When neighbours play a role: a systems-level analysis of protein interactions conditioning cancer driver mutation effects

Márcia FS Vital[§1], João AI Miranda[§1,2], Margarida Carrolo[§3], António Quintela[3], Francisco R Pinto[1]

[1] BioISI, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal
[2] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, 1000-029, Lisbon, Portugal
[3]CUF Oncology – Hospital CUF Descobertas, Lisbon, Portugal
[§] These authors have contributed equally to this work
Corresponding author: Francisco R Pinto, frpinto@ciencias.ulisboa.pt

Abstract

Background: Cancer is driven by the accumulation of somatic mutations, including driver mutations that confer a selective advantage to cancer cells. Driver proteins operate within complex interaction networks, and their activity is conditioned by neighbour proteins. Understanding the interplay between driver mutations and the expression of their neighbour proteins can provide insights into cancer biology and potential therapeutic targets.

Methods: We assessed associations between expression of neighbour proteins and driver mutation status, comparing both between and within cancer types. We further evaluated if neighbours were enriched in significant associations with multiple drivers and characterised the impact of neighbour expression on overall survival for all cancer types.

Results: We found a significant correlation between the number of driver associations a neighbour gene has and the number of sign-coherent survival associations, particularly for neighbours enriched in positive associations, where high neighbour expression correlated with increased driver mutations and poorer survival. We identified 247 neighbours simultaneously enriched in positive driver and survival associations and 39 neighbours simultaneously enriched in negative driver and survival associations.

Conclusions: Our study systematically identified neighbours associated with driver mutation status. Complementary evidence from survival analysis and the literature suggests that neighbours enriched in driver associations are promising drug target candidates.

Introduction

Over the last decades, our understanding of oncogenesis has significantly expanded. The discovery of cancer driver genes in the 1970s has contributed to the advancement of targeted anticancer therapies and, more broadly, to the identification of genomic biomarkers for predicting prognosis and treatment response [1]. Likewise, the arrival of genomic sequencing of tumours has led to a remarkable collection of disease-driving mutations, offering valuable insights into cancer genomics [2].

Although 91% of tumours carry at least one gene driver mutation [3], these mutations alone are insufficient to fully explain the process of cell transformation into tumour cells [4]. Furthermore, research in healthy individuals has shown that some cancer driver mutations are already present in non-transformed cells across different somatic tissues, supporting the belief that mutations in cancer driver genes *per se* are not enough to promote tumorigenesis [5, 6].

The cellular context in which a driver mutation occurs, namely the physical interactors that are present, can condition the impact of the mutation on the cell phenotype [7]. Protein-protein interactions (PPIs) play a crucial role in both physiological and pathological processes, such as cell proliferation, signal transduction and apoptosis [8, 9]. Consequently, mapping of cancer-related PPIs is an ongoing effort to better understand disease and contribute to the development of precision medicine approaches [10–13].

In the last decades, PPIs have gained growing attention and emerged as therapeutic targets for both haematological and solid tumours [14, 15]. Examples of therapeutic approaches targeting cancer driver PPIs include inhibitors of MDM2/p53, KRAS/PDEδ and c-Myc/Max interactions [14].

Despite these advances, the systematic investigation of the relationship between driver mutations, neighbour gene expression, and patient survival remains largely unexplored. This study aims to address this gap by analysing the association between driver mutations and the expression of their neighbour genes in a pan-cancer cohort. We hypothesize that the expression of neighbour genes can modulate the phenotypic effects of driver mutations and influence patient survival. Moreover, we hypothesize that when such neighbour effects exist, we can identify them through correlations between neighbour expression and driver mutation status in cohorts of cancer patients. By integrating genomic, transcriptomic and interactomic data, we aim to identify novel therapeutic targets and gain a deeper understanding of the complex interplay between driver mutations and their protein interaction context in cancer.

Methods

Cancer drivers

A list of canonical and candidate cancer driver genes was retrieved from the Network of Cancer Genes & Healthy Drivers 7.1 [16].

Protein physical interaction data

Human PPI were retrieved from five sources: String (only PPI with a confidence score equal or greater than 0.4) [17], Biogrid [18], HuRI [19], Omnipath [20] and APID [21]. Interactions involving complexes were expanded to include all possible pairwise interactions. Protein physical interactors (neighbours) of each cancer driver were extracted from all five interactome sources.

Cancer patient data

Cancer patient data from the TCGA Pan-Cancer cohort was accessed through the UCSC Xena online tool (https://pancanatlas.xenahubs.net) [22]. In particular, gene-level non-silent somatic mutation data [23], batch effects normalised mRNA data, sample phenotype data and patient clinical data [24] were retrieved. Patients with more than 85 mutated cancer driver genes were discarded as they were considered mild outliers using the Tukey's fence method [25]. Missing values in the gene expression data were imputed with the minimum expression value observed for the respective gene across all patients.

Comparison of paired tumour and normal samples

665 patients in the pan-cancer cohort had paired tumour and normal tissue samples. For each of these patients, we computed the difference in gene expression between the tumour sample and the matching normal tissue. For each driver-neighbour PPI, we fitted a linear regression model where the dependent variable was the tumour-normal neighbour gene expression difference. The independent variables were 1) the mutation status of the driver and 2) the cancer type of the patients. Only data from cancer types that had at least one patient with a positive mutation status were included in the model. Additionally, the model was only applied when there were at least 2 cancer types included in the data used for model fitting. A significant ($p<0.05$) regression coefficient of the driver mutation status was considered as evidence that the driver mutation induced a change in the expression of the neighbour. These driver-neighbour PPIs were excluded from further analyses.

Statistical associations between neighbour expression and driver mutation status

We used data from 7770 patients (which passed the mutation outlier filter and had mutation and gene expression data for tumour samples) to compute, for each driver-neighbour PPI, 1) the average neighbour expression per cancer type and 2) the fraction of patients with mutated driver per cancer type. We then computed the Spearman correlation coefficient between these two variables and the corresponding statistical test to evaluate if the computed coefficient was significantly different from 0. The correlation was only computed when there were at least 3 cancer types with a driver mutation frequency greater than 0. Driver-neighbour PPIs with a significant correlation ($p<0.05$) were considered statistically associated in a between cancer type comparison (BCT).

Using the same data, we fitted a linear regression model for each driver-neighbour PPI, where the dependent variable was the neighbour gene expression and the independent variables were 1) the driver mutation status and 2) the patient cancer type. The model was fitted using only data from cancer types that had at least one patient with a positive mutation status. Additionally, the model was only applied when at least 3 cancer types were included. A significant ($p<0.05$) regression coefficient of the driver mutation status was considered as evidence that the driver-neighbour PPI was statistically associated in a comparison within cancer types (WCT).

