen and *KIT*-kinase inhibitors positive control associations were not statistically significant anymore (Supplementary Fig. 11), some less plausible associations with the “other”



(heterogeneous) drug group were identified instead. Additionally, the effect sizes of the 4 known associations were larger for the "default" DiffInvex with intronic baseline (Supplementary Fig. 11). These observations support the utility of the intronic-based empirical baseline for tracking the changes in tumor mutation rates and/or spectra between individuals.

Finally, we assessed the robustness of DiffInvex results towards sequencing technical factors (mappability, depth of read coverage) and a relevant biological factor (mutation clonality). Here, we used a dataset consisting of 5226 Hartwig and PCAWG samples with 107.55 million mutations as they were uniformly processed with the same HMF bioinformatics pipeline[20].

Fig. 3 | Chemotherapy-associated mutational driver genes identified by DiffInvex. a Drug sensitivity and resistance associations identified in the pan-cancer analysis of 8,591 cancer genomes, while controlling for confounding between joint drug treatment. At FDR < 25%, there are 13 drug resistance and 2 drug sensitivity associations. Associations are colored by the drug type. The x-axis represents the effect size upon treatment: the natural log fold-enrichment of coding exonic mutation rates in tumor samples pre-treated with the putative drug type compared to other samples, normalizing for the difference in intronic mutation burdens between two groups of samples. Obtaining FDR via DiffInvex statistical test is described in Methods. **b** Effect sizes of associations with treatment, derived from the DiffInvex analysis in individual tumor types (listed with acronyms next to corresponding points), shown here for the significant pan-cancer associations. **c** Mutation functional impact profile of putative drug-associated genes in candidate tumor types. The upper-half of the plot (positive values) represents the fractions of different mutation types (nonsense, splicing, high-impact missense, effectively synonymous (low-impact missense), synonymous, and intronic) in the candidate gene in tumor samples treated with a certain the putative drug type, as listed on the x-axis. The lower half of the box plot (negative values) represents the fractions of these mutation types in that gene in other tumor samples (treatment-naive samples and samples pre-treated with other drug types). The effectively synonymous ("eff.syn.") mutations are missense mutations with AlphaMissense pathogenicity score <0.39. Tier 1 encompasses known drug-gene mutation associations, where Tier 1a does so in known tumor types while Tier 1b includes additional tumor types. Tier 2 has candidate new gene-drug associations with DiffInvex FDR < 25% in the pan-cancer analysis. Please see Supplementary Fig. 6 for similar plots for the more tentative Tier 3a and 3b associations. Source data are provided as a Source Data file.

We first applied DiffInvex to the 5226 samples (for GEM mappability, 3660 samples subset thereof) and determined the selection coefficients for the full cohort. Then, we divided the mutations into two half-datasets, based on a specific criterion (mappability, read depth, mutation clonality) and applied DiffInvex to each half separately (Supplementary Figs. 12–14). We observed a strong positive correlation between the effect sizes of the full dataset and the half-datasets with R^2 ranging from 0.69 to 0.89 (Supplementary Figs. 12–14, top panels). These R^2s were broadly similar to the maximum attainable correlations in that test, which does not reach 1.0 because of the noise introduced by randomly selecting a subset of mutations (Supplementary Figs. 12–14, bottom panels). Thus, results appear minimally affected overall by technical artefacts due to read alignment/mutation calling, and by mutation clonality. As a side note, we also observed that subclonal mutations are increased in tumors of pre-treated patients, consistent with treatments generating the mutations and/or clonal diversification in later stages of tumor evolution (test based on purity-adjusted variant allele frequencies, Supplementary Fig 13b).

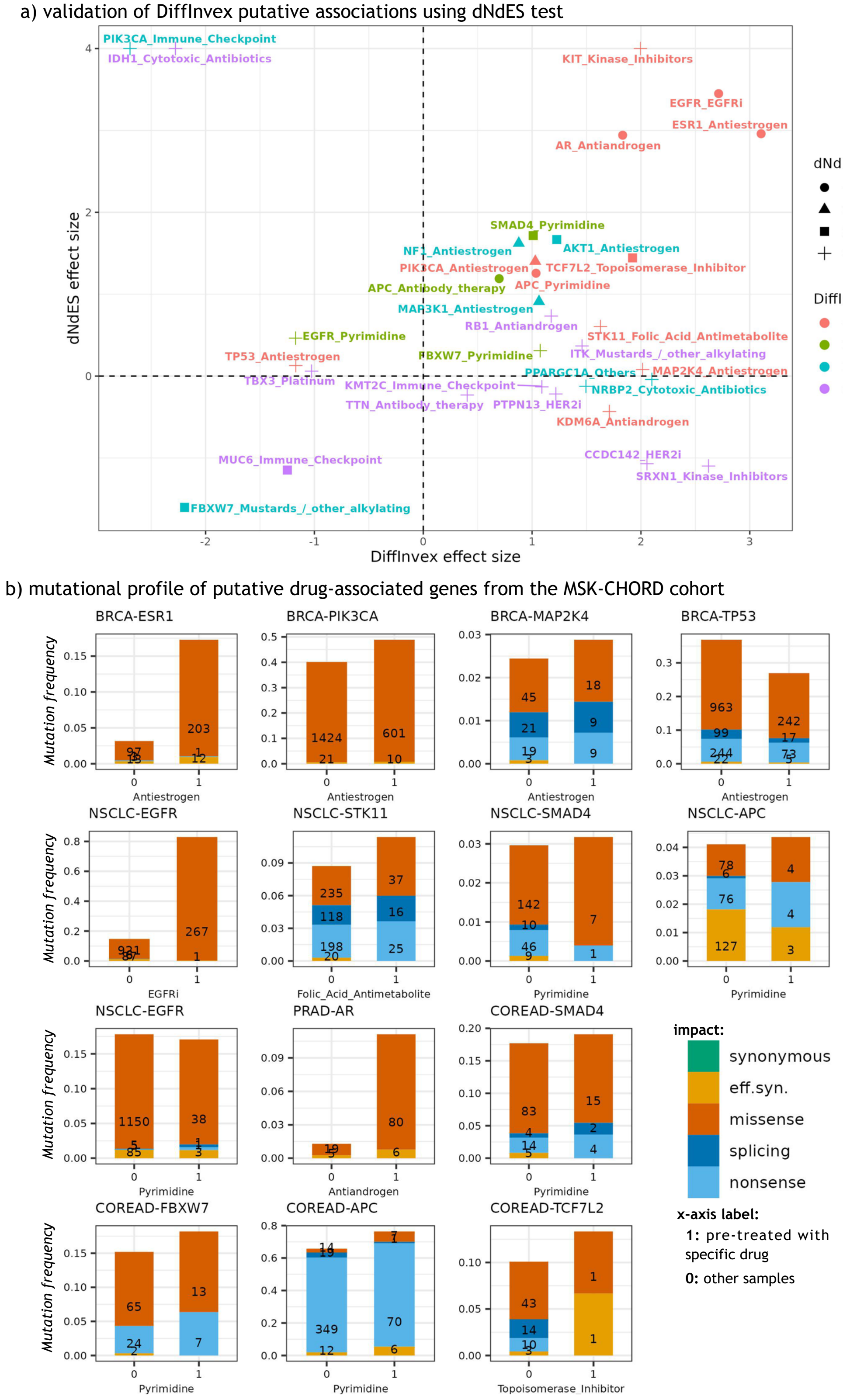
It is worth mentioning that the DiffInvex framework can be used to control for additional confounders (if any) by splitting mutations in each sample into bins based on e.g. the quality control measurements and adding the bins as a covariate into the regression model.

Classification of the differentially selected genes by cancer type and prior knowledge

Next, we asked in which individual tumor type these pan-cancer drug-target associations were most relevant, by performing the association testing in individual cancer types and monitoring the effect sizes (Fig. 3b; statistical power was limited to report significance of associations in individual cancer types). As an overall trend, the highly cancer-type-specific but strong effect size associations appear to be rare. Instead, the more common case is that the conditional selection upon drug treatment has moderate effect sizes, but they tend to be observed with some consistency across several cancer types (Fig. 3b).

Based on these association testing results, additionally taking into consideration validation analyses described in the following sections, we divided the observed gene-drug-cancer type combinations into 5 tiers by reliability (Fig. 3c and Supplementary Fig. 6d).

- Tier 1a includes known drug-gene mutation associations in known tumor types such as *ESR1* mutations-antiestrogen resistance in BRCA[21], *EGFR* mutation-EGFRi resistance in NSCLC[22], *KIT* mutations-kinase inhibitor resistance in GIST[23], and *AR* mutations-antiandrogen resistance in PRAD cancer type[24]. These well-known associations are recapitulated in our analyses where they serve as a positive control.
- Tier 1b contains known gene mutation-drug resistance (or sensitivity) pairs but includes other tumor types. In other words, these are bona fide drug resistance drivers, however, their tissue spectrum is redefined, such that additional tumor types could benefit from understanding these drug resistance mechanisms and, possibly, countering them. For example, we find that the *EGFR* mutation-treatment by EGFRi association, which is well-known in NSCLC (effect size [natural log ratio of mutation rates] = 2.25), is relevant to COREAD as well at a suggestive significance threshold (effect size = 0.91). Similarly, our analysis indicates that the *KIT* mutations associated with kinase inhibitors are relevant for STAD (effect size = 3.44) as a Tier 1b hit in our analysis, besides the known Tier 1a relevance to drug resistance in GIST (effect size = 0.98).
- Tier 2 has candidate new gene-drug associations with solid statistical support (here, pan-cancer FDR < 25%; we note again that our *p*-values and therefore also FDRs are likely conservative, see above); these can be associated to certain cancer types by considering the distribution of effect sizes across individual cancer types (Fig. 3b). The Tier 2 encompasses the following observed hits: *PIK3CA* mutations-antiestrogen resistance and *MAP2K4* mutations-antiestrogen resistance, both in BRCA; *KDM6A* mutations-androgen deprivation resistance in PRAD, *TCF7L2* mutation-topoisomerase inhibitor and *APC* mutation-antibody therapy resistance, both in COREAD; *STK11* mutations-folic acid antimetabolites resistance in NSCLC; *SMAD4* mutations-pyrimidine analog resistance in PAAD. Some of these mutation-drug associations were supported by a statistical test on the functional impact of mutations as well as replicated in independent cohorts of pre-treated tumors, further increasing confidence therein (see sections "Validation of gene-drug associations using dNdES test for differential functional impact" and "Validation of gene-drug associations using exome or panel sequencing in independent cohorts" below).
- Tier 3 includes additional putative gene mutation-drug associations, but with limited evidence of the relevance of those gene mutations in drug resistance (or sensitivity) in individual tumor types. This divides into two categories, firstly Tier 3a that includes putative associations of a frequently-mutated cancer driver gene (pan-cancer FDR < 25%, Fig. 3), however, these associations indicate different responses (sensitivity or resistance) in different tumor types (Fig. 3b). For example, *TP53* mutations indicated sensitivity to antiestrogen drugs in the pan-cancer analysis (Fig. 3a, left side), but the effect varied in direction across cancer types (Supplementary Fig. 6d), therefore, they are marked as lower confidence than Tier 2 hits. Tier 3a also includes associations of *APC* mutations-pyrimidine analog resistance, and *EGFR* mutations-pyrimidine analog sensitivity in NSCLC and COREAD tumors.
- Tier 3b includes more preliminary gene mutation-drug associations, defined as those hits with only a tentative FDR < = 80% in our main pan-cancer analysis (Fig. 3a, Supplementary Fig. 6d), which could however be validated using the functional impact test and/or in independent sequencing cohorts (*MAP3K1*-antiestrogen in BRCA, *RB1*-antiandrogen in PRAD and *KMT2C*-immune



checkpoint in NSCLC) (see section below and Fig. 4). Here, alongside *TP53* and *EGFR*, we note an additional hit denoting a sensitizing mutation, indicating negative selection on *FBXW7* mutations associated with treatment using mustards and other alkylating agents, which validated in the functional impact test at a tentative FDR (see below; Fig. 4).

