Integrative Computational Framework, *Dyscovr*, Links Mutated Driver Genes to Expression Dysregulation Across 19 Cancer Types

Sara Geraghty[1], Jacob A. Boyer[1,3], Mahya Fazel-Zarandi[1], Nibal Arzouni[1], Rolf-Peter Ryseck[1], Matthew J. McBride[1,2], Lance R. Parsons[1], Joshua D. Rabinowitz[1,3,4], Mona Singh[1,5,6,*]

SUMMARY: Though somatic mutations play a critical role in driving cancer initiation and progression, the systems-level functional impacts of these mutations—particularly, how they alter expression across the genome and give rise to cancer hallmarks—are not yet well-understood, even for well-studied cancer driver genes. To address this, we designed an integrative machine learning model, Dyscovr, that leverages mutation, gene expression, copy number alteration (CNA), methylation, and clinical data to uncover putative relationships between nonsynonymous mutations in key cancer driver genes and transcriptional changes across the genome. We applied Dyscovr pan-cancer and within 19 individual cancer types, finding both broadly relevant and cancer type-specific links between driver genes and putative targets, including a subset we further identify as exhibiting negative genetic relationships. Our work newly implicates–and validates in cell lines–*KBTBD2* and mutant *PIK3CA* as putative synthetic lethals in breast cancer, suggesting a novel combinatorial treatment approach.

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544
[2] Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ 08854
[3] Ludwig Cancer Institute, Princeton Branch, Princeton University, Princeton, NJ 08554
[4] Department of Chemistry, Princeton University, Princeton, NJ 08544
[5] Department of Computer Science, Princeton University, Princeton, NJ 08544
[6] Lead Contact
*Correspondence: mona@cs.princeton.edu

KEYWORDS: cancer systems biology, machine learning, gene regulation

HIGHLIGHTS:

- Integrative framework Dyscovr links mutations within cancer drivers to downstream expression changes
- Dyscovr uncovers known and novel targets of cancer-driver genes
- Dyscovr reveals clinically important negative genetic interaction pairings
- Web platform to explore uncovered driver gene-target relationships

eTOC BLURB: An integrative computational framework, Dyscovr, links mutated cancer driver genes to expression changes in putative target genes within and across 19 TCGA cancer types. Dyscovr's results include experimentally verifiable synthetic lethal driver-target pairings.

INTRODUCTION

The personalized medicine approach to cancer treatment has largely focused on targeting an individual's altered cancer driver genes. This approach has shown significant promise, particularly with the emergence of drugs targeting specific driver genes[1]. However, not all patients with a targetable alteration respond to the corresponding therapy. For many who do, their tumors eventually develop resistance, often by reactivating the driving pathway. Understanding the effects of driver gene mutations on the transcription of downstream genes throughout the genome would expedite the identification of genes that could be co-targeted along with the driver gene, ultimately improving the likelihood of durable success of these personalized treatments. However, given the intricate nature of each patient's cancer–characterized by unique molecular and environmental contexts–this remains a challenging task.

Much prior work in cancer regulatory genomics has centered on uncovering the set of regulators that determine the expression of a gene[2–8] or identifying cancer genes by determining whether they have a large impact on the gene expression of their known targets or interactions[9–

[15] or whether they are proximal in networks to dysregulated genes[16,17]. These approaches typically rely on *a priori* knowledge of global regulatory interactions, such as from protein-protein interaction (PPI) networks or transcription factor (TF)-target databases, to inform downstream linear or more complex models. Using networks as a prior can be useful in reducing multiple testing burden, but means that these models cannot uncover the impact of frequent gene-level somatic alterations on the expression of *all* dysregulated genes in the tumor genome. Another class of methods aims to determine how somatic alterations in cancer change the activity of regulators such as TFs[18–22]; in some cases, these methods also use inferred TF activity to make predictions about subsequent gene expression changes[22]. These approaches, however, do not attempt to link somatic alterations directly to quantitative expression changes in *individual* target genes, instead tending to focus on changes in regulatory activity and expression more broadly. This inability to predict precise changes in the expression of single genes across the cancer genome–particularly for indirect targets involved in druggable downstream cancer pathways–limits our understanding of the global transcriptional impact of somatic mutations and our ability to discover clinically actionable synthetic lethal pairings (i.e., gene pairs whose co-targeting leads to cell death).