Significant neighbour identification through random permutations

To evaluate the statistical significance of the number of driver-neighbour associations per neighbour protein, the computation of BCT and WCT associations was repeated 100 times with randomly permuted gene expression datasets. Gene expression values were organized in a matrix where each row corresponded to a gene and each column corresponded to a patient. Randomly permuted matrices were produced by randomly shuffling the column order without changing the row orders. In each shuffle, the number of significant driver associations (x_i) of each type (BCT or WCT) for each neighbour *i* was recorded, allowing the computation of the average ($< x_i^r >$) and standard deviation ($sd(x_i^r)$) of this number according to a random null model. These parameters were used to compute z-scores (equation 1) for the original observations and for each of the 100 analyses with randomly permuted data.

$z_i = \frac{x_i - < x_i^r >}{sd(x_i^r)}$ (equation 1)

For any given z-score threshold (t) we computed the False Discovery Rate (FDR, equation 2) using the average number of neighbours with a z-score greater than the threshold in the 100 analyses with the randomly permuted data ($< \#(z_i^r > t) >$) as an estimate of the expected false positives.

$FDR(t) = \frac{< \#(z_i^r > t) >}{\#(z_i > t)}$ (equation 2)

Equation 2 estimates FDR by excess, as the estimated number of false positives is done with the total number of neighbours instead of using the unknown proportion of neighbours that follows the null hypothesis (H_0: the neighbour does not have more driver associations than expected by chance). For each type of driver association (BCT or WCT), we determined the lowest value of t for which $FDR(t) < 0.05$.

Survival analysis

Using the overall survival times included in the TCGA Pan-Cancer cohort clinical data, we fitted a Cox proportional hazards regression model for each neighbour and each of the 32 cancer types included in the cohort. The models included the neighbour expression as an independent variable, codified as a binary variable (high/low expression for individuals above/below the median neighbour expression value within each cancer type). Regression coefficients significantly different from 0 ($p<0.05$) supported a statistical association between that neighbour expression and overall survival for that cancer type. A positive sign of the coefficient indicated that higher neighbour expression was associated with shorter overall survival times. To evaluate if a given neighbour has survival associations with more cancer types than expected by chance, we applied a binomial test (equation 3).

$$p = \sum_{s=x}^{32} P_{bin}(s, 32, p_{success})$$ (equation 3)

In equation 3, x is the number of survival associations observed for a given neighbour and $P_{bin}(s, n, p_{success})$ is the binomial distribution function that gives the probability of observing s successes in n Bernoulli trials, knowing that in each trial the probability of observing a success event is $p_{success}$. The binomial test was applied to each neighbour, for all survival associations (using $p_{success} = 0.05$) or only positive or negative survival associations (using $p_{success} = 0.025$). Binomial test p-values were adjusted to control FDR using the Benjamini-Yekutieli procedure [26].

Code availability

All the code used to perform the analyses presented in this manuscript are available at https://github.com/GamaPintoLab/driver_neighbours.

Results

Rationale and computational workflow

In this study, we sought to find physical interactors (hereafter called "neighbours") of cancer driver proteins that influence how driver mutations cause cancer. We reasoned that if a neighbour affects the phenotypic consequences of driver mutations in a cancer cell, it should be possible to identify a correlation between the neighbour gene expression and the presence of driver mutations in cancer patient data. Pharmacologically targeting these neighbours might lead to an attenuation or even reversion of the driver mutation effects. With this in mind, we designed a computational workflow (Figure 1) to analyse cancer patient genomic and transcriptomic data (from the TCGA Pan-Cancer cohort [3, 23] and evaluate Cancer Driver PPIs (known cancer driver genes extracted from NCG [16] and PPIs from Biogrid [18], String [17], HuRI [19], Omnipath [20] and APID [21]).

The expression of neighbours can vary between tissues and between individuals. Therefore, we followed two approaches to detect neighbours associated with driver mutation status: one comparing neighbour expression between cancer types (BCT) and the other comparing neighbour expression within cancer types (WCT) (Figure 1 A). But the detection of these associations does not imply a causal relationship. Three types of causal relationships can originate these associations: 1) the driver mutation induces changes in neighbour expression; 2) the neighbour expression influences the phenotypic outcome of the driver mutation; or 3) a third variable influences simultaneously the occurrence of mutations in the driver and the neighbour expression. We aim to find candidate neighbours with significant associations with cancer drivers due to the second type of causal relationship. The third type of causal relationship can only be excluded with experimental tests, but the first type can be excluded with available cancer patient data

(Figure 1 B). Using a cohort of patients in which gene expression was measured in both tumour and adjacent normal tissue, we could infer the effect of the driver mutation by quantifying the difference in the neighbour expression between tumour and normal tissues. If the difference is the same in patients with and without driver mutations, the mutation does not have a significant impact on the neighbour expression. Only these neighbour-driver PPIs were kept for further analysis.