Overall, the associations in Tiers 3a and 3b should be considered cautiously, given their more modest FDRs, however, they may be reasonable to consider for some forms of downstream follow-up work.

To explore the functional impact of these mutations, we further visually inspected their distribution across driver hotspots in the candidate drug resistance genes, contrasting the pre-treated and

Fig. 4 | Validation of DiffInvex drug-gene associations in a functional impact test and in independent MSK-CHORD data set. a A test for increased mutation functional impact associated with treatment, dNdES (ES, effectively synonymous) of the pan-cancer associations previously identified by DiffInvex. The scatter plot shows the DiffInvex and dNdES effect sizes of the putative hits. The dNdES effect size is the natural log fold-enrichment of high-impact mutations (nonsense, splicing, and high-impact missense, i.e., those with AlphaMissense score >= 0.39) in tumor samples pre-treated with the particular drug type compared to other samples, normalizing for the change in low-impact mutations (synonymous and effectively synonymous i.e., low-impact missense). Associations are colored by the FDR range in the pan-cancer analysis by DiffInvex in Fig. 3a. Here, only the DiffInvex gene mutations-drug associations with DiffInvex FDR < 90% were tested using the dNdES test. In this plot, we capped the dNdES selection coefficients at the value of 4. Obtaining FDR via the DiffInvex statistical test is described in Methods. **b)** Replication of some of the DiffInvex significant associations using panel sequencing data from the MSK-CHORD dataset. In each panel, the title is "tumor type-gene" and the x-axis is the putative drug type. The bar plot shows the mutational impact profile for putative drug-associated genes of pre-treated samples with that drug (x-axis = 1) and of other samples (x-axis = 0) in the candidate tumor types. The effectively synonymous ("eff.syn.") mutations are the low-impact missense mutations that have AlphaMissense pathogenicity score <0.39. Source data are provided as a Source Data file.

treatment-naive tumors (lollipop plots in Supplementary Figs. 15–17). In the Tier 1a hits – known drug-resistance mutations in *EGFR*, *KIT* and *AR* – do exhibit specialized drug-resistance mutations as expected (Supplementary Fig. 15). In contrast, in the Tier 2 and Tier 3 genes the distributions of driver mutations seem broadly consistent across the treated and non-treated tumor genomes (see example of *PIK3CA* in antiestrogen treated vs untreated BRCA in Supplementary Fig. 16a). These genes and encoded proteins are not the known targets of the drugs we find they are associated with, and would appear to confer resistance via an increase in fitness rather than by specialized resistance mechanisms. While this is the general trend, there might be individual cases that run counter to that (e.g., the Tier 2 gene *MAP2K4* and Tier 3b gene *MAP3K1* with antiestrogen resistance). In case of *MAP2K4* gene, exon 5 seems to be enriched in splice-site mutations in pre-treated tumors, compared to treatment-naive tumors, which could disrupt RNA splicing and protein-coding sequence (Supplementary Fig. 17a). Additionally, in case of *MAP3K1*, exon 10 versus exon 14 appears differentially affected by nonsense mutations, thus generating truncations of different lengths (Supplementary Fig. 17b). However, larger cohorts are needed to test the significance of the *MAP2K4* and *MAP3K1* individual treatment-associated mutations.

Validation of gene-drug associations using dNdES test for differential functional impact

To further filter out false positives in these Tiers 3a and 3b, and also to validate the more confident Tiers 1a, 1b, and 2 above, we next devised and implemented a complementary statistical test for differential selection associated with drug therapy. This test is based on the rationale that if the mutation rate increases after therapy is biologically meaningful, the functional impact of the mutations on the protein sequence after treatment should also increase compared to the untreated group.

Our differential functional impact test, named dNdES, can confirm the DiffInvex hits by using a baseline that does not rely on introns and gene flanks. For each gene, the dNdES test compares the frequency of nonsynonymous to "effectively synonymous" (ES) mutations under different conditions, here implying, again, tumor samples treated with a drug type versus other tumor samples (treatment-naive samples and samples pre-treated with other drug types). We annotated exonic mutations with AlphaMissense, a state-of-the-art predictor for mutation pathogenicity[25], thus classifying the missense mutations with a score <0.39 as ES and grouping them together with synonymous mutations to form a baseline for estimating functional impact (see Methods). Reassuringly, the positive side of the log nonsynonymous to effectively-synonymous ratio (log dNdES) was broadly correlated to the exonic-to-intronic ratio (log dEXdIN from DiffInvex; Supplementary Fig. 18a), considering general selection on known cancer genes. This correlation remains observed for different cutoff values of the AlphaMissense score, where the cutoff of 0.39 resulted in a slope ~=1 for this correlation on cancer genes (Supplementary Fig. 18c). Of note, dNdES ratios have a higher spread compared to DiffInvex ratios (Supplementary Fig. 18b), likely because of more sparse mutation occurrence in the ES loci used as a neutral baseline, compared to the more abundant intronic loci in the default DiffInvex (dEXdIN) approach.

Next, we applied the dNdES test to identify conditionally-selected genes, comparing the functional impact of exonic mutations in pre-treated tumors relative to treatment-naive tumors. Encouragingly, the effect sizes of conditional selection between the dNdES and DiffInvex tests correlate (Fig. 4a, Supplementary Data 4), and many DiffInvex putative associations have been validated in the dNdES test (Fig. 4a). Out of 15 DiffInvex significant associations (FDR < 25%), 6 associations were also statistically significant using dNdES test (FDR < 25%), 2 were dNdES-significant at a permissive FDR 50%, and additional 4 associations had a consistent direction of response (sensitivity or resistance) as they did in the discovery DiffInvex.

In particular, the known positive control, Tier 1 hits including *ESR1*-antiestrogen, *EGFR*-EGFR inhibitors and *AR*-antiandrogen had strong effect sizes and significance in the dNdES test (Fig. 4a). Additionally, the pan-cancer study of differential dNdES validated some Tier 2 and Tier 3a associations (FDR < 25%) including *APC*-pyrimidine analogs, *PIK3CA*-antiestrogens, and *APC*-antibody therapy drugs; at a permissive FDR < 50%, dNdES additionally supported *SMAD4*-pyrimidine analogs and *TCF7L2*-topoisomerase inhibitors. However, other associations did not reach significance in the (less-powered) dNdES test, including associations of *KIT*-kinase inhibitors, *STK11*-folic acid antimetabolites, *MAP2K4*-antiestrogen, and *FBXW7*-pyrimidine analogs. In some cases, this might be attributed to the low number of mutations used as a baseline. For example, in testing the known *KIT* mutations-kinase inhibitors association, no synonymous or effectively synonymous mutations were observed in *KIT* in the samples pre-treated with kinase inhibitor drugs (Fig. 3c), and so this association could not be validated, but was not invalidated either.

Additionally, in Tier 3b hits, two associations were significant in the dNdES test at FDR < 25% (*NF1* and *MAP3K1*-antiestrogen resistance), and one association at dNdES at FDR < 50% (*AKT1* mutations - antiestrogen resistance).

Other Tier 3b associations did not reach significance, even though some did exhibit the correct direction of effect in the dNdES test, such as *RB1*-antiandrogen resistance, and the *ITK* gene mutations associated with resistance to mustards & other alkylating agents (this drug group excludes platinum). The remaining Tier 3a and Tier 3b candidates for drug resistance mutations were invalidated by dNdES, in particular the *SRXN1* and *CCDC142* candidates.

Regarding the tentative sensitizing mutations (negative conditional selection associations), these were rare in the initial DiffInvex test – only 4 significant at FDR < 50% – of which 3 were not supported in the dNdES test, including *EGFR*, *PIK3CA* and *IDH1* apparently sensitizing mutations (Fig. 4a). One of the 4 was clearly supported in dNdES: the *FBXW7* mutations sensitizing association with mustards and other alkylating agents (Fig. 4). Technically the *MUC6*-immunotherapy association also meets criteria for Tier 3b although this is very low-confidence in DiffInvex (FDR > 50%, Fig. 4a). We note that the *TP53*-antiestrogen link and the *EGFR*-pyrimidine link do have some support in external datasets; see below. As a caveat, the dNdES test relies on AlphaMissense and may not be informative in case of genes or sites where AlphaMissense does not perform well. This might explain some of the cases with ~0 dNdES effect size (Fig. 4a), such as the link between *NRBP2* mutations and cytotoxic antibiotic resistance. The above results provided support for the DiffInvex approach overall, and supported many individual putative hits identified by DiffInvex by detecting an increase in impactful mutations in pre-treated tumors.

Replication of gene-drug associations using a large panel sequencing cohort

Next, we sought to replicate the DiffInvex drug resistance gene associations using panel sequencing and whole-exome sequencing (WES) data from independent studies with available treatment information. We obtained the panel sequencing data of the well-curated MSK-CHORD cohort with treatment information[26]. After filtering, we retained 22,914 samples that include 7238 non-small-cell lung, 5300 colorectal, 4851 breast, 2898 pancreatic, and 2627 prostate tumor samples, including treatment-naive and pre-treated samples (see Methods). For each putative gene-drug association we identified in the DiffInvex discovery cohorts, as a validation, we compared in the replication cohort the frequency of mutations in this gene in the pre-treated samples with that drug versus the other tumor samples in the same cancer type. Our validation analysis considered only the frequencies of coding mutations and splice site mutations; since these were not WGS datasets, DiffInvex as such could not be applied. As a second type of validation, we stratified the missense mutations into low-impact (ES) and high-impact missense, again based on AlphaMissense scores as earlier; synonymous mutations were not available for this dataset so the ES group here reflects low-impact missense mutations only. We compared the proportions of different mutation impact categories as in the dNdES test above. This analysis confirmed many associations for different tumor types (Fig. 4b and Supplementary Fig. 19).