Here, we introduce an interpretable computational framework that uncovers links between nonsynonymous somatic mutations in cancer driver genes and changes in expression in individual genes across the genome. This framework, Dyscovr (named for its ability to “discover” mutational links to transcriptional DYSregulation in cancer), is conceptually similar to approaches for identifying expression quantitative trait loci[23], but is focused on gene-gene associations, aggregates somatic mutations at different sites within a gene, considers the effects of multiple mutated genes in the same model, and integrates numerous other factors that are critical for the cancer context.

To hone in on gene-gene links with the most clinical potential, we developed a subsequent approach to identify which of Dyscovr’s hits are the most likely to exhibit synthetic lethality with

driver genes. This model takes advantage of cancer cell line data from the Cancer Dependency Map (DepMap)[24] to discover cases where mutant tumor suppressors or inhibition of mutant oncogenes synergistically interact with knockout of a putative target gene to inhibit cell growth. Among our uncovered links, we investigate *PIK3CA* mutations and their predicted downregulatory effect on *KBTBD2,* which we put forth as a cancer-relevant positive regulator of the PI3K-AKT pathway and insulin signaling. Further, we demonstrate experimentally that targeting *KBTBD2* enhances the efficacy of PI3K inhibitors in breast cancer cell lines.

Overall, we find that Dyscovr is a powerful tool for relating cancer driver mutations to target gene dysregulation across the cancer genome. Dyscovr is available open source (github.com/Singh-Lab/Dyscovr), and all of Dyscovr's predictions are easily browsable and downloadable via a user-friendly web interface (dyscovr.princeton.edu).

RESULTS

Overview of Dyscovr Framework

Briefly, Dyscovr integrates matched mutation, CNA, methylation, and expression data from primary tumor samples from The Cancer Genome Atlas (TCGA) (Fig. 1A) to disentangle the effects of each of these molecular phenotypes on transcription (Methods I). We first identify the set *S* of cancer driver genes[25] that frequently (≥5%) possess nonsynonymous mutations in the given patient cohort (Fig. 1B). For each candidate gene in the human genome, we apply the Dyscovr linear regression framework to simultaneously estimate the effect of the nonsynonymous mutation status of each driver gene *d* in *S* on that gene's expression. In each regression, we also include other factors that may influence gene expression, including the CNA and methylation status of each driver; the target gene's mutation, CNA, and methylation status; and sample-level covariates (e.g., patient age, gender, treatment status, tumor subtype, estimated fraction of infiltrating immune cells, genotypic variation, etc.) (Fig. 1C).

For each regression performed, Dyscovr extracts the fit coefficients for the nonsynonymous mutation status of each driver gene *d*, performs per-driver multiple hypothesis testing correction, and outputs a ranked set of driver mutation-target expression correlations with associated *q*-values, estimated magnitudes, and directionalities. We apply this framework both pan-cancer and within each of the 19 TCGA cancer types with at least 75 samples possessing all required data types, and use $q < 0.2$ as our significance threshold. Because *q*-values tend to be smaller with an increased number of tumor samples considered, we report results for the pan-cancer analysis using a $q < 0.01$ threshold.

Dyscovr Prioritizes Cancer-Related Genes and Known Functional Interactors.

Across primary samples spanning 32 cancer types in the TCGA, four annotated cancer driver genes[25] are mutated at greater than 5% frequency overall and in at least two cancer types: *TP53*, *PIK3CA*, *KRAS*, and *IDH1* (Table S1). We first used Dyscovr to uncover pan-cancer correlations between nonsynonymous mutations in these four driver genes and the expression of 16,447 putative target genes with sufficient data and variability across patient samples (see Methods IV.A). For each of *TP53*, *PIK3CA*, *KRAS*, and *IDH1*, our models identified hundreds to thousands of downstream target genes, or 'hits', whose expression is significantly correlated to their nonsynonymous mutation status. *TP53*, the most highly mutated gene across TCGA samples, accounts for the largest number of hits at a $q < 0.01$ significance threshold. In contrast, when *TTN*–a gene mutated at greater than 5% frequency in 22 of these 32 cancers due to its long length–is included in the model, it has only three hits at $q < 0.01$ (Fig. 2A, Fig. S1A). Additionally, when randomizing data, no driver gene has significant hits ($q < 0.2$), suggesting that Dyscovr's hits reflect genuine biological signal (Fig. S1B).