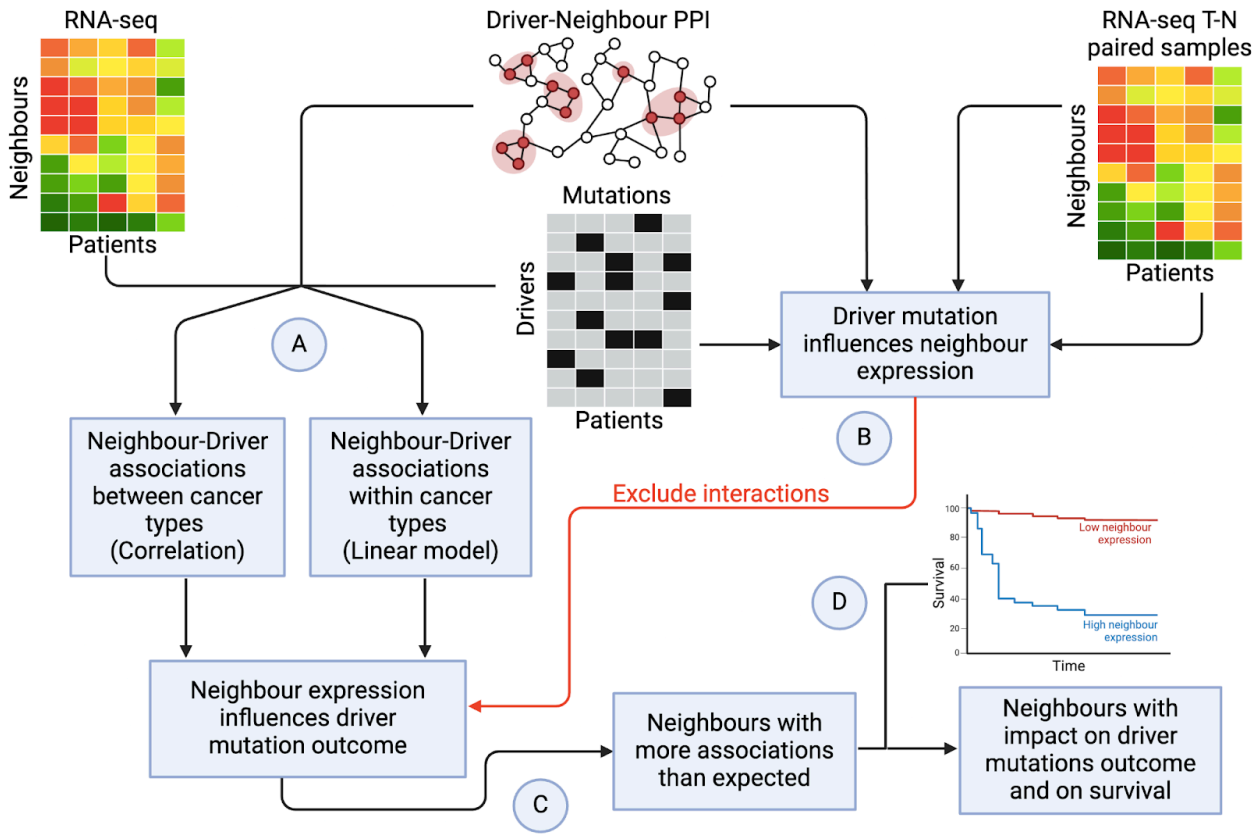

Figure 1—Computational analysis workflow. (A) Using cancer patient mutation data, PPI between cancer drivers and neighbour proteins and gene expression data from the same patients we evaluated two types of statistical association between neighbour expression and driver mutation status, one comparing between cancer types and other comparing within cancer types. (B) To avoid statistical associations where change in gene expression is caused by driver mutations, we also evaluated associations between the difference in neighbour expression between tumour and normal matching tissues and driver mutation status. Statistically significant associations found with paired tumour-normal samples were excluded from further analyses. (C) Next, we selected neighbours that had more driver associations than expected by chance, using analyses with randomly permuted expression datasets to estimate the random expectation. (D) Finally, we performed Cox proportional hazard regression models to evaluate the impact of neighbour expression on patient overall survival for all cancer types included in the cohort. This allowed us to select neighbours that were simultaneously enriched in driver and survival associations with coherent effect signs (positive or negative association with driver mutations concurrently with a decrease or an increase in overall survival times, respectively).

Due to the very large number of neighbour-driver PPIs under evaluation (362263) it is very difficult to apply a statistical cut-off that ensures a low fraction of false positives without drastically losing statistical

power (meaning that many false negatives are produced). Therefore, we focused on identifying neighbours that were involved in a larger number of significant associations with drivers than expected by chance (Figure 1 C). Lastly, we performed survival analysis to determine whether the identified neighbours might have an impact on disease progression and severity (Figure 1 D).

Neighbour-Driver associations are detectable between and within cancer types

We detected BCT associations by computing the Spearman rank correlation coefficient between the fraction of individuals with mutations in a driver and the mean expression of the driver neighbour for each cancer type. Figure 2A shows one example of a positive BCT association between neighbour expression and driver mutation.