Overall, we observed a good agreement between the DiffInvex WGS discovery cohort and the MSK-CHORD pan-cancer analysis as a validation cohort (Supplementary Data 5), when comparing effect sizes for the DiffInvex-significant gene-drug associations (FDR < 25%). In particular, we compared the ratio of mutation rates in each gene in the samples that were pre-treated with a drug, to mutation rates in that gene in other samples; these effect sizes correlated at $R^2 = 0.815$ between DiffInvex discovery and MSK-CHORD validation (Supplementary Fig. 19a). Similarly, a strong correlation ($R^2 = 0.822$) was shown for the ratio of high-impact (missense, nonsense or splice) to low-impact ES mutations in associated drug pre-treated samples to other samples between the DiffInvex discovery data and MSK-CHORD validation data (Supplementary Fig. 19b).

Furthermore, we considered if the individual gene-drug associations from DiffInvex were replicated in analysis of the matching tumor types from MSK-CHORD panel sequencing validation (Fig. 4b). In breast cancer, the positive control *ESR1*-antiestrogen association, as well as the discovered resistance associations herein *PIK3CA*-antiestrogen, and *MAP2K4*-antiestrogen were replicated, as well as the *TP53*-antiestrogen sensitizing association. For example, 48% (601 out of 1250) of pre-treated tumors with antiestrogen bear high-impact missense *PIK3CA* mutations, while 39.5% (1424 out of 3601) other tumors have high-impact missense *PIK3CA* mutations (Fig. 4b). In case of *MAP2K4*-antiestrogen treatment resistance, the frequency of high-impact (missense, nonsense or splice) mutations was modestly increased to 2.8% in antiestrogen pre-treated patients, compared to 2.4% in other patients, and as additional support the former did not have any low-impact missense mutations, while the latter did (Fig. 4b). Additionally, *TP53* mutation-antiestrogen sensitivity (which did not validate in dNdES functional impact test on pan-cancer WGS, Fig. 4a) was confirmed with 26.6% (332 out of 1250) of antiestrogen pre-treated patients harbor high-impact mutations, compared to 36.2% (1306 out of 3601) of other patients have high-impact mutations.

Similarly, many gene-drug associations were validated in non-small lung cancer, including the positive control *EFGR*-EGFR inhibitors association, as well as the discovered associations herein *STK11*-folic acid antimetabolites resistance, *SMAD4*-pyrimidine analogs, *APC*-pyrimidine analogs and *EGFR*-pyrimidine analogs. For example, 11.3% (78 out of 685) of pre-treated tumors with folic acid antimetabolites bear high-impact missense, nonsense or splice *STK11* mutations, while 8.5% (551 out of 6553) other tumors have such *STK11* mutations; consistently, the relative frequency of ES (low impact missense) mutations in *STK11* is reduced in the treated group (Fig. 4b). The frequency of high-impact *SMAD4* mutations was modestly higher in pre-treated with pyrimidine analogs drugs (3.2% versus 2.8%), and similarly, *APC* was confirmed in that 3.2% (8 out of 252) pyrimidine analog pre-treated patients bear high-impact mutations, compared to 2.3% of other patients (Fig. 4b). In contrast, *EGFR*-pyrimidine analogs sensitizing association (which did not validate in dNdES functional impact test in pan-cancer WGS, Fig. 4a) was observed in the lung validation data however with only a very modest difference in the correct direction (15.8% versus 16.6% patients with high-impact mutations).

Additionally, in prostate cancer validation analysis, the *AR*-antiandrogen resistance was confirmed with 10.3% pre-treated versus ~1% other tumors have high-impact mutations in that gene.

Next, we turn to colorectal cancer validation. In the MSK-CHORD colorectal cancer cohort, we noted variability in the ratio of mutation rates between different subgroups of patients (COREAD implying tumors classified as colorectal without further subdivision, and additionally the metadata supplied separate classification for some tumors into colon and rectum (READ), defined by “Cancer Type Detailed” field of the clinical information of the MSK-CHORD cohort) (Supplementary Fig. 20). In COREAD patients, the associations of *SMAD4*-pyrimidine analogs and *FBXW7*-pyrimidine analogs resistance were replicated (Fig. 4b). For example, the frequency of high-impact *SMAD4* mutations was modestly higher in patients pre-treated with pyrimidine analog drugs (19.1%, 21 out of 110), compared to others (16.8%, 101 out of 599). The frequency of high-impact *FBXW7* mutations was 18.2% in pyrimidine analog pre-treated patients, higher than 14.9% in other patients. These two associations were supported in the colon cancer patient subset in specific, by the increase of the high-impact mutations percentage in the pre-treated patients with pyrimidine analog drugs (Supplementary Fig. 20). Additionally, the *APC*-pyrimidine analogs and *TCF7L2*-topoisomerase inhibitors resistance associations were indeterminate in the colorectal validation data: they did trend in the correct direction, however *APC* with modest effect size and slightly increased relative proportion of lower-impact mutations (counter to expectation), and *TCF7L2* had very low mutation numbers (Fig. 4b).

The MSK-CHORD panel sequencing data also confirmed two lower confidence DiffInvex associations: *MAP3K1*-antiestrogen (DiffInvex FDR < 50%) in breast cancer and *RB1*-antiandrogen (DiffInvex FDR < 75%) in prostate cancer (Supplementary Fig. 19c).

Replication of associations in additional genomic data comparing untreated and pre-treated tumors

Additionally, we obtained other panel and whole-exome sequencing data from smaller independent studies with available treatment information. This includes the NSCLC and CRC cohorts from the GENIE Biopharma Collaborative (BPC) project[27], as well as the BRCA cohort from the metastatic breast cancer (MBC) genomics project[28]. We obtained the mutation profiles of 1747, 1469, and 186 samples from NSCLC, COREAD, and BRCA cohorts, respectively, which were either treatment-naive or pre-treated (see Methods).

We validated DiffInvex gene-drug associations for non-small cell lung cancer and colorectal cancer using the GENIE cohort (Supplementary Fig. 21a). In NSCLC patients, this includes the positive control *EGFR*-EGFR inhibitor resistance, as well as the putative *EGFR*-pyrimidine analogs treatment sensitivity and, interestingly, the link between *KMT2C* mutations and immune checkpoint inhibitors resistance (which did not validate in the dNdES test in the original pan-cancer WGS analysis). In colorectal patients, we note support for our Tier 1b association between *EGFR* mutations and pre-treatment with EGFR inhibitors. In this case, the frequency of high-impact mutations was 5% in pre-treated colorectal patients with EGFR inhibitor drugs, compared to 1.7% in other colorectal patients. Further, the link between *APC* mutations and resistance to pyrimidine analogs treatment or antibody therapy (here, commonly, this implies the Bevacizumab drug: 349 out of 398 cases of the antibody therapy in our data set are bevacizumab) was supported in the GENIE colorectal cohort (Supplementary Fig. 21a).

Our analysis also confirmed many antiestrogen associations in the external breast MBC dataset[28]. This included the positive control *ESR1*-antiestrogen associations, as well as the discovered associations herein *PIK3CA*-antiestrogen, *TP53*-antiestrogen, and *MAP3K1*-antiestrogen (Supplementary Fig. 21b). For example, in case of *PIK3CA*-antiestrogen resistance, 30% pre-treated tumors with antiestrogen drugs have high-impact missense *PIK3CA* mutations, while 18% of other tumors bear high-impact missense mutations. Another example, we highlight *TP53* mutation-antiestrogen treatment sensitivity, which had high DiffInvex significance. In pre-treated patients with antiestrogen drugs, 88.9% of the *TP53* mutations are high-impact, while in other tumors, 97.3% of the *TP53* mutations are high-impact. Thus, the frequency of *TP53* high-impact mutations was lower in antiestrogen patients (18.6%), compared to other patients (26%).

Moreover, results from other smaller-scale studies in breast cancer patients showed support of some DiffInvex associations. Using a longitudinal study of 48 patients on aromatase inhibitors (a type of antiestrogen therapy), Lopez-Knowles et al.[29] reported that *MAP2K4* mutations were observed in 1 patient post-treatment with antiestrogen drugs, but not pre-treatment. Additionally, Brady et al[30]. showed that out of 3 patients treated with antiestrogens, two patients gained *ESR1* mutations and one patient gained *MAP2K4* mutation upon treatment. Similarly, Huang et al[31]. suggested that *PIK3CA* mutations contributed to fulvestrant resistance in ER-positive breast cancer as they observed mutations in 3 out of 4 fulvestrant-resistant patients. In another study using ctDNA from 141 advanced breast cancer patients that underwent first-line standard treatment, Liao et al.[32] showed that *PIK3CA* and *TP53* variants were associated with drug resistance. *PIK3CA* and more generally the PI3K pathway has convergent clinical and experimental evidence supporting its causal involvement in breast cancer resistance to endocrine therapies (reviewed in Araki et al.[33] and Rusquec et al.[34]). In a clinical study, the inhibitor of the *PIK3CA*-encoded p110α, alpelisib, potentiated the effects of fulvestrant therapy significantly only in the *PIK3CA*-mutant breast tumors[35]. Analysis of compiled data from three earlier clinical trials indicated that tumors with *PIK3CA* mutations had a lower response rate to antiestrogen therapy[36]. Our large-scale genomics analysis supports that *PIK3CA* mutations preferentially occur after therapy, implicating this gene in antiestrogen resistance. Overall, in breast cancer, we identified associations of *PIK3CA*, *MAP2K4*, and *MAP3K1* mutations with resistance to antiestrogen drugs as well as associations of *TP53* mutations with sensitivity to antiestrogen drugs, which were replicated in the MBC cohort and/or in other smaller-scale studies.