Our results also suggest Dyscovr prioritizes each driver's known targets or functional partners: for *TP53, PIK3CA, KRAS,* and *IDH1*, each gene's hits were significantly enriched in either the driver's transcriptional targets from DoRothEA[26], if a TF (i.e. *TP53*), or in the

transcriptional targets of downstream TFs that have been shown the literature to be intermediaries in enacting their effects (i.e. *PIK3CA, KRAS, and IDH1*) (Fig. 2B, Table S2). *TP53* also has exceptionally well-classified downstream targets from a variety of other sources, and we find that the hits Dyscovr identifies for *TP53* are significantly enriched for curated *TP53* targets[27] (Fig. 2C) as well as for targets identified in the TF-target databases TRRUST[28] and hTFtarget[29] (Table S3). *TP53*'s hits are also significantly enriched in genes from KEGG's P53 Signaling Pathway (hsa04115)[30], Reactome's Transcriptional Regulation by *TP53* Pathway (R-HSA-3700989) (Fig. 2C) and *TP53* Regulates Metabolic Genes Pathway (R-CFA-5628897)[31] (Table S3). Taken together, these strong enrichments across a variety of sources suggest that Dyscovr is effectively capturing transcriptional changes in downstream genes, including both direct transcriptional targets and pathway targets that may lie further downstream of the mutational event.

In the case of *TP53*, *IDH1*, and *PIK3CA*, each driver genes' hits are also statistically enriched in known cancer-related genes, such as those from the Cancer Gene Census (CGC)[32] (Fig. 2D, Table 1), highlighting Dyscovr's ability to prioritize genes with cancer-relevant roles. This is further supported by gene set enrichment analysis: we find that our driver genes' targets are also statistically enriched in cancer-related pathways from Gene Ontology (GO)[33] and KEGG[30] (Fig. S1C; Table S4). For *TP53*, for example, KEGG pathways include cell cycle (q = 7.84E-09), transcriptional misregulation in cancer (q = 1.82E-03), and p53 signaling pathway (q = 2.19E-03), as well as various metabolic pathways (Table S4).

Though an advantage of Dyscovr is its ability to estimate transcriptional changes in putative targets individually, we find that many of the targets that Dyscovr prioritizes also interact with one another. When each driver gene's top hits are overlaid on the STRING functional protein-protein interaction network[34], subclusters of dysregulated processes emerge. In the case of *TP53*'s top hits, a cluster of interconnected, cell cycle-related genes linked directly to *TP53* or to one another are visible (Fig. 2E). Similar clusters can be observed for the other drivers, such as upregulated MAPK signaling in the case of *KRAS* and downregulated CaMK kinase cascade

signaling in the case of *IDH1* (Fig. S2) suggesting that Dyscovr's results can be used to visualize how driver mutations disrupt broad functional networks.

When top hits for each of these drivers are examined individually, known genetic partners are apparent (Fig. 2F). These include *TP53* and *MDM2*, which form an autoregulatory feedback loop[35], and *KRAS* mutation and upregulation of *DUSP4/6* and *ETV4/5*, which are members of the ERK/MAP cascade downstream of *KRAS*[36]. Another example is *PIK3CA* mutation and upregulation of *PIK3R3*, a regulatory subunit of the PI 3-kinase (PI3K) of which *PIK3CA* is also a member. Dyscovr also identifies novel genetic relationships with partners outside of the drivers' immediate pathways, such as *PIK3CA* mutation and upregulation of sphingolipid metabolism genes *SGPL1* and *SPTLC2*. This class of genes has sparked recent interest due to its relevance to cancer diagnosis and prognosis, as well as its potential to provide new antitumor targets[37]. These links, which are available for browsing on the Dyscovr website, present intriguing, previously unstudied therapeutic opportunities.

Dyscovr Uncovers Driver Mutation-Target Expression Correlations for 19 Individual TCGA Cancer Types.