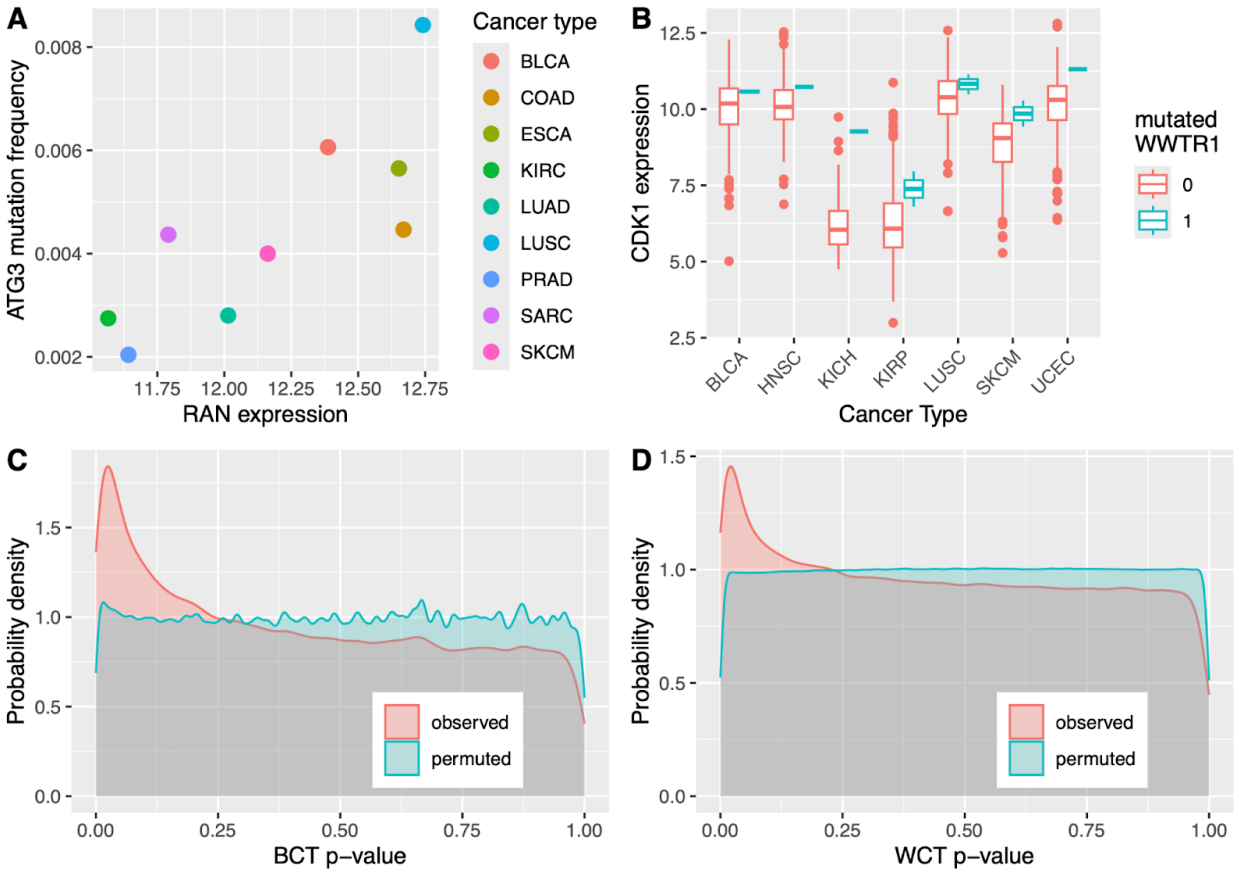

Figure 2—Driver-neighbour associations. (A) Example of a significant BCT statistical association between neighbour (*RAN* in the example) expression and the frequency of driver mutation (*ATG3* in the example). Each dot represents data from a different cancer type. (B) Example of a significant WCT statistical association between neighbour (*CDK1* in the example) expression and driver (*WWTR1* in the example) mutation status. Neighbour expression values are represented as box plots according to cancer type and driver mutation status. In (A) and (B) expression values are log2(normalised TPM + 1). (C) Probability density distributions of Spearman correlation test p-values obtained using observed data or randomly permuted expression datasets. (D) Probability density distributions of mutation status linear model coefficient p-values obtained using observed data or randomly permuted expression datasets.

This correlation and the associated statistical test were computed for all driver-neighbour PPIs. The obtained distribution of p-values showed an enrichment for lower p-values, contrasting with the uniform distribution obtained when the same tests were applied with permuted expression datasets (Figure 2C). This departure from the uniform distribution confirms the existence of driver-neighbour associations between neighbour expression and driver mutation status. To detect WCT associations, we applied linear models to predict neighbour expression using driver mutation status and cancer types as predictor variables. Figure 2B shows one example of a positive WCT association between neighbour expression and driver mutation. When model coefficients for driver mutation status are significantly different from zero, we can infer that neighbour expression is consistently distinct between individuals with and without driver mutation for multiple cancer types. The distribution of driver mutation status coefficient p-values for all driver-neighbour PPIs shows a similar bias to lower values (Figure 2D) implying that driver-neighbour associations are also detectable for neighbours with variable expression within cancer types. Considering significant associations with a $p<0.05$, we obtained 41471 BCT and 33668 WCT associations. However, some of these associations could be due to the influence of driver mutation on neighbour expression. To exclude these cases, we fitted a linear model to predict the difference in neighbour expression between tumour and normal paired tissue samples using as independent variables driver mutation status in the tumour tissue and cancer type. Excluding PPIs that had a model coefficient for driver mutation status significantly different from 0 ($p<0.05$), indicating that variations in neighbour gene expression can be caused by driver mutation, we retained 37331 neighbour-driver BCT associations (of which 63.6% are positive associations) and 29302 WCT associations (of which 52% are positive associations).

Neighbours enriched in BCT associations tend to have sign-coherent effects

The list of driver-neighbour PPIs evaluated included 15071 neighbours, but only around 56% of them were involved in significant BCT associations and 60% in WCT associations. Moreover, within the neighbours involved, the number of significant associations with drivers has a large variability. Part of that variability is explained by a linear relationship with the number of PPIs with driver proteins (Figure 3A-B). On average, each neighbour has significant driver associations with approximately 15% (for BCT) or 10% (for WCT) of his interacting drivers. We then evaluated if some neighbours were involved in more driver associations than expected by chance. For that, we repeated all analyses with 100 datasets where expression profiles were randomly permuted between all patients. In this way, the eventual associations between neighbour expression and driver mutation status were abolished, and the resulting distribution of the number of associations per neighbour was used to evaluate a null model random expectation. For each neighbour, the number of significant associations was transformed into a z-score. Similar z-scores were computed for the permuted datasets, allowing us to set a z-score threshold (3.1 for BCT and 2.65 for WCT) above which a false discovery rate lower than 0.05 was ensured. This resulted in the identification of 1390 and 1435 neighbours enriched in BCT and WCT driver associations, respectively. Interestingly, a high proportion of these neighbours enriched in driver associations had the majority of

their associations with a coherent sign, either positive or negative. This coherent-sign neighbours are dominant within the BCT-enriched neighbours, while for WCT-enriched neighbours, mixed signs are more common (Figure 3C).