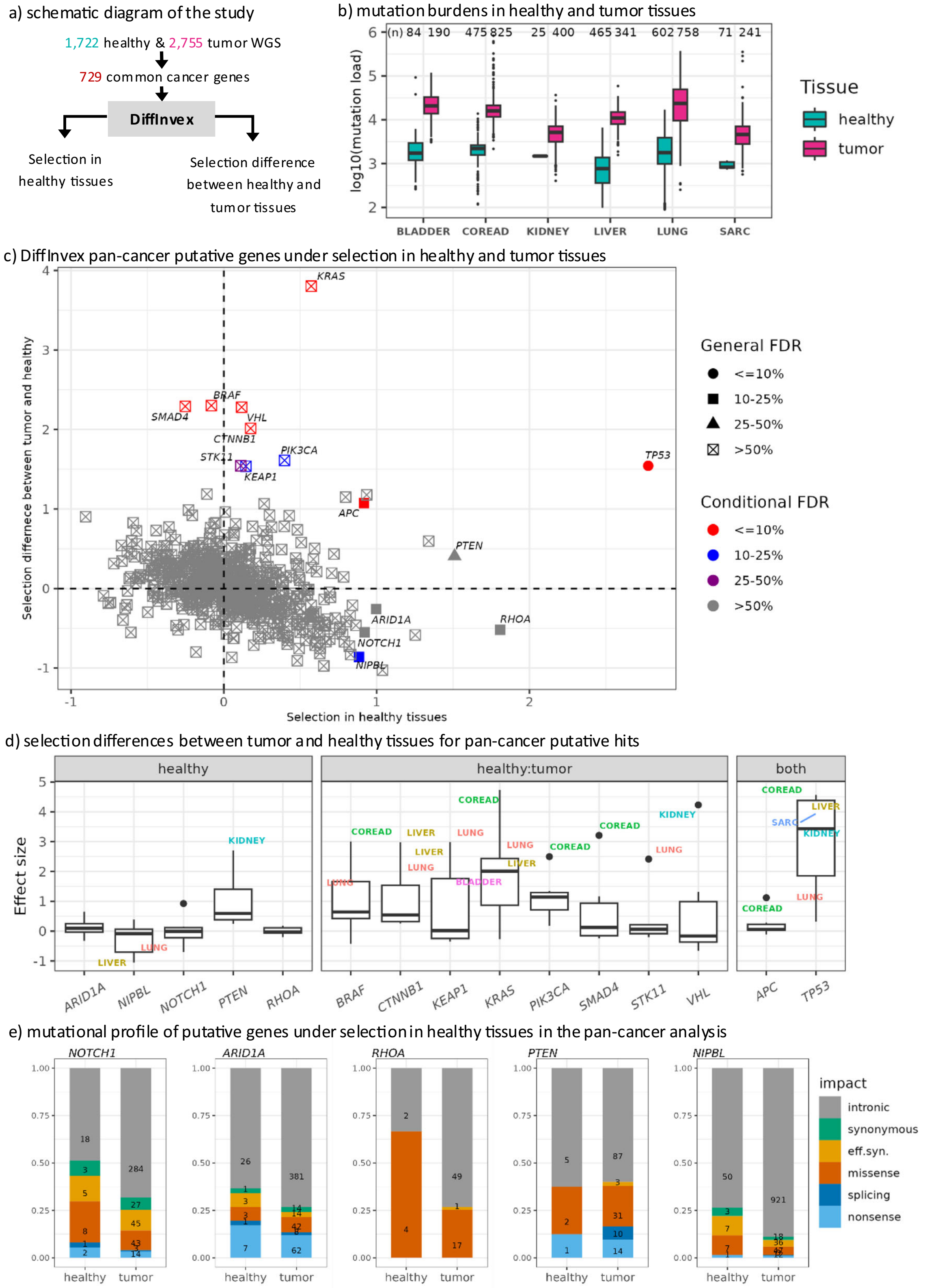
Further, smaller-scale studies provide support for some identified hits in colon and prostate cancer. For prostate cancer, Grasso et al.[37] reported that *KDM6A* was altered in pre-treated metastatic castration-resistant prostate cancer (CRPC) with 3 copy number gains, 2 losses, and 1 point mutations in CRPC compared to 0 alterations in treatment-naive prostate cancer. Using cell lines, Das et al.[38] showed that CRC with *APC* mutations developed resistance to 5-FU (pyrimidine analog by our categorization). Using CRC patients and cell lines, two studies suggested that loss of *SMAD4* in CRC patients induces resistance to 5-FU-based therapy[39,40], apparently consistent with our association of *SMAD4* selected mutations with pyrimidine analogs pre-treatment.

Joint analysis of selection pressures on driver genes in healthy and cancerous tissue genomes

The DiffInvex framework can be applied to study different somatic evolutionary scenarios where selection differentials are of interest. A salient question concerns the evolution in healthy somatic cells that, while not appearing transformed, may nonetheless undergo clonal expansions driven by mutations. Such mutations would not be considered cancer drivers in a strict sense, as the malignant transformation can occur independently of them, or even despite them (reviewed in Acha-Sagredo et al.[41], Herms and Jones[42], and Fiala and Diamandis[43]). The latter case, implying a protective role, is probable with *NOTCH1* gene mutations. These are found more commonly in the apparently healthy cells of the human esophagus that have undergone clonal expansion, than in the esophageal cancer[44], and similarly *NOTCH1* is commonly mutated in healthy skin but the incidence of mutations does not appear higher in non-melanoma skin cancers[45]. The presence of certain mutations in cancers may simply be a remnant of the non-tumoral stages of somatic evolution, therefore arguing against their role in tumorigenesis. There is a rising availability of WGS data from various healthy human tissues, facilitating a systematic, pan-tissue analysis of driver gene potential in healthy human cells. We applied DiffInvex to 1722 WGS of healthy somatic cells from lung, liver, colon, muscle, fat, kidney, and bladder (see Methods for data sources), comparing them to 2755 WGS of matched cancer types to test differential selection (Fig. 5a). We considered a permissive set of 729 cancer genes collected from the CGC and the TCGA pan-cancer analysis catalog[46], asking whether the previously-observed genomic signatures of selection on these genes may have derived in part or fully from noncancerous somatic evolution.

As anticipated, mutation burdens in the cancers were considerably higher than those in healthy cell WGS (Fig. 5b), likely reflecting increased amounts of evolutionary time and/or mutator phenotypes in tumors. DiffInvex accounts for these differences, as well as any differences in mutation spectra and regional mutation rates that might have occurred between cancer and normal. Quantile-quantile plots for the 729 cancer genes indicated some deflation (Supplementary Fig. 22a), suggesting a conservative bias of the test. These cancer genes showed a shift to positive selection in both healthy and tumor tissues, compared to passenger genes (Supplementary Fig. 22b). We identified 6 genes under positive selection in healthy tissues at FDR < 25% (7 genes at FDR < 50%; x axis in Fig. 5c), and 9 genes under conditional positive selection i.e. increased in cancer versus normal tissues at FDR < 25% (10 genes at FDR < 50%; y axis in Fig. 5c) (Supplementary Data 6). This modest number of hits is likely to result from a limiting amount of healthy tissue WGS data, and the low mutation burden therein (Fig. 5b), constraining the statistical power of tests for somatic selection.

This multi-tissue analysis of differences in selection between healthy and tumoral tissues reveals that some prominent driver genes are indeed clearly under preferential selection in cancer compared to normal: *KRAS*, *BRAF*, *SMAD4*, *VHL*, *CTNNB1*, *PIK3CA*, *STK11*, *KEAP1*, *TP53*, and *APC* (Fig. 5c). The tissue distribution of the conditional-selection effects in our analysis broadly matches known tissue spectra for these genes (Fig. 5d), validating our methodology. All these genes, except the latter two, were not under significant positive selection in healthy tissues (and *APC* only modestly so, Fig. 5c) in the pan-tissue analysis. The breakdown per tissue of the first 8 genes yields no significant hits in healthy genomes, even at a permissive threshold of nominal $p < 0.05$ (Supplementary Fig. 23a middle panel; we note distributions are slightly shifted towards positive values for some genes). This means that genomic signatures of selection in these bona fide driver genes are particular to the



cancerous transformation, and they largely do not drive clonal expansion in healthy tissues. The central tumor suppressor gene *TP53* was selected both in healthy cells, and it additionally exhibited additional differential selection in cancerous cells (Fig. 5d and Supplementary Fig. 23a), suggesting that *TP53* mutations may either transform cells, or result in expansion of apparently normal cells, depending on context.

Genes exhibiting signatures of positive selection across healthy somatic cell types

Importantly, the analysis revealed 4 genes (*NOTCH1, ARID1A, RHOA, and NIPBL*) that are under significant positive selection in healthy genomes at FDR < 25% (Supplementary Fig. 23a left-panel), but do not exhibit significant differential positive selection in cancer genomes (i.e., their signatures of positive selection are not significantly stronger

Fig. 5 | Difference in positive selection on driver genes between healthy and tumor tissues. a Schematic diagram of the study for testing common genes under selection in healthy and tumor tissues using 1722 healthy and 2755 tumor WGS. **b** Differences in the tumor mutation burden (here, number of point mutations per sample) between the healthy and tumor tissues. COREAD is colorectal cancer, matched with colon healthy cells; SARC is sarcoma cancer matched with muscle and fat healthy cell WGS. For healthy kidney cells, the WGS are shown as a single point, as patient IDs were not available. **c** Genes under selection in healthy tissues and under differential selection between healthy and tumor tissues identified in a pan-tissue analysis of 1722 healthy and 2755 tumor WGS. Gene names were shown for genes under selection in healthy (general FDR < 25%, effect size on x axis) or genes under differential selection between healthy and tumor tissues (conditional FDR < 25%, effect size on y axis), based on the DiffInvex test. Obtaining FDR via DiffInvex statistical test is described in Methods. **d** DiffInvex effect sizes of the difference between healthy and tumor tissues, from analyses in individual tissues (listed with acronyms next to corresponding points). Genes that were significant in the pan-cancer analysis are shown, grouped based on the significant of positive selection in healthy tissue (left panel), tumor tissue (middle panel) and both (right panel). Tumor names are shown for tumor types with p-value < 0.05. In **b** and **d**, the horizontal line within the box indicates the median, while the lower and upper hinges (bounds) of the box represent the first (25th percentile) and third quartiles (75th percentile), respectively. **e** Mutation functional impact profile of genes putatively under positive selection in healthy tissues, compared between healthy and tumor samples. The effectively synonymous ("eff.syn.", in text abbreviated as "ES") mutations here are the missense mutations with low AlphaMissense score (Methods). Source data are provided as a Source Data file.

in cancer compared to normal cell genomes). As a validation, this includes the known example of the *NOTCH1* gene reported in esophagus and skin cells[44,45]; importantly, our multi-tissue WGS dataset does not include esophagus nor skin, indicating that *NOTCH1* mutations are associated predominantly with healthy somatic evolution across many human tissues. A further notable gene is the commonly-mutated gene *ARID1A*. Its selection effect size in healthy tissues was positive (significant at FDR < 25%), while the differential tumor-vs-healthy selection was negative (n.s., Fig. 5c), suggesting there is not tumor-specific selection for *ARID1A* in the tissues we analyzed. This finding is consistent with previous reports, where mutations in *ARID1A* were identified in somatic evolution of noncancerous liver cells, and are protective against liver injury in a mouse model[47]. *ARID1A* is also recurrently somatically mutated in endometriosis[48], a common and benign condition that rarely progresses to cancer.