To discover nonsynonymous driver mutation-target gene expression correlations in a cancer-specific context, we next applied Dyscovr individually to the 19 TCGA cancer types with at least 75 samples possessing all data types. Each TCGA cancer type has its own unique set of cancer driver genes[25] mutated in at least 5% of samples (Fig. S3A); as such, the landscape of mutated driver-dysregulated target pairs identified varies by cancer type (Fig. 3A). While the absolute number of significant hits at a fixed q-value threshold is in part dependent upon the number of available samples, all 19 cancer types possess at least 10 significant hits at a $q < 0.2$ threshold, a substantial number given the large multiple testing burden.

On the whole, Dyscovr finds that transcriptional effects of driver mutations display similarity across cancer types, as the pairwise Spearman's rank correlations of mutational

coefficients fit by Dyscovr across all putative targets are largely positive (Fig. S3B). These results align with the hypothesis that mutations in cancer driver genes tend to affect similar downstream genes and processes across tissues. We find that there is still a great deal of tissue specificity, however, as Dyscovr's results more strongly replicate when the same tissue type is being compared. In the case of breast cancer, Dyscovr's hits for TCGA-BRCA and external dataset METABRIC[38] display high Spearman's rank correlations for the two most frequently mutated drivers *TP53* and *PIK3CA* (Fig. 3B, Fig. S3C).

We were particularly interested in genetic targets that are commonly dysregulated by a mutated driver across multiple cancer types, as these mechanisms might be broadly targetable (Fig. 3C). We identify some shared target genes with known mechanisms-of-action relating to mutations in the given driver–e.g. activating *PIK3CA* mutations have been shown to result in upregulation of glypican (GPC) family members, such as *GPC4,* and consequent tumorigenesis in gliomas[39]–though many other correlations have not been previously described. The link we identify between *PIK3CA* mutations in breast cancer, cervical cancer, and low-grade glioma and overexpression of *FYCO1*, for example, is not well explored, despite the fact that *FYCO1* has been shown to have important roles in migration and invasion of tumor cells[40].

Cell Viability Analysis Using DepMap Data Highlights Dyscovr Hits Displaying Putative Negative Genetic Interactions.

To further narrow Dyscovr's set of hits to those with the most clinical potential, we honed in on targets that exhibit putative negative genetic interactions such as synthetic lethality (SL) with nonsynonymous mutations in the corresponding driver genes. SL refers to cases where inactivation of one gene renders the other essential[41]. In this vein, we sought cases where loss of activity of a cancer driver gene is correlated with greater dependence of the cell upon a given target gene (i.e. decreased cell viability when the target is inhibited), which would suggest potential synergy. Loss of driver activity can occur either when expression of that driver is low, or

when a nonsynonymous mutation disrupts tumor suppressor gene function. Conversely, high driver activity can occur when expression of that driver is high, or when a nonsynonymous mutation hyperactivates oncogene function. Because we expect nonsynonymous mutations to impact the activity of tumor suppressors and oncogenes in different ways, we treated them separately in our analysis, looking to identify cases that showed positive correlation between oncogene mutation status and cell viability upon target knockout and a negative correlation between tumor suppressor mutation status and cell viability upon target knockout (Fig. 4A).

We tested for these genetic interactions using CRISPRi knockdown, somatic mutation, and gene expression cell line data from the DepMap database[24]. We developed a regression framework to relate driver gene nonsynonymous mutation status and expression to cell viability of a given target upon CRISPRi knockdown, accounting for disease type (Fig. 4B, Methods VIII), and applied it to hits identified for each of our TCGA pan-cancer driver genes with sufficient mutational diversity in CCLE cell lines (*TP53*, *PIK3CA*, and *KRAS*). By identifying cases where driver activity–captured by both a driver's expression and its mutation status–is linked to the essentiality of a given target gene, we can systematically identify cases where inhibition of the target in combination with a mutant tumor suppressor or inhibition of a mutant oncogene is likely to have a disproportionately negative effect on cell growth.