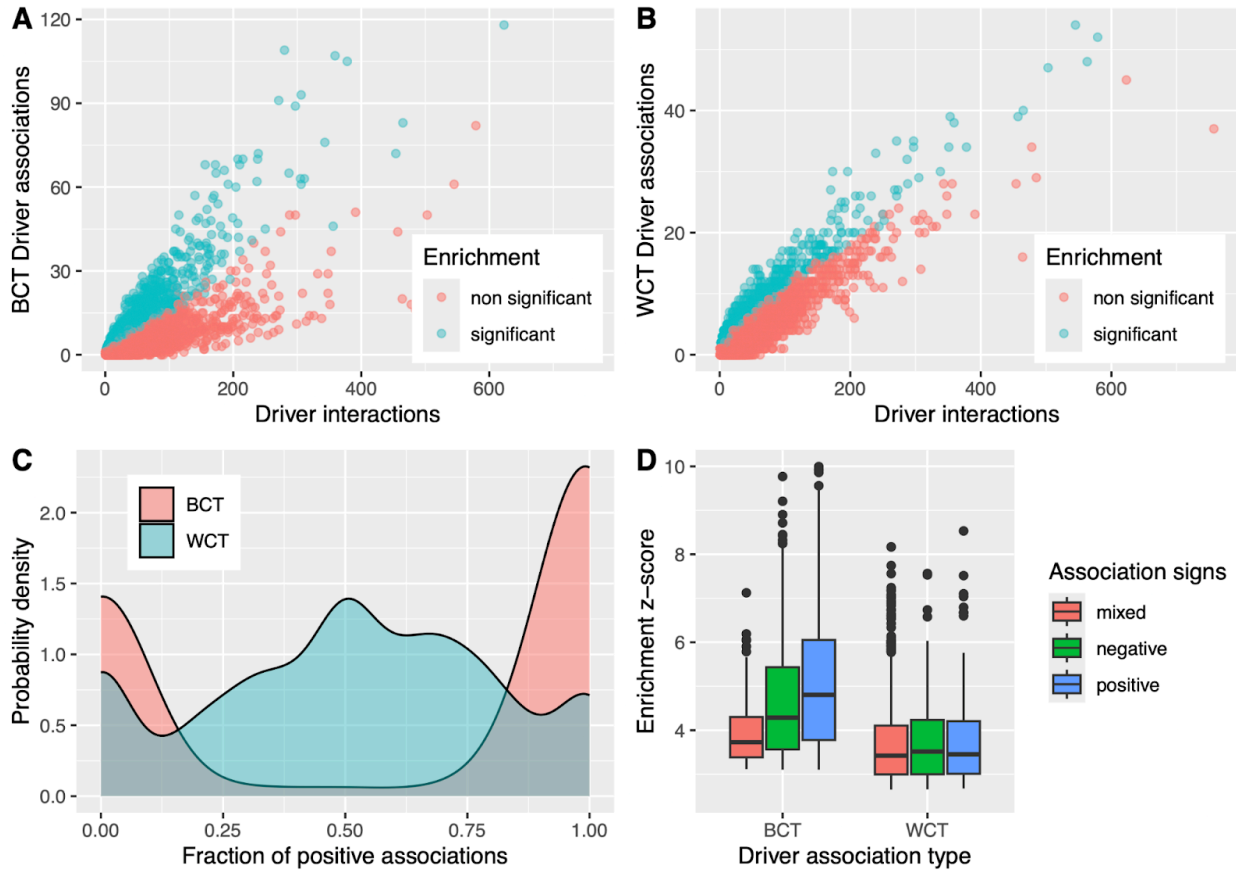

Figure 3—Neighbours enriched in driver associations. (A) Relation between number of PPIs with drivers and number of between cancer types (BCT) driver associations. (B) Relation between number of PPIs with drivers and number of within cancer types (WCT) driver associations. In (A) and (B), each dot represents a neighbour protein and colour indicates if the neighbour is enriched or not in driver associations. (C) Probability density distribution of the fraction of positive driver associations per neighbour (meaning that a higher neighbour expression is associated with a higher likelihood of driver mutation). Distributions for the two types of driver associations (BCT and WCT) are plotted separately. (D) Relation between driver association enrichment z-score and the typical sign of driver associations for each neighbour (mixed—fraction of positive associations between 10 and 90%; negative—fraction of positive associations below 10%; positive—fraction of positive associations above 90%).

We then classified enriched neighbours as “positive” (when at least 90% of their driver associations are positive), “negative” (when at least 90% of their driver associations are negative) or “mixed” (remaining cases). A higher expression of positive neighbours is associated with the tumorigenic effect of driver mutations. So we can infer that these neighbours have a cellular function that is intrinsically oncogenic. Conversely, negative neighbours should have tumour suppressor roles. Mixed neighbours have a context-dependent role according to which drivers are mutated. Within WCT-enriched neighbours, the

three sign classes had similar z-score distributions, meaning that the enrichment in driver associations was equally strong independently of the association signs. However, for BCT-enriched neighbours, the z-score distribution showed an increasing trend, with mixed neighbours showing the weaker association enrichments and positive neighbours showing the strongest association enrichments (Figure 3D). This suggests that sign-coherence is a characteristic property of BCT-enriched neighbours.

Neighbours enriched with driver associations tend to have cancer survival impacts that are coherent with the association signs

In order to verify whether associations between neighbour expression and driver mutation status had a significant impact on the clinical outcome, we assessed the impact of neighbour expression on overall survival across the 32 cancer types included in the TCGA Pan-Cancer cohort. For each cancer type cohort, we classified each patient into a high or low neighbour expression group, using the median neighbour expression in the cohort as a threshold. We then fitted a Cox proportional hazard regression model for each neighbour-cancer type pair. Significant model coefficients ($p<0.05$) were considered as evidence that neighbour expression was associated with changes in overall survival. The sign of the model coefficient indicated the direction of the neighbour effect on survival: positive coefficients were associated with a higher hazard ratio (shorter survival times), while negative coefficients were associated with a lower hazard ratio (and therefore, with better survival outcomes). First, we observed that neighbours with significant survival associations in more cancer types tended to have more driver associations, as shown by the distribution of the z-score for the number of driver associations according to the number of cancer types where a significant survival impact was detected (Fig. 4A). Next, we correlated the number of driver associations with the number of sign-coherent survival associations for each neighbour (Figure 4B). By sign-coherent, we mean that the survival impact has the expected sign according to the sign of the driver associations. Positive neighbour-driver associations tell us that high neighbour expression promotes the tumorigenic effects of driver mutations and, consequently, poorer survival outcomes (with positive coefficients in the survival model). Correlations between the numbers of driver and survival associations were measured separately for each type of driver association (BCT and WCT) and for neighbours that were enriched (or not) in driver associations. Indeed, all measured correlations were significantly positive, suggesting that neighbours with more driver associations tend to have sign-coherent survival effects in more cancer types. Moreover, correlations are stronger for neighbours that are significantly enriched in driver associations. Lower correlations among non-enriched neighbours may be due to a larger contribution of false-positive driver associations. Neighbours with WCT driver associations show weaker correlations. Also, for negative associations, the increase in correlation for neighbours enriched in WCT associations is not significant. Both observations can be explained by the fact that neighbours enriched in WCT associations are less sign-coherent (Figure 3C). In an effort to identify the neighbours with a robust potential to be further studied as drug targets, we applied binomial tests to find neighbours that were enriched with positive or negative survival associations. We were able

to confirm that neighbours enriched in positive driver associations were more than three times more likely to be also enriched in positive survival associations (Figure 4C, OR=5.19 CI95%=[4.50-5.99] for BCT and OR=3.46 CI95%=[2.83-4.20] for WCT). Enrichment in negative associations showed the same trend but with weaker strength (Figure 4D, OR=1.42 CI95%=[1.11-1.81] for BCT and OR=1.41 CI95%=[1.00-1.95] for WCT). Finally, we ended up with a list of 247 neighbours simultaneously enriched in positive driver associations and in positive survival associations (204 from BCT, 17 from WCT and 26 from both, supplementary table 1), and 39 neighbours simultaneously enriched in negative driver associations and in negative survival associations (29 from BCT, 6 from WCT and 4 from both, supplementary table 2).