In addition to *NOTCH1* and *ARID1A*, we also identified the less-common driver genes *RHOA* and *NIPBL* as under positive selection in healthy cells at FDR < 25% (Fig. 5c). Their differential selection in tumors is not positive and in fact tends towards negative values (n.s. for *RHOA*; Fig. 5c) suggesting these genes would also not be bona fide cancer drivers, but instead their role is restricted to noncancerous somatic evolution. The break-down of differential selection per tissue reveals that no tissues are significant (at nominal $p < 0.05$), and the distributions of effect sizes across tissues narrowly spread around 0 for the tumor-differential selection (Fig. 5d; *NIPBL* is shifted towards negative values). The final example is the common tumor suppressor gene *PTEN*, where we suggest substantial selection effect size in normal cells (however, at permissive FDR < 50%). Additionally, we see differential positive selection in *PTEN* in cancer as expected of a bona fide driver; however, intriguingly with modest effect sizes (n.s., Fig. 5d). This latter result should be interpreted cautiously, given that our analysis did not include the uterus and brain healthy tissues/cancers, where *PTEN* mutations are known to be more common drivers.

We also considered, as a supplementary test to the DiffInvex test, the functional impact of exonic mutations, based on AlphaMissense scores, and whether it differs in these genes between healthy cell WGS and cancer WGS. Contrasting the mutational profile of the healthy-associated genes (*NOTCH1*, *ARID1A*, *RHOA*, *PTEN* and *NIPBL*) between healthy and tumor samples (Fig. 5e and Supplementary Fig. 23b) revealed various pathogenic mutations in healthy cells. For *NOTCH1*, the ratio of high-impact to low-impact (ES, "effectively synonymous") mutations was somewhat higher in healthy tissues (1.38) compared to tumor tissues (0.83) (Fig. 5e), supporting that mutations in *NOTCH1* are protective against cancer. For *ARID1A*, we observe similar proportions of impactful exonic mutations between healthy and cancer (here, slightly shifted towards cancer; Fig. 5e). This is broadly in line with the DiffInvex results (Fig. 5c), indicating that, unlike *NOTCH1*, the damaging mutations in *ARID1A* are not protective against cancer, however also supporting that they are unlikely to be generally tumorigenic either. The *NIPBL* displayed a similar pattern, arguing against a protective role. In *RHOA* and *PTEN*, all coding exonic mutations in healthy WGS were high-impact missense or nonsense (Fig. 5e), however, as a caveat, the low numbers of coding exonic mutations in our healthy WGS data collection preclude statistical testing of the selection differential between healthy and cancer using the dNdES test. Similarly, the analysis across individual tissues (Supplementary Fig. 23b) has some suggestive results, including *NOTCH1* in the lung, but is less powered. Overall, these data support that mutations in the 5 identified genes promote clonal expansions in healthy tissues, but that their somatic selection does not predispose to cancer transformation.

Discussion

Mechanisms by which cancer evolves by changing selective pressures on driver genes are important to understand, so that this knowledge may be applied to prevent (or target) cancer evolutionary paths to evading treatment. Genome sequences of tumors that were pretreated by chemotherapies should reflect selective pressures to resist the therapy and should therefore contain resistance-conferring alterations. A rigorous comparison of genomes of tumors before versus after treatment is, in principle, a powerful tool to identify drug resistance mutations[49]. However, an important challenge are the mutation rate heterogeneity and the changes in mutation spectra caused by therapy[16,50], which means that changes in mutation burdens of individual genes could result from neutral processes. We rigorously addressed these challenges by the DiffInvex framework that draws on non-coding mutations to establish an accurate mutation risk baseline. In addition, DiffInvex accounts for confounding between various co-administered treatments, thus suggesting putative drug resistance driver mutations from a large cancer WGS dataset. As with all genomics studies, ours too is limited by statistical power to identify the less commonly mutated therapy resistance driver genes. This is in principle addressable by including additional cancer WGS data, e.g., recently generated from Genomics England with currently ~14 k WGS[51], however, the lack of genomes from pre-treated tumors in that particular study remains a bottleneck. Additional genomic data will enable testing statistical significance in individual tumor types, which may uncover the "long-tail" of drug resistance-conferring point mutations, relevant to rarer tumor types or subtypes. Additionally, our implementation of the "dN/dES" supporting test, which validates hits by testing differential functional impact, will also benefit from additional genomes and, possibly, by future algorithms for mutation effect prediction to supplant the currently employed AlphaMissense. Of note, DiffInvex does not distinguish between the cases where the therapy exposure caused the mutation, or the mutation was pre-existing at a low clonal frequency and then selected by therapy. Future developments of related methods might use multiregion sequencing or simply higher-depth WGS to establish clonal versus subclonal status of mutations, which can be analyzed by DiffInvex as a covariate, thus shedding light on this issue.

We acknowledge an important limitation, is that drug resistance in tumors may in many cases evolve not via point mutations in coding gene regions, but instead via noncoding mutations, copy number alterations, and/or epigenetic changes, which DiffInvex does not currently model. Indeed, our results suggest that point mutations are only one type of mechanism generating drug resistance in tumors: in addition to the well-known resistance drivers to targeted therapies (*EGFR*, *ESR1*, *AR*, and *KIT* mutations), other chemoresistance drivers we identified have lower effect sizes. Moreover, they are largely mutations in general cancer driver genes like *PIK3CA*, *APC*, *MAP3K1*, *SMAD4*, *STK11* and *MAP2K4*, rather than mutations in genes particular to drug resistance phenotypes (we do not exclude these exist, e.g. *TFPT* is a tentative hit that awaits replication in larger cohorts; its association with platinum resistance seems plausible given its roles in DNA nucleotide excision repair). Thus, cancer seems to more commonly evolve drug resistance by increasing the activity of known oncogenic pathways, thus gaining fitness, rather than employing specialized mechanisms such as altering drug target proteins, at least as far as point mutations are concerned. Regardless of the mechanism granting resistance, the mutated genes linked with chemoresistance and validated in additional cohorts are nonetheless useful as markers for stratifying patients for therapy in a genomically-informed manner.

The relative rareness of negative conditional selection (i.e. sensitizing) associations in our analysis – 4 detected at a permissive FDR, with only 1 of those (*FBXW7*) validating in the dNdES test – might stem from limited power, requiring a fairly commonly mutated gene in untreated tumors to begin with. Another reason could be that DiffInvex can detect only those cases of sensitizing mutations that are overcome during evolution, specifically by reverting that mutation or deleting the allele containing it. Removal of sensitizing mutations may not be the common mechanism of acquiring resistance post-treatment; resistance may instead be gained by events that circumvent the effects of the sensitizing mutation by altering other genes, while the original mutation remains in the genome, which prevents it from being detected by DiffInvex.

In addition to studying tumor evolution under therapy, we demonstrated another application of our DiffInvex method to the rapidly-growing datasets of healthy cell WGS, identifying changes in selection occurring during the transformation from normal to cancerous cells. Here, too, the ability for DiffInvex to control for shifts in mutational spectra is relevant, since certain mutational processes are common in cancerous cells and rarer in normal ones, e.g. APOBEC mutagenesis and the SBS17 mutational process[52]. We were motivated by the known example of the *NOTCH1* mutations that appeared particular to the expanded, but phenotypically normal cells[44] and are under increased selection in the normal esophagus[53]. Building on this, our analysis further identified *ARID1A* and several other genes under significant positive selection in normal cells that does not increase in tumoral cells across various tissues. This argues that the mutations observed in these genes in cancer genomes may be a remnant of the normal somatic evolution stage, rather than causal contributors to carcinogenesis. Some of these “normal drivers” might even be somewhat protective against further cancer development, countering the fitness benefits from future oncogenic mutations. A caveat in this analysis is that healthy cell WGS data is currently limited in which tissues are represented, urging caution in interpreting inferences for individual genes. This concerns the *PTEN* example that was predominantly associated with the healthy state in our current analysis, which does not include brain or uterus healthy WGS, two tissues where tumors have *PTEN* as a common driver. We also acknowledge that, in principle, additional genes might be under selection in the healthy state that were not previously reported as cancer driver genes in any context, e.g. because the mutations therein are strongly protective against transformation, a possibility that we did not address here as it was reported as uncommon[41]. Going forward, we foresee that DiffInvex or related frameworks could be successfully applied to modeling cancer evolution in comparing precancerous lesions with tumors, using multiregion sampling of tumors or single-cell WGS, or from longitudinal sampling of cancer genomes before versus after therapy. This will identify fitness effects conferred by specific genes in various phases of (pre)cancer development, improving our understanding of how somatic cells evolve, and revealing additional avenues for therapeutic intervention.

Methods

Data collection and processing

We collected WGS somatic mutation data from nine different studies (Supplementary Data 1). First, we obtained somatic SNVs for 4858 WGS from the Hartwig Medical Foundation study (https://www.hartwigmedicalfoundation.nl/en/)[3]. Second, we downloaded somatic SNVs from 2808 WGS from the Pan-cancer Analysis of Whole Genomes (PCAWG) study that were re-processed by the Hartwig pipeline at the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/releases/PCAWG/Hartwig)[49]. Third, we downloaded somatic SNVs from 570 WGS from the Personal Oncogenomics (POG) project from BC Cancer (https://www.bcgsc.ca/downloads/POG570/)[5]. Fourth, we obtained somatic SNVs for 724 WGS from the DECIDER study (https://www.deciderproject.eu/)[6]. Fifth, we downloaded somatic SNVs for 133 ovarian WGS from the British Columbia Ovarian Cancer Research Program (herein referred to as OVCARE)[54]. Sixth, we downloaded somatic SNVs from the breast and prostate projects that were not included in the PCAWG study from the ICGC data portal (https://dcc.icgc.org/releases/release_25/Projects). These ICGC samples included 352 samples from BRCA-EU project[55], 183 samples from PRAD-CA project, 91 samples from PRAD-UK project, and 152 samples from EOPC-DE project. We also downloaded read alignments (BAM files) of WGS samples for 907 tumors from the MMRF COMMPASS project[56] and for 548 tumors from the Clinical Proteomic Tumor Analysis Consortium 3 (CPTAC-3) project[57] from the GDC data portal (https://portal.gdc.cancer.gov/). Finally, we downloaded the BAM files for 610 samples from the Mutographs project[58] from the European Genome-Phenome archive (https://ega-archive.org/datasets/EGAD00001006732). Somatic variants of MMRF-COMMPASS, CPTAC-3 and Mutographs projects were called using Strelka2 caller with SomaticEVS score ≥6[59]. Then, we performed a liftOver of mutation calls from the GRCh38 to the hg19 reference genomes for those samples, as well as samples from the DECIDER project.