For each driver, we applied this method to the set of target genes from Dyscovr that were found to be significant both pan-cancer and within at least one individual cancer type ($q < 0.2$). Using this pipeline, we identify a much smaller set of driver-target pairs which exhibit putative SL relationships (Fig. 4C, Table S5). As is the case for Dyscovr's full set of hits, identified putative SLs are enriched for cancer-related GO pathways such as DNA damage checkpoint signaling, regulation of cell death and apoptotic process, and cell cycle checkpoint signaling (Fig. 4D). Though gold standard experimental sets of SL genes are scarce, *KRAS* has a set of 23 genes from the SynLethDB 2.0 database[42] that are confident SL partners from experimental sources (confidence >0.7) and intersect the genes tested in our analysis. We see enriched overlap of

these genes with those identified as significant *KRAS* SL partners by our pipeline, albeit not at the level of statistical significance (hypergeometric p = 9.72E-02).

Closer examination of the putative SL candidates reveals both known pairings from the literature as well as novel pairings. These include genes involved in top enriched GO pathways (Fig. 4E), as well as candidates involved in an assortment of other cancer-related processes, such as DNA damage response, cellular metabolism, and angiogenesis (Table S5). The top hit for *KRAS*, for example, is fibroblast growth factor receptor (*FGFR)* adaptor protein *FRS2*; *FGFR1*, which signals via *FRS2*, has been found to mediate adaptive drug response in *KRAS*-mutant lung cancers, leading to success of a combinatorial treatment approach[43]. Notably, Dyscovr identified this link between mutant *KRAS* and downregulation of *FRS2* in lung cancer (q = 0.13), but not in colon, pancreatic, or uterine: this is in close alignment with previous evidence that this combinatorial treatment strategy is effective only in *KRAS*-mutant lung cancers and not other *KRAS*-mutant cancer types[44]. Other of *KRAS's* top predicted SLs have also been linked to *KRAS* mutations in a cancer context, such as *ARFGEF2*[45] and *NPAS2*[46]*,* while others such as *CDCP1* and *SKP2* have not been previously associated with *KRAS* mutation, but have been associated with tumorigenesis[47–49]*.*

In the case of *TP53*, we identify examples of cases where *TP53*-mutant tumors are more sensitive to knockdown of a given target gene. *USP28*, the top hit, is an oncogene that regulates a variety of tumorigenic processes in cancers like squamous cell carcinoma, including cellular proliferation, DNA damage repair, and apoptosis. Overexpression of *USP28* has been associated with poorer outcomes, leading to the development of therapeutics targeting this gene[50]. However, *USP28* has also been put forth as a candidate tumor suppressor, as its deubiquitinating functions play a role in stabilizing tumor suppressor *TP53 in vivo*[50]. Our analysis aligns with this dual functionality, suggesting that targeting *USP28* is most effective in contexts where *TP53* function has already been disrupted via nonsynonymous mutation and *USP28's* tumor suppressive functions are therefore no longer important. Several others of TP53's top SL hits have bodies of

evidence directly supporting a SL relationship, including oncogene *CHEK2*[51] (which has also been shown to modulate resistance to epirubicin in tandem with *TP53* in breast cancer[52]); *LBR* (which has been shown to promote cellular proliferation in the absence of *TP53*[53], and is also a member of the increasingly well-studied class of lamin B-related diseases[54]); *ACO1/IRP1* (a key modulator of iron homeostasis that is involved in a well-characterized iron-*TP53* feedback loop in cancer[55,56]); and *CUL9* (a tumor suppressor that promotes *TP53*-dependent apoptosis[57] and regulates cell proliferation, senescence, apoptosis and genome integrity via *TP53*[57]). Many of *TP53*'s other top hits have been reported in the cancer literature, though without a mechanistic relationship to *TP53* mutations, such as *NCDN*[58], *PSME4*[59,60], *FAM189B*[61], *CKAP2L*[62,63], *GHR*[64], and *MYO9B*[65].

Many of mutant *PIK3CA*'s top SL candidates have been shown to act via or downstream of the PI3K-AKT signaling pathway, such as *TM4SF1,* which regulates breast cancer cell migration and apoptosis[66], *SLC7A2*, which mediates recruitment of myeloid-derived suppressor cells and tumor immunosuppression[67], and *TCF7L2*, a WNT pathway effector shown to mediate colorectal cancer cell migration and invasion[68] (though WNT signaling dysregulation has been implicated in other *PIK3CA*-mutant cancer types as well[69]). Other identified hits are involved in insulin signaling and regulation, which activates PI3K-AKT signaling; these include well-described *PIK3CA*-interactors *IRS2* and *IGF1R*[70]*,* which are also candidate oncogenes in a variety of cancer types[71–75].