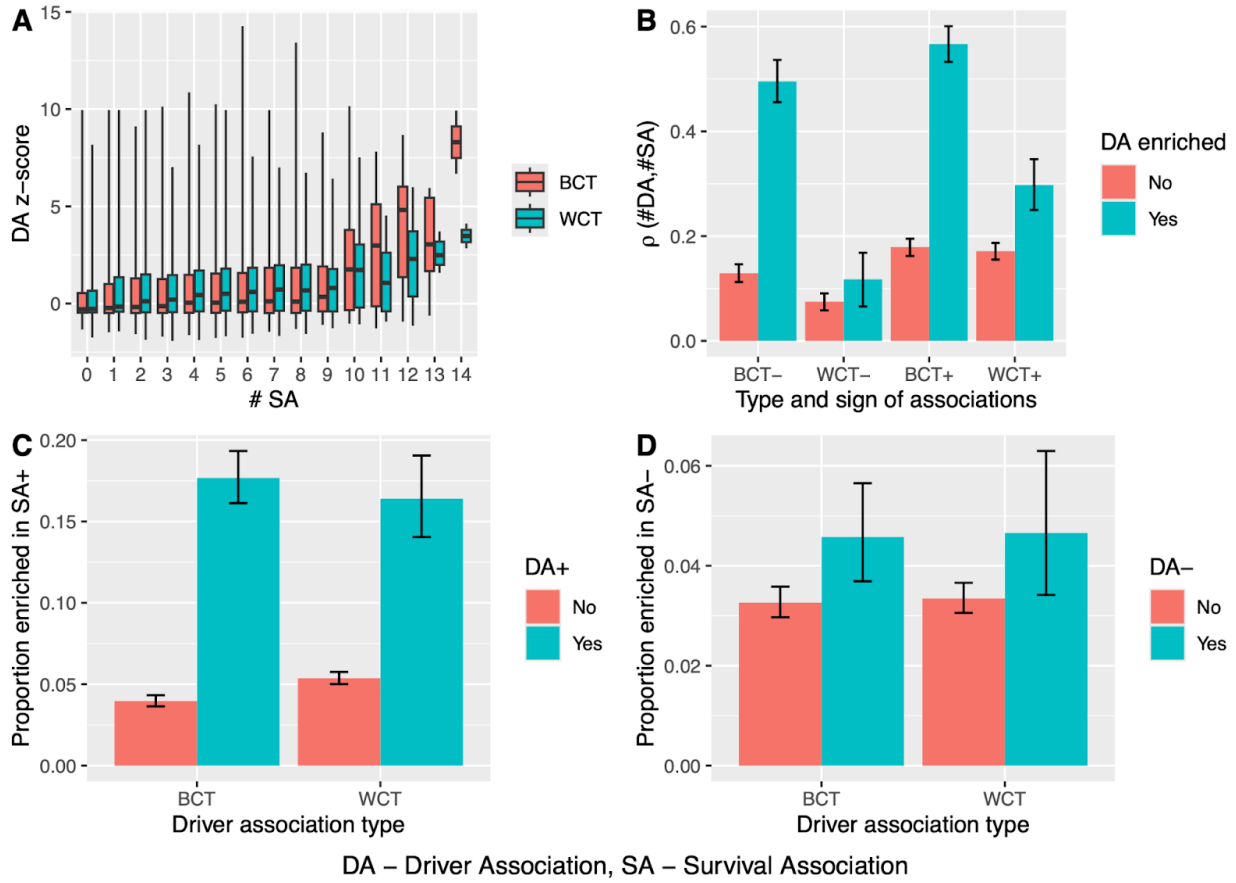

Figure 4—Association between driver and survival associations. (A) Boxplots representing the distribution of neighbour z-scores for the number of significant driver associations as a function of the number of cancer types for which a significant survival association was detected. (B) Spearman's correlation coefficient between the numbers of driver associations and survival associations of the same sign for neighbours with and without enrichment in driver associations. (C) Proportion of neighbours enriched in positive survival associations (SA+) according to their enrichment (or no enrichment) in positive driver associations (DA+). (D) Proportion of neighbours enriched in negative survival associations (SA-) according to their enrichment (or no enrichment) in negative driver associations (DA-). In (B), error bars represent 95% confidence intervals of Spearman's correlation coefficient estimated by bootstrapping, while in (C) and (D), error bars represent 95% confidence intervals for proportions estimated with the normal approximation.

Most neighbours simultaneously enriched in driver and survival associations have evidence of causing effects in cancer cells that are coherent with driver association sign

The candidate neighbour sets were selected through significant statistical associations with driver mutation status and overall survival. Both associations were evaluated with observational data from a large cancer patient cohort. Therefore, these associations alone are not enough to prove that these neighbours have any causal effect on cancer cells. A fraction of these neighbours are themselves cancer drivers (65 out of 247 positively associated neighbours and 14 out of 39 negatively associated neighbours, according to NCG [16]). Being cancer drivers implies that changes in their activities induce causal effects on cancer cells. Surprisingly, nine tumour suppressor genes (*BRCA1*, *BLM*, *BUB1B*, *CBFB*, *FANCA*, *FANCD2*, *FEN1*, *POLD1* and *POLE*) are among the neighbours enriched in positive associations with drivers and decreased survival in multiple cancer types. The pro-tumorigenic effect of high expression of some of these genes has been previously reported [27–31]. Interestingly, eight of these genes are involved in DNA repair pathways. Loss of function mutations in DNA repair genes may lead to accumulation of mutations and chromosomal instability, promoting cancer development and explaining their tumour suppressor roles. But when they are not mutated, higher expression may be associated with more aggressive tumours, where these pathways may be activated to prevent excessive chromosomal instability that could lead to cell death [32]. Higher DNA repair capabilities may also increase the resistance to some therapies, which may cause decreased survival times [33].

For neighbours not previously identified as cancer drivers, we conducted a bibliographic search to find experimental evidence (functional tests using overexpression, knockdown, or knockout in cancer cell lines) of causal effects that were coherent with the dominant sign of driver and survival associations. We successfully identified such experimental evidence for 87% (158 out of 182, supplementary table 1) of positively associated neighbours and for 52% (13 out of 25, supplementary table 2) of negatively associated neighbours.

We note that, for some neighbours, we were able to identify evidence with opposing effects. As an example, *CIRBP*, a neighbour enriched in negative driver associations, has been shown to both inhibit proliferation of pancreatic cancer cells [34] and promote breast cancer cell malignancy [35]. These dual roles could be justified by a mixture of positive and negative driver associations. This is not the case for *CIRBP*, for which only negative driver associations were detected. This suggests that *CIRBP* may have other PPIs with cancer-related proteins that explain its pro-tumorigenic role in breast cancer.