We filtered these tumor WGS to exclude tumor samples with missing treatment or tumor stage information. Then, we also excluded samples with tumor purity <20%, and ultramutated samples from microsatellite instability (MSI) tumors and colorectal tumors with POLE mutations, and finally also the problematic Hartwig and PCAWG samples that were reported in earlier studies using the same data[20,49]. For the PCAWG cohort, we excluded samples tagged with PCAWG blacklist, No pipeline output, Corrupt HMF pipeline 5 run, No PURPLE 2.53 or LINX 1.14 output, and SV INV outlier. In the Hartwig cohort, we excluded one sample with <50 SNVs and 108 samples from unknown primary sites. Additionally, we filtered out one duplicate PCAWG patient (DO217844) that was also included in the Hartwig cohort. In the MUTOGRAPH cohort, we additionally filtered out 7 ESCA and 2 COREAD ultramutated samples with > 1 million SNVs. From the CPTAC3 cohort, we kept 336 samples from NSCLC, PAAD, RCC, and UCEC (out of 436 samples with purity >= 20%). Finally, we selected one sample or one time point per patient for DECIDER and Hartwig patients with multiple tumor samples. In the case of DECIDER cohort, we kept one time point per patient by merging samples from different sites, removing redundant mutations, at the most recent biopsy time point, resulting in 202 genomes from DECIDER. In the Hartwig cohort, if available, we kept one pre-treated sample per patient.

For the validation datasets, we obtained the mutation calls of the panel sequencing data and their well-curated clinical data from 25,040 tumor samples from the MSK-CHORD cohort (https://www.cbioportal.org/study/summary?id=msk_chord_2024). We also obtained the somatic SNVs of panel sequencing data from 2004 COREAD samples (https://www.synapse.org/#!Synapse:syn30991602) and 1551 NSCLC samples (https://www.synapse.org/#!Synapse:syn27056179) from the GENIE Biopharma Collaborative (BPC) project (https://genie.cbioportal.org/)[7]. Additionally, we downloaded somatic SNVs of whole-exome sequencing data from 379 BRCA samples from the metastatic breast cancer genomics project (https://www.cbioportal.org/study/clinicalData?id=brca_mbcproject_2022)[28]. In these cohorts, we kept one sample per patient, at the most recent biopsy time point. Additionally, in the MSK-CHORD colorectal cohort, we kept the three main groups (Colorectal Adenocarcinoma: COREAD, Colon Adenocarcinoma: Colon and Rectal Adenocarcinoma: Rectal) based on the "Cancer Type Detailed" field of the clinical information. This resulted in 22,914 MSK-CHORD samples, including 7238 non-small-cell lung, 5300 colorectal, 4851 breast, 2898 pancreatic and 2627 prostate tumor samples (counting both treatment-naive and pre-treated samples). In the GENIE cohort, similarly we kept one sample per patient, at the most recent biopsy time point, resulting in 1747 NSCLC samples (413 pre-treated and 1334 treatment-naive) and 1469 COREAD samples (468 pre-treated and 1001 treatment-naive). For the BRCA longitudinal cohort[28], we excluded blood samples and tissue samples with missing treatment information. For patients with more than one sample, we kept only the most recent sample. This resulted in 109 treatment-naive and 77 pre-treated BRCA samples in the validation dataset.

For the datasets of healthy tissues, we obtained the mutation calls of WGS for 6 tissues from different studies including somatic SNVs of liver tissue from Brunner et al.[60], lung tissue from Yoshida et al.[61], colon from Lee-Six et al.[62], liver and colon from Blokzijl et al.[63], fat, muscle, and kidney from Franco et al.[52,64]. and bladder from Lawson et al.[65]. We filtered out these mutations to remove duplicated mutations per donor as there were many patients with multiple samples. We also excluded samples with less than 100 SNVs. Finally, we filtered out indels. This resulted in 3,745,713 mutations from 1722 samples, including: 321,276 mutations from 84 bladder samples, 1,126,055 mutations from 475 colon samples, 37,078 mutations from 25 kidney samples, 491,568 mutations from 465 liver samples, 1,705,896 mutations from 602 lung samples and 63,840 mutations from 71 fat and muscle samples (considered jointly for comparison with the sarcoma cancer type).

Baseline for neutral mutation rate estimation

To estimate the neutral mutation rate in protein-coding sequence (CDS) regions, we utilized the mutations in their neighboring regions including introns, untranslated regions (UTRs), and upstream and downstream gene flanking regions. We utilized the Ensembl annotation package (EnsDb.Hsapiens.v75) to get the coordinates of genomic regions (CDSs, exons, introns, and UTRs) in the hg19 reference genome. Starting with the HGNC symbol of the gene, we selected the most-expressed transcript[66] if available. Otherwise, we chose the longest transcript. We then defined the target regions (regions of interest) as the CDSs and the adjacent five nucleotides of each intron (to account for splice site disrupting mutations), while the background regions included intronic, UTRs, and intergenic regions, but excluding the coding exonic regions of neighbor genes. In the case when the total intronic regions were shorter than the minimum required background region size (user-defined parameter, default value = 50,000 nucleotides), we included adjacent genomic loci from the upstream and downstream intergenic regions into the background.

We implemented various filters to exclude genomic loci that could confound the estimation of neutral mutation rate. This includes microsatellite repeats of 6 or more nucleotides, CTCF and cohesin (RAD21) overlapping binding sites, the ENCODE blacklist of problematic regions of the genome[67], non-uniquely mappable sites according to the Umap k100 alignability track (https://bismap.hoffmanlab.org/)[68], target regions of Activation-induced cytidine deaminase (AID) somatic hypermutation process (see Supek et al.[69]. and references therein), and promoter hypermutated sites (NYTTCCG motifs within promoters)[70]. Finally, we filtered out the loci in the background regions bearing hotspots that have 3 or more mutations in our cohort, suggestive of locally high mutation risk and/or technical artefacts.

Next, to minimize the variation in mutation risk arising from differing mutational signatures between the target regions and background regions, we employed a locus sampling approach to match their trinucleotide or pentanucleotide (the default) compositions as well as DNA methylation status, within each gene. We iteratively removed nucleotide positions from the background regions to increase similarity in composition to the target regions, until reaching a tolerance <0.01 (Euclidean distance between the relative trinucleotide/pentanucleotide frequencies in the DNA sequence composition of each region).

The DiffInvex test for somatic selection and conditional selection

We outline the rationale underlying the DiffInvex test below.

- The use of the intronic mutation rates as a neutral baseline to infer positive selection signatures and accurately identify driver genes was established in the InVEx test[13]. The InVEx test provided the rationale that exonic mutation rates, normalizing for background mutagenesis modeled from intronic rates, are an estimate of driver potential. Here, we build on the InVEx principle, extending its use to being able to test for conditional (differential) selection, where two or more conditions are compared to assess differences in selection strength on a gene. Within the DiffInvex framework, this test also stringently controls for changes in global mutation rates and spectra/signatures, as well as for any confounding factor of interest (e.g. cancer type).
- The rationale underlying the application of the DiffInvex test to identifying mutational drivers of cancer treatment resistance or sensitivity, is that various cases of drug resistance in cancer are known to be acquired via point mutations in coding regions of genes (see positive controls, "Tier 1a" genes such as *ESR1* and *EGFR* gene mutations). Our analysis systematically searches for additional examples of this type of occurrence in a large dataset of WGS-sequenced tumors.
- The limitations of the DiffInvex analyses include that it does not address cases of drug sensitivity/resistance that occur by non-coding point mutations (e.g. promoter or enhancer), except intronic splice-site, which is covered. As a second limitation, we do not address cases of drug sensitivity/resistance that occur by somatic copy number alterations and other structural variants, nor by epigenetic changes. Copy number alterations and epigenetics are likely to constitute important drug resistance mechanisms to be addressed by future methods.
- The DiffInvex implementation encompasses two tests. Firstly, the DiffInvex test (also called "dEXdIN"), which contrasts exonic coding mutation rates, against the mutation rate in intronic regions (excl. splice sites) and flanking intergenic regions, assumed to be neutral. Secondly, as an independent validation method, we provide the "dNdES" test, which tests the ratio of functional to non-functional coding exonic variants ("ES" stands for "effectively synonymous", here consisting of synonymous plus the missense mutations that were determined as neutral by AlphaMissense functional impact scores).

Implementation of the DiffInvex test via a regularized Poisson regression

After defining the target and background regions for each gene, we utilized Poisson regression to model the mutation counts in each region and to determine the effect estimates of conditional selection using the following model:

$$\log(\#\text{mutations}) \sim B_0 + B_1 * \text{isTarget} + \sum\nolimits_{i\in \text{confounders}} B_i * \text{confounder}_\text{i} + \sum\nolimits_{j\in \text{drugs}} B_j * \text{drug}_\text{j} + \sum\nolimits_{k\in drugs} B_k * \text{isTarget} : \text{drug}_\text{k} + \log(r) \tag{1}$$

where the regression coefficient B_0 represents the base mutation rate and is included as the intercept of the model. The isTarget is a binary indicator variable to distinguish mutations in target regions (isTarget = 1), which are being tested for selection from the mutations in the background regions (isTarget = 0), which represent the neutrally accumulated mutations. The corresponding regression coefficient B_1 represents the natural log fold change of mutation rates in the target regions (CDS and splicing sites), compared to mutation rates in the background regions (introns, UTRs and, optionally, intergenic regions). Therefore, for each gene, positive B_1 indicates that the gene is under positive selection in cancer, as there is enrichment of mutations in target regions compared to the baseline (background regions), while negative B_1 suggests that it is under negative selection in cancer, as there is mutation depletion in target regions. The differences in trinucleotide or pentanucleotide compositions between the two regions are controlled by locus sampling to remove parts of the background region (see above).

The confounder$_i$ represents the variables that could confound our analyses and should be controlled for, and the regression coefficients B_i reflect their effects (for binary variables, the fold-difference in mutation risk between the 2 levels of the variable, on the natural log scale). In our analyses, we adjusted for several confounders, including tumor type in the pan-cancer study, the source study, the tumor stage (primary or metastatic), and the patient sex. Similarly, the regression variable drug$_\text{j}$ represents the treatments that were given to different patients in our cohort (drug$_\text{j}$ = 1 meaning this drug j was given to patients prior to treatment, and 0 if not given prior to therapy) and the coefficient B_j reflects their effects on the background mutation rate (on the natural log scale).