Dyscovr Pipeline Reveals Clinically Actionable SL Pairs. To demonstrate Dyscovr's ability to identify candidate SL pairs, we focused on an under-studied member of a cullin3-RING E3 ubiquitin ligase complex, kelch repeat and BTB domain-containing protein 2 (*KBTBD2)*. Kelch repeat and BTB domain-containing proteins are adaptors which provide substrate specificity to the E3 ligase complex[75,76]. Physiologically, *KBTBD2* has been shown to regulate insulin signaling in adipocytes by controlling stability of PI3K regulatory subunit, p85α[77,78]. The function of *KBTBD2*

in cancer remains unexplored, though it was identified as a significant negative genetic interactor with mutant *PIK3CA* (q = 6.2E-02) and was predicted by Dyscovr to be downregulated in relation to *PIK3CA* mutation pan-cancer (q = 6.5E-04) and in breast cancer (q = 4.8E-02), suggesting a cancer-relevant role. In addition, *KBTBD2* followed a similar pattern as several other putative positive regulators of PI3K signaling such as *IGF1R*, *IRS2* and *FURIN* that were also downregulated in *PIK3CA*-mutant tumors, likely the result of increased pathway output and hence negative feedback (Fig. 5A). Further implicating *KBTBD2* as a positive regulator of PI3K signaling in human cancer, co-dependency analysis showed that cells with a high dependence on *KBTBD2* were most likely to be sensitive to a variety of *IGF1R* inhibitors (Fig. 5B). Highlighting *KBTBD2* as a potential oncogene, we observed a significant reduction in survival rates across tumors in which *KBTBD2* was highly expressed, both pan-cancer (Fig. 5C) and in breast cancer (Fig. 5D).

To interrogate the function of *KBTBD2 in vitro*, we used siRNA to ablate its expression in the ER+, PI3K mutant cell line, MCF7. Knockdown of *KBTBD2* alone resulted in a 75-85% reduction in cell growth (Fig. 5E, Fig. S4A). Interestingly, knockdown of *KBTBD2* at baseline did not alter signaling through the PI3K pathway as analyzed by canonical substrates (Fig. S4B), though we did notice a slight increase in p85α expression–a regulatory subunit of PI3K involved in modulating insulin sensitivity–following ablation of *KBTBD2*. Strikingly, *KBTBD2* knockdown significantly enhanced the effects of a clinically employed PI3K inhibitor, alpelisib[79]—with 1μM alpelisib blocking growth by approximately 25%, and the addition of *KBTBD2* ablation converting this effect into cytotoxic cell death (Fig. 5F). This effect was even more pronounced with 10μM alpelisib. To examine the effects on pathway output, we knocked down *KBTBD2* in MCF7 cells followed by treatment with alpelisib for time *t*. We analyzed phosphorylation of PI3K effector *AKT*, as well as the downstream output of mTOR complex I (*mTORC1*), which is regulated by PI3K signaling and is a frequently activated and therapeutically targeted oncogene across many cancer types[80]. As expected, suppression of *mTORC1* substrate phosphorylation correlates with response to PI3K inhibitors[78]. When *KBTBD2* was knocked down, we observed a deeper inhibition

of *mTORC1* targets, *pS6* and *p4EBP1*. Cyclin D1, also under *mTORC1* control[81], was also better suppressed when *KBTBD2* expression was ablated (Fig. 5G). We also tested a mutant-selective PI3K alpha inhibitor, RLY-2608[82]. In this case, *KBTBD2* knockdown was more effective at blocking growth than the drug at any concentration. Additionally, the combination of RLY-2608 with *KBTBD2* knockdown produced only a minor additive effect on growth inhibition. Interestingly, in terms of pathway inhibition, knockdown of *KBTBD2* enhanced the inhibitory effects of RLY-2608, with *pS6*, *p4EBP1*, and cyclin D1 all more deeply suppressed than in the control siRNA cells (Fig. S4C). Taken together, these results demonstrate that *KBTBD2* inhibition holds exciting potential to enhance the effects of *PIK3CA* inhibitors in *PIK3CA*-mutant breast cancers, warranting further investigation. This example suggests that the Dyscovr platform can not only predict driver mutation-expression pairs, but also identify a subset of such predictions as clinically informative or even pharmacologically actionable.