Finally, we searched the Open Target platform [36] and found that 38 of our selected neighbours (34 neighbours enriched in positive driver and survival associations and 4 enriched in negative driver and survival associations, supplementary tables 1 and 2) are known drug targets. 25 out of these 38 were not previously recognized cancer drivers. 31 were targeted by drugs developed for cancer treatment, confirming in this way the therapeutic potential of neighbours enriched in driver associations. Interestingly,

7 neighbours (*COL6A3*, *E2F1*, *E2F2*, *GMPS*, *LAMB3*, *PH4B* and *VAMP2*) were targeted by drugs for non-cancer diseases, suggesting an opportunity for drug repurposing.

Discussion

This study investigated the relationship between driver mutations, neighbour gene expression, and patient survival in cancer. We were able to identify significant associations between neighbour expression and driver mutation status that were not explained by a direct mutation effect. These associations were detectable both in comparisons between and within cancer types. Moreover, we could identify neighbours that had more driver associations than expected by chance. Many of these enriched neighbours tended to have sign-coherent associations with driver mutation status, either mostly positive or mostly negative associations. One key finding was the significant correlation between the number of driver associations and the number of sign-coherent survival associations for each neighbour. This correlation is particularly strong for neighbours enriched in driver associations. Furthermore, the correlations are generally stronger for BCT-enriched neighbours compared to WCT-enriched neighbours. Neighbours enriched in driver associations were significantly more likely to be enriched in survival associations, particularly for neighbours with positive associations, where high neighbour expression correlated with both increased driver mutations and poorer survival.

The fact that direct neighbours of cancer drivers are associated with patient overall survival is not surprising, as it has been observed that physical interactors of cancer drivers have an increased probability of being cancer drivers themselves [13]. There are also previously known examples of neighbours that condition the effects of driver mutations, such as EGFR and c-CBL [37] or TP53 and MDM2 [38]. In both examples, the neighbour proteins are E3 ubiquitin ligases that control the degradation rate of the driver proteins, therefore regulating their abundance and activity. The novelty in this work resides in the finding that such neighbour effects can leave a statistical signature in genomic and transcriptomic datasets of cancer patient cohorts. This signature allows for the systematic discovery of neighbours with significant driver associations that can be relevant prognostic biomarkers or therapeutic targets.

Not all neighbours that condition driver mutations effects are discoverable in this way. For the statistical driver association to be detectable, it is necessary that the neighbour has a natural expression variability, either between tissues or between individuals. On the other hand, similar statistical associations with driver mutation status and with patient overall survival can also exist for other proteins that are not neighbours of cancer drivers. The advantage of detecting an association for a neighbour protein is that the PPI supports the existence of a molecular mechanism explaining the statistical association.

Our study identified 247 neighbours simultaneously enriched in positive driver and survival associations and 39 neighbours simultaneously enriched in negative driver and survival associations. These

neighbours represent promising candidates for further investigation as potential drug targets. In fact, 31 of them are already known drug targets for cancer therapy. Additionally, a literature search revealed experimental evidence supporting the causal roles of a substantial proportion of these neighbours.

We opted to focus on neighbours that have most of their driver associations with the same sign. Their role in cancer development is less context-dependent, facilitating the detection of significant survival associations. Neighbours with a mixture of both positive and negative driver associations are not less valid. In fact, they can be fundamental to fully understand the role of drivers that behave both as tumour suppressors or as oncogenes, depending on the cellular context. But to untangle these context-dependent effects, either more data or experimental approaches are needed.

While the statistical associations observed in this study provide compelling evidence for the role of proteins that interact directly with cancer driver proteins, further experimental validation is essential to establish causality. Future research should focus on characterising the functional roles of the identified neighbours and exploring their potential as therapeutic targets. This could involve in vitro and in vivo studies to assess the impact of modulating neighbour gene expression on cancer cell growth, survival, and response to therapy. The impact of neighbour modulation should be stronger when the cancer cells harbour mutations in drivers interacting with that neighbour. The confirmation of these neighbours as drug targets, specially effective for subsets of patients carrying particular driver mutations, would expand the toolkit of targeted therapies.

It is important to acknowledge certain limitations of this study. The analysis relied on observational data from the TCGA pan-cancer cohort. The main advantage of using observational data is that measurements are made in real patient tissues instead of animal or cell models. However, it is not possible to control all confounding variables in observational studies. This limitation implies that we cannot control for a situation in which some individuals were exposed to unknown factors that could affect simultaneously driver mutations and neighbour expression and, therefore, generate an artificial association between driver mutation and neighbour expression.

Using observational data also implies that the number of cases with a particular combination of attributes is not controlled by the researchers. This limits the application of more complex statistical models. In the linear models used to detect associations between driver mutation status and neighbour expression, it would be interesting to include interaction terms between mutation status and cancer type. This would allow us to detect more driver-neighbour associations that would be conditional on the cancer type. Similarly, it would be informative to add interaction terms between neighbour expression and driver mutation status in the Cox proportional hazard regression models. If these interactions were significant, it would suggest that the impact of neighbour expression on survival was dependent on the existence of mutations in one of its interacting drivers. But given that for most drivers there is only a reduced number

of mutated individuals among the less prevalent cancer types, the statistical power is not sufficient for a precise estimation of the referred interaction coefficients.

RNA-seq data was used to estimate the abundance of neighbour proteins in cancer cells. Correlation between mRNA and protein levels can be rather low, especially when comparing values of different genes in the same tissue or cell type [39]. However, when comparing values for the same gene across different samples, mRNA-protein correlations are much higher [40]. In the present study, mRNA expression values being compared are always for the same neighbour gene across different individuals and cancer types. Therefore, we expect that most of these changes in gene expression reflect similar changes in protein levels.

Other limitations arise due to tumour heterogeneity and the diverse composition of the tumour microenvironment. RNA-seq was performed on the bulk of cells collected from the tumour tissue. We cannot be sure that all tumour cells carry the same mutations and have similar gene expression profiles. And part of the expression signal may come from non-tumour cells that are infiltrated in the tumour tissue. This issue could be solved by utilizing single-cell genomic sequencing and mRNA sequencing for each tumour sample. In this way, we could identify neighbour gene expression in the same cells that carried a mutation in the interacting driver. Besides the additional layer of complexity in the data analysis, we are not aware of the availability of this type of data for large cohorts of cancer patients.

The PPI datasets used might have their own limitations. It is possible that interaction data is incomplete, meaning that drivers may have unknown interactions with other proteins. Additionally, some interactions are only established with mutated drivers [41] and are not included in the interaction databases used here. Although our focus was on protein physical interactions, relevant interactions may be established with other types of molecules, such as non-coding RNAs. Finally, we only considered binary interactions. Some context-dependent effects of driver mutations may be explained only by higher-order interactions, involving more than one interactor simultaneously.