The terms isTarget : drug$_\text{k}$ represent the interaction terms of the target-region indicator variable isTarget with the given treatment indicator variable (drug$_k$); the corresponding regression coefficient B_k reflects the conditional selection effects by estimating the difference in the natural log fold change of target region to background region mutation rates between the two conditions of the drug$_k$ (0: drug treatment was absent, 1: drug treatment was given). So, for drug$_k$, positive B_k indicates that the tested gene is a putative drug-resistance gene, as there is an increase of log ratio of target-to-background region mutations when patients who were pre-treated with that drug (drug$_k$ = 1) compared to patients untreated with that drug (drug$_k$ = 0). In contrast, negative B_k indicates that the tested gene is a putative drug-sensitivity gene as there is a decrease of log ratio of target-to-background region mutations when patients were pre-treated with that drug (drug$_k$ = 1) compared to patients untreated with that drug (drug$_k$ = 0).

Finally, we include the number of nucleotides-at-risk r as an offset in the regression model to normalize mutation counts to relative mutation rates (per nucleotide per sample). This allows DiffInvex to account for variations in DNA length of the target and the background regions (after tri/pentanucleotide sampling), as well as the number of samples of each condition (e.g. number of primary versus metastatic samples, or number of samples pre-treated by one drug versus untreated ones). The adjustment for nucleotides-at-risk r ensures that the effect sizes (coefficients) are not biased by gene length or nucleotide composition, however, we note that longer genes naturally provide greater statistical power to detect significant selection, because the standard error of the estimates will be reduced in longer genes.

We implemented the regression model as a generalized linear model with a log link function, further regularizing the regression coefficients by using a weakly informative prior to stabilize estimates from sparse data[71] (using the R function bayesglm, arm package version 1.11_2), as applied to cancer genomes in Besedina et al.[72]. The statistical test applied on the regression coefficients (log enrichments) was the Z-test, two-tailed, as implemented in the R function summary().

Multiple testing correction was performed using the FDR method separately for the selection coefficients (B_1) and for the conditional selection coefficients (B_k). FDR correction for the selection coefficients were performed on all protein-coding genes. Then, we used the putative genes under selection in cancer (selection FDR < 25%) and known cancer genes (TCGA and CGC) for the FDR correction for the conditional selection coefficients and downstream analyses. In our analyses, we applied the DiffInvex model in two scenarios to obtain the conditional selection effects of each gene-drug interaction. In the first scenario, we tested the associations between gene mutations and each drug type individually, without controlling for the conjoint drug treatment problem:

$$\log(\#\text{mutations}) \sim B_0 + B_1 * \text{isTarget} + B_2 * \text{tumorType} + B_3 * \text{cohort} + B_4 * \text{isMetastatic} + B_5 * \text{Sex} + B_6 * \text{drug} + B_7 * \text{isTarget} : \text{drug} + \log(r) \tag{2}$$

where drug represents the drug type to be tested (e.g. platinum drugs) and B_7 reflects its conditional selection effect (in the natural log scale). Here, for each gene, we applied the DiffInvex model 15 times (one for each drug type). Results for this scenario were shown in Supplementary Fig. 6a, and they likely result in many spurious associations.

Therefore, we devised the second testing scenario, where we tested the associations between gene mutations and all drug types in the same model to control for the conjoint drug treatment problem:

$$\log(\#\text{mutations}) \sim B_0 + B_1 * \text{isTarget} + B_2 * \text{tumorType} + B_3 * \text{cohort} + B_4 * \text{isMetastatic} + B_5 * \text{Sex} + \sum_{j\in \text{drugs}} B_j * \text{drug}_j + \sum_{k\in \text{drugs}} B_k * \text{isTarget} : \text{drug}_k + \log(\text{r}) \tag{3}$$

where B_k reflects its conditional selection effect (in the natural log scale) of the drug type drug$_k$. We applied the DiffInvex model once for each gene, and results for this scenario were shown in Fig. 3a for the pan-cancer analysis, and Fig. 3b shows the coefficients for the selected associations in the cancer type-specific analyses. We evaluated the goodness-of-fit of the regression models using the pseudo-R^2 based on deviance (R^2 deviance)[73].

The dNdES statistical test for conditional mutation impact on coding regions

We also implemented the dNdES test, to test if a putative DiffInvex gene (either drug resistance-associated or drug sensitizing) is under conditional selection upon drug treatment, using a different test on the mutations that does not rely on the intronic baseline, and therefore provides a validation method for DiffInvex. The dNdES functional impact test compares the frequency of nonsynonymous "N" (high-impact missense, nonsense, splicing) to effectively synonymous "ES" (low-impact missense and synonymous) exonic mutations. For that, we first annotated the exonic mutations using Annovar tool[74] and AlphaMissense method[25]. The AlphaMissense provides a score for missense mutation pathogenicity. We defined the AlphaMissense score = 0.39 as the threshold for dividing missense mutations into high-impact and low-impact missense mutations. This threshold resulted in approximately unity slope in the linear regression model between dEXdIN (exonic to intronic mutations) and dNdES (nonsynonymous to effectively synonymous mutations, defined by AlphaMissense) of cancer genes as shown in Supplementary Fig. 18c. We used this threshold for annotating the missense mutations from the WGS data (discovery cohort) and from the panel and whole-exome sequencing data (validation cohort).

We again utilized Poisson regression to model the mutation counts in exonic regions (target, background) and to determine the effect estimates of selection and conditional selection, for each drug separately, using the following model:

$$\begin{aligned}\log(\#\text{mutations}) \sim & B_0 + B_1 * \text{isTarget} + B_2 * \text{tumorType} + B_3 * \text{context} \\ & + B_4 * \text{drug} + B_5 * \text{isTarget} : \text{drug} + \log(\text{r})\end{aligned} \tag{4}$$

where regression coefficient B_0 represents the base mutation rate and is included as the intercept of the model. Regression coefficient isTarget is a target variable to distinguish mutations in target regions (isTarget = 1) that is currently being tested for selection and mutations in the background regions (isTarget = 0) that represent the neutrally accumulated mutations. In this test, we further classified the exonic sites (importantly, not the mutations), into target regions (sites with 2 or 3 of their possible mutations are nonsynonymous with AlphaMissense >= 0.39) and background regions (sites with 0 or 1 of their possible mutations are nonsynonymous with AlphaMissense >= 0.39). The regression coefficient B_1 represents the natural log fold difference of mutation rates in the target regions, compared to mutation rates in the background regions. The tumorType variable represents the tumor type to control for in the pan-cancer study and B_2 reflects its effect (natural log fold difference, relative to one arbitrarily chosen tumor type). Similarly, the context variable represents the trinucleotide context (e.g. CCA, CCC, CCG, CCT, ...) of mutations that could confound our analyses and B_3 reflects its effect (natural log fold difference, relative to one arbitrarily chosen trinucleotide). The variable drug, the interaction term isTarget : drug and the offset r are explained in the DiffInvex test above. We implemented the regression model as a generalized linear model with a log link function (using the R function glm, stats package). The statistical test applied on the regression coefficients (log enrichments) was the Z-test, two-tailed, as implemented in the R function summary(). False discovery rate (FDR) correction was applied to account for multiple comparisons.An overview of main differences between DiffInvex and other methods for assessing conditional selection in somatic genomes, cancerEffectSizeR and Coselens, is provided as Supplementary Note 2.

Test for differential somatic selection between healthy and tumor tissues

We again utilized DiffInvex for testing somatic selection in healthy tissues and the difference in selection between healthy and tumor tissues. We used 1722 WGS of 6 healthy tissues and 2755 WGS of their matched cancer types. We excluded mutations in chromosomes X and Y in this analysis as there were not available for some studies of healthy tissues[52,63,64]. Here we applied the DiffInvex model to obtain the general selection and conditional selection effect sizes of each gene that represent the selection in healthy tissues and the difference between selection in healthy and tumor tissues, respectively:

$$\begin{aligned}\log(\#\text{mutations}) \sim & B_0 + B_1 * \text{isTarget} + B_2 * \text{tissue} \\ & + B_3 * \text{isCancer} + B_4 * \text{isTarget} : \text{isCancer} + \log(r)\end{aligned} \tag{5}$$

where isCancer is a binary indicator variable to distinguish between healthy tissues (isCancer = 0) and tumor tissues (isCancer = 1). The regression coefficient B_1 represents the selection in healthy tissues (in the natural log scale), and the regression coefficient B_4 reflects the conditional selection effect, the selection difference between healthy and tumor tissues (in the natural log scale).

Reporting summary

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

Data availability

In this study, published tumor and healthy datasets were reanalyzed. We obtained the WGS somatic mutation calls for the Hartwig Medical Foundation study [https://www.hartwigmedicalfoundation.nl/en/] under request DR-260 (Hartwig data is available under restricted access via https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/), the PCAWG study re-processed by the HMF pipeline at the International Cancer Genome Consortium (ICGC) data portal [https://dcc.icgc.org/releases/PCAWG/Hartwig], the Personal Oncogenomics (POG) project from BC Cancer [https://www.bcgsc.ca/downloads/POG570/], the DECIDER data from within a collaborative project [https://www.deciderproject.eu/] also available under restricted access via EGA accession number EGAS00001006775, the breast, prostate and ovarian WGS projects (BRCA-EU, PRAD-CA, PRAD-UK, EOPC-DE and OVCARE) from the ICGC data portal [https://dcc.icgc.org/releases/release_25/Projects] and the British Columbia Ovarian Cancer Research Program under accession number EGAS00001002390. We downloaded bam files for CPTAC-3 project (available under restricted access via dbGaP accession phs001287.v17.p6 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001287.v17.p6]), and MMRF-COMMPASS project (available under restricted access via dbGaP accession phs000748.v7.p4 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000748.v7.p4]) from the NCI Genomic Data Commons data repository [https://portal.gdc.cancer.gov/]. Mutographs project is at the European Genome-Phenome archive EGAD00001006732 [https://ega-archive.org/datasets/EGAD00001006732]. We downloaded the somatic mutations of panel sequencing data for the MSK-CHORD study from cBioPortal [https://www.cbioportal.org/study/summary?id=msk_chord_2024] (available under the Creative Commons BY-NC-ND 4.0 licence), panel sequencing data for the GENIE study [https://www.synapse.org/Synapse:syn27056172] and somatic mutations of WES data for the metastatic breast cancer genomics project from cBioPortal [https://www.cbioportal.org/study/clinicalData?id=brca_mbcproject_2022], dbGaP accession phs001709 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001709.v2.p2]. We downloaded the curated WGS somatic mutation calls for healthy datasets for Brunner et al. study [Mendeley data identifier: https://doi.org/10.17632/ktx7jp8sch.1] (BAM files are available with accession number EGAD00001004578), Yoshida et al. study [Mendeley data identifier: https://doi.org/10.17632/b53h2kwpyy.2], Lee-Six et al. study [https://github.com/HLee-Six/colon_microbiopsies/tree/master/files_added_post_publication] (sequencing data are available at the EGA under the accession number EGAD00001005193), Blokzijl et al. studies (mutation calls are available as supplementary tables therein), Franco et al. study (mutation calls are available as supplementary tables therein) and Lawson et al. study (mutation calls are available as supplementary tables therein). The relevant data generated in this study are provided in the Supplementary Tables or the Source Data file. Source data are provided with this paper.