DISCUSSION

We have introduced an integrative framework to link nonsynonymous mutations in cancer driver genes to transcriptional dysregulation in target genes across the genome. Our method that uses this framework, Dyscovr, draws on a wide array of data types, including CNA and methylation data and clinical features, to determine the unique contributions of each feature to transcriptional dysregulation in cancer. When applied to over 6,000 primary tumor samples from the TCGA, both in a pan-cancer context and within 19 individual TCGA cancer types, Dyscovr reveals thousands of novel correlations with potential clinical relevance. These correlations are highly interpretable, can be replicated across patient cohorts, and are enriched in cancer-relevant genes and pathways. When assessed using gold standard sets of known targets for widely mutated cancer driver gene *TP53*, Dyscovr successfully prioritizes these targets. All Dyscovr's correlations for the TCGA are downloadable and searchable by gene and cancer type through its

website (dyscovr.princeton.edu). Altogether, Dyscovr's integrative approach sheds new light on ways that a patient's driver mutational landscape influences downstream processes with a specificity that lends itself to experimental and clinical applications.

For decades, scientists have conducted expression quantitative trait loci (eQTL) analyses to relate cancer mutations to gene expression changes[83]. While valuable, traditional eQTL analyses are not directly applicable to the task at hand, as they do not traditionally account for target-level sources of expression variability that are common in cancer, such as CNAs, methylation changes and mutations, or other features of tumors such as immune cell infiltration. In our models, cancer subtype, target CNA status, target methylation status, and level of immune cell infiltration are the strongest determinants of a target gene's expression (Fig. S5A), suggesting that attempts to relate mutation status to target gene expression without accounting for these factors may lead to erroneous conclusions. This is particularly problematic given the increasingly important role that patients' driver mutations play in dictating their consequent treatment plan, as targeted sequencing panels become routine in the clinic[84]. An advantage of our framework is that it produces fit coefficients with meaningful magnitudes and directionalities, as well as significance measures. This enables meaningful ranking and thresholding of results that is often unachievable with modern black box machine learning methods (which have also been shown to underperform simple linear models on similar tasks[85]). This interpretability allows clinicians to decipher the role that driver mutations play in downstream tumorigenic processes, independent of other mutations, molecular alterations, and patient background.

In this work, we also show how commonly mutated drivers may be jointly targeted with previously understudied genes to increase the effectiveness of driver-targeted therapies, with the putative SLs identified by Dyscovr serving as potentially clinically useful synergistic genetic targets. This is demonstrated by Dyscovr's prioritization of known SL pairings (e.g. *KRAS* and *FRS2*[44]), enrichment among putative SL genes in cancer-related processes, and the experimental validation of a previously unstudied relationship between mutated *PIK3CA* and *KBTBD2*

expression. In *PIK3CA*-mutant breast cancer cell lines, joint inhibition of *PIK3CA* and *KBTBD2* resulted in more suppressed cell growth than *PIK3CA* inhibition or *KBTBD2* knockout alone (Fig. 5). We anticipate that many of Dyscovr's other, untested correlations may hold similar therapeutic value.

There are several possibilities for future expansions of the Dyscovr framework. For one, we restricted Dyscovr to cancer driver genes mutated at sufficiently high frequencies across available samples, which poses challenges for small cohorts or highly mutationally diverse cancer types. As multiomic sequencing continues to rapidly advance and more tumors are sequenced, we can apply Dyscovr to larger patient cohorts and uncover relationships for less frequently mutated driver genes. Ideally, these cohorts will also have greater representation of patients from across all ethnic backgrounds, as our current work with the TCGA is limited by the overrepresentation of patients from European backgrounds, a well-described problem in the field of cancer genomics[86]. We also look forward to the possibility that additional sources of transcriptional dysregulation in