Despite these limitations, the findings of this study provide valuable insights into the complex interplay between drivers and their neighbours. It also demonstrates a new approach to analyse genomic and transcriptomic data from cancer patient cohorts that uncovers neighbours enriched in driver associations as potential therapeutic targets and prognostic biomarkers.

Funding: BioISI Centre grants UIDB/04046/2020 (DOI: 10.54499/UIDB/04046/2020) and UIDP/04046/2020 (DOI: 10.54499/UIDP/04046/2020) through Fundação para a Ciência e Tecnologia and an initiation to investigation fellowship (BII) under the BioISI Junior Programs (Márcia F. S. Vital).

References

1. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260:170–3.
2. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;:1–18.
3. Consortium TIP-CA of WG, Aaltonen LA, Abascal F, Abeshouse A, Aburatani H, Adams DJ, et al. Pan-cancer analysis of whole genomes. Nature. 2018;578:82–93.
4. Stransky B, Souza SJ de. Modeling tumor evolutionary dynamics. Front Physiol. 2013;3:480.
5. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574:532–7.
6. Martincorena I, Roshan A, Gerstung M, Ellis P, Loo PV, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Sci (N York, NY). 2015;348:880–6.
7. Schneider G, Schmidt-Supprian M, Rad R, Saur D. Tissue-specific tumorigenesis: context matters. Nat Rev Cancer. 2017;17:239–53.
8. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. Cell. 2011;144:986–98.
9. Cusick ME, Klitgord N, Vidal M, Hill DE. Interactome: gateway into systems biology. Hum Mol Genet. 2005;14 suppl_2:R171–81.
10. Kim M, Park J, Bouhaddou M, Kim K, Rojc A, Modak M, et al. A protein interaction landscape of breast cancer. Science. 2021;374:eabf3066.
11. Swaney DL, Ramms DJ, Wang Z, Park J, Goto Y, Soucheray M, et al. A protein network map of head and neck cancer reveals PIK3CA mutant drug sensitivity. Science. 2021;374:eabf2911.
12. Tabar MS, Francis H, Yeo D, Bailey CG, Rasko JEJ. Mapping oncogenic protein interactions for precision medicine. Int J Cancer. 2022;151:7–19.
13. Bouhaddou M, Eckhardt M, Naing ZZC, Kim M, Ideker T, Krogan NJ. Mapping the protein–protein and genetic interactions of cancer to guide precision medicine. Curr Opin Genet Dev. 2019;54:110–7.
14. Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. Signal Transduct Target Ther. 2020;5:213.
15. White AW, Westwell AD, Brahemi G. Protein–protein interactions as targets for small-molecule therapeutics in cancer. Expert Rev Mol Med. 2008;10:e8.
16. Dressler L, Bortolomeazzi M, Keddar MR, Misetic H, Sartini G, Acha-Sagredo A, et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. Genome Biol. 2022;23:35.
17. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51:D638–46.
18. Oughtred R, Rust J, Chang C, Breitkreutz B, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci : A Publ Protein Soc. 2021;30:187–200.
19. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature. 2020;580:402–8.
20. Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17:MSB20209923.
21. Alonso-López D, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, et al. APID database: redefining protein–protein interaction experimental evidences and binary interactomes. Database: J Biol Databases Curation. 2019;2019:baz005.
22. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting

cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38:675–8.
23. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018;6:271-281.e7.
24. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173:400-416.e11.
25. Tukey JW. Exploratory Data Analysis. 1977.
26. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics. 2001;29:1165–88.
27. Chang H-J, Yang U-C, Lai M-Y, Chen C-H, Fann Y-C. High BRCA1 gene expression increases the risk of early distant metastasis in ER+ breast cancers. Sci Rep. 2022;12:77.
28. Lan C, Yamashita Y, Hayashi H, Nakagawa S, Imai K, Mima K, et al. High Expression of Bloom Syndrome Helicase is a Key Factor for Poor Prognosis and Advanced Malignancy in Patients with Pancreatic Cancer: A Retrospective Study. Ann Surg Oncol. 2022;29:3551–64.
29. Fu X, Chen G, Cai Z, Wang C, Liu Z, Lin Z, et al. Overexpression of BUB1B contributes to progression of prostate cancer and predicts poor outcome in patients with prostate cancer. OncoTargets Ther. 2016;9:2211–20.
30. Hsu C-H, Ma H-P, Ong JR, Hsieh M-S, Yadav VK, Yeh C-T, et al. Cancer-Associated Exosomal CBFB Facilitates the Aggressive Phenotype, Evasion of Oxidative Stress, and Preferential Predisposition to Bone Prometastatic Factor of Breast Cancer Progression. Dis Markers. 2022;2022:8446629.
31. Bravo-Navas S, Yáñez L, Romón Í, Pipaón AC. Elevated FANCA expression determines a worse prognosis in chronic lymphocytic leukemia and interferes with p53 function. FASEB J. 2019;33:10477–89.
32. O'Connor MJ. Targeting the DNA Damage Response in Cancer. Mol Cell. 2015;60:547–60.
33. Kiwerska K, Szyfter K. DNA repair in cancer initiation, progression, and therapy—a double-edged sword. J Appl Genet. 2019;60:329–34.
34. Gao H, Xie R, Huang R, Wang C, Wang Y, Wang D, et al. CIRBP Regulates Pancreatic Cancer Cell Ferroptosis and Growth by Directly Binding to p53. J Immunol Res. 2022;2022:2527210.
35. Indacochea A, Guerrero S, Ure&ntildea M, Araujo F, Coll O, LLeonart ME, et al. Cold-inducible RNA binding protein promotes breast cancer cell malignancy by regulating Cystatin C levels. RNA. 2021;27:rna.076422.120.
36. Ochoa D, Hercules A, Carmona M, Suveges D, Baker J, Malangone C, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023;51:D1353–9.
37. Crotchett BLM, Ceresa BP. Knockout of c-Cbl slows EGFR endocytic trafficking and enhances EGFR signaling despite incompletely blocking receptor ubiquitylation. Pharmacol Res Perspect. 2021;9:e00756.
38. Vassilev LT. Small-molecule antagonists of p53-MDM2 binding: research tools and potential therapeutics. Cell cycle (Georget, Tex). 2004;3:419–21.
39. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Letters. 2009;583:3966–73.
40. Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol. 2016;12:883.
41. Mo X, Niu Q, Ivanov AA, Tsang YH, Tang C, Shu C, et al. Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. Cell. 2022;185:1974-1985.e12.