Code availability

The DiffInvex computational framework is freely available at https://github.com/AISKhalil/diffinvex under the MIT license[75]. For citation of the code used in this study, please reference https://doi.org/10.5281/zenodo.15161052.

References

1. Gandhi, U. H. et al. Outcomes of patients with multiple myeloma refractory to CD38-targeted monoclonal antibody therapy. *Leukemia* **33**, 2266–2275 (2019).

2. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The different mechanisms of cancer drug resistance: a brief review. *Adv. Pharm. Bull.* **7**, 339–348 (2017).
3. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
4. ICGC/TCGA pan-cancer analysis of whole genomes consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
5. Pleasance, E. et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
6. Lahtinen, A. et al. Evolutionary states and trajectories characterized by distinct pathways stratify patients with ovarian high grade serous carcinoma. *Cancer Cell* **41**, 1103–1117.e12 (2023).
7. De Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
8. Liu, R. et al. Systematic pan-cancer analysis of mutation–treatment interactions using large real-world clinicogenomics data. *Nat. Med.* **28**, 1656–1661 (2022).
9. Iranzo, J., Gruenhagen, G., Calle-Espinosa, J. & Koonin, E. V. Pervasive conditional selection of driver mutations and modular epistasis networks in cancer. *Cell Rep.* **40**, 111272 (2022).
10. Mandell, J. D., Cannataro, V. L. & Townsend, J. P. Estimation of neutral mutation rates and quantification of somatic variant selection using cancereffectsizeR. *Cancer Res.* **83**, 500–505 (2023).
11. Cannataro, V. L., Gaffney, S. G. & Townsend, J. P. Effect sizes of somatic mutations in cancer. *JNCI J. Natl. Cancer Inst.* **110**, 1171–1177 (2018).
12. Cannataro, V. L., Mandell, J. D. & Townsend, J. P. Attribution of cancer origins to endogenous, exogenous, and preventable mutational processes. *Mol. Biol. Evol.* **39**, msac084 (2022).
13. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
14. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
15. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
16. Szikriszt, B. et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 99 (2016).
17. Szikriszt, B. et al. A comparative analysis of the mutagenicity of platinum-containing chemotherapeutic agents reveals direct and indirect mutagenic mechanisms. *Mutagenesis* **36**, 75–86 (2021).
18. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nat. Commun.* **10**, 4571 (2019).
19. Mas-Ponte, D. & Supek, F. Mutation rate heterogeneity at the subgene scale due to local DNA hypomethylation. *Nucleic Acids Res.* **52**, 4393–4408 (2024).
20. Martínez-Jiménez, F. et al. Genetic immune escape landscape in primary and metastatic cancer. *Nat. Genet.* **55**, 820–831 (2023).
21. Robinson, D. R. et al. Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat. Genet.* **45**, 1446–1451 (2013).
22. Kobayashi, S. et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
23. Serrano, C. et al. Complementary activity of tyrosine kinase inhibitors against secondary kit mutations in imatinib-resistant gastrointestinal stromal tumours. *Br. J. Cancer* **120**, 612–620 (2019).
24. Taplin, M. E. et al. Mutation of the androgen-receptor gene in metastatic androgen-independent prostate cancer. *N. Engl. J. Med.* **332**, 1393–1398 (1995).
25. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
26. Jee, J. et al. Automated real-world data integration improves cancer outcome prediction. *Nature* https://doi.org/10.1038/s41586-024-08167-5 (2024).
27. AACR Project GENIE: Biopharma Collaborative. *American Association for Cancer Research (AACR)* https://www.aacr.org/professionals/research/aacr-project-genie/bpc/.
28. Wagle, N. et al. The metastatic breast cancer (MBC) project: Accelerating translational research through direct patient engagement. *J. Clin. Oncol.* **35**, 1076–1076 (2017).
29. Lopez-Knowles, E. et al. Molecular characterisation of aromatase inhibitor-resistant advanced breast cancer: the phenotypic effect of ESR1 mutations. *Br. J. Cancer* **120**, 247–255 (2019).
30. Brady, S. W. et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* **8**, 1231 (2017).
31. Huang, D., Tang, L., Yang, F., Jin, J. & Guan, X. PIK3CA mutations contribute to fulvestrant resistance in ER-positive breast cancer. *Am. J. Transl. Res.* **11**, 6055–6065 (2019).
32. Liao, H. et al. Identification of mutation patterns and circulating tumour DNA-derived prognostic markers in advanced breast cancer patients. *J. Transl. Med.* **20**, 211 (2022).
33. 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* **25**, 392–401 (2018).
34. Du Rusquec, P., Blonz, C., Frenel, J. S. & Campone, M. Targeting the PI3K/Akt/mTOR pathway in estrogen-receptor positive HER2 negative advanced breast cancer. *Ther. Adv. Med. Oncol.* **12**, 175883592094093 (2020).
35. André, F. et al. Alpelisib for *PIK3CA* -mutated, hormone receptor–positive advanced breast cancer. *N. Engl. J. Med.* **380**, 1929–1940 (2019).
36. Ellis, M. J. et al. Phosphatidyl-inositol-3-kinase alpha catalytic subunit mutation and response to neoadjuvant endocrine therapy for estrogen receptor positive breast cancer. *Breast Cancer Res. Treat.* **119**, 379–390 (2010).
37. Grasso, C. S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
38. Das, D. et al. 5-Fluorouracil mediated anti-cancer activity in colon cancer cells is through the induction of adenomatous polyposis coli: Implication of the long-patch base excision repair pathway. *DNA Repair* **24**, 15–25 (2014).
39. Zhang, B. et al. Loss of Smad4 in colorectal cancer induces resistance to 5-fluorouracil through activating Akt pathway. *Br. J. Cancer* **110**, 946–957 (2014).
40. Papageorgis, P. et al. Smad4 inactivation promotes malignancy and drug resistance of colon cancer. *Cancer Res.* **71**, 998–1008 (2011).
41. Acha-Sagredo, A., Ganguli, P. & Ciccarelli, F. D. Somatic variation in normal tissues: friend or foe of cancer early detection? *Ann. Oncol.* **33**, 1239–1249 (2022).
42. Herms, A. & Jones, P. H. Somatic Mutations in normal tissues: New perspectives on early carcinogenesis. *Annu. Rev. Cancer Biol.* **7**, 189–205 (2023).
43. Fiala, C. & Diamandis, E. P. Mutations in normal tissues—some diagnostic and clinical implications. *BMC Med.* **18**, 283 (2020).
44. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
45. Fowler, J. C. et al. Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discov.* **11**, 340–361 (2021).
46. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
47. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621.e12 (2019).
48. Anglesio, M. S. et al. Cancer-associated mutations in endometriosis without cancer. *N. Engl. J. Med.* **376**, 1835–1848 (2017).
49. Martínez-Jiménez, F. et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
50. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).

51. Sosinsky, A. et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 genomes cancer programme. *Nat. Med.* **30**, 279–289 (2024).
52. Franco, I. et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biol.* **20**, 285 (2019).
53. Glasmacher, K. A. et al. Mutation of NOTCH1 is selected within normal esophageal tissues, yet leads to selective epistasis suppressive of further evolution into cancer. Preprint at https://doi.org/10.1101/2023.11.03.565535 (2023).
54. Wang, Y. K. et al. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nat. Genet.* **49**, 856–865 (2017).
55. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
56. Computational Science - ICCS 2020: 20th international conference, Amsterdam, The Netherlands, June 3-5, 2020: Proceedings, 3. Part III. (Springer, Cham, 2020).
57. Edwards, N. J. et al. The CPTAC Data Portal: A resource for cancer proteomics research. *J. Proteome Res.* **14**, 2707–2713 (2015).
58. Perdomo, S. et al. The mutographs biorepository: A unique genomic resource to study cancer around the world. *Cell Genomics* **4**, 100500 (2024).
59. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
60. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
61. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
62. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
63. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
64. Franco, I. et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat. Commun.* **9**, 800 (2018).
65. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
66. Tung, K.-F., Pan, C.-Y., Chen, C.-H. & Lin, W. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci. Rep.* **10**, 16245 (2020).
67. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
68. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky677 (2018).
69. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547.e23 (2017).
70. Elliott, K. et al. Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLOS Genet.* **14**, e1007849 (2018).
71. Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2**, https://doi.org/10.1214/08-AOAS191 (2008).
72. Besedina, E. & Supek, F. Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nat. Commun.* **15**, 6139 (2024).
73. Cameron, C. A. & Windmeijer, F. A. G. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econom.* **77**, 329–342 (1997).
74. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
75. Khalil, A. & Supek, F. DiffInvex identifies changes in driver gene repertoires during transformation to cancer and upon chemotherapy. Zenodo https://doi.org/10.5281/ZENODO.15161052 (2025).

Acknowledgements

Work in the lab of F.S. was supported by a European Research Council ERC CoG 'STRUCTOMATIC' (101088342), European Commission Horizon2020 project 'DECIDER' (965193) and Horizon Europe project 'LUCIA' (101096473), the Spanish government project 'REPAIRSCAPE' (PID2020-118795GB-I00), CaixaResearch project 'POTENT-IMMUNO' (HR22-00402), an ICREA professorship, EMBO YIP funding, the SGR funding of the Catalan government (SGR 00616), and a Novo Nordisk Fonden starting grant. We thank Elizaveta Besedina for assistance with developing the DiffInvex Poisson regression model, and Daniel Naro for retrieving and processing (alignment and variant calling) the WGS data from the CPTAC-3, MMRF-COMMPASS, OVCARE and Mutographs studies.

Author contributions

A.K. collected and curated the data, devised methods and implemented the DiffInvex computational framework, performed all analyses, visualized the data, and interpreted the results. F.S. conceptualized the analysis, devised the overall methodology and supervised the study. F.S. and A.K. drafted the manuscript jointly.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59397-8.

Correspondence and requests for materials should be addressed to Fran Supek.

Peer review information *Nature Communications* thanks Ruping Sun, and 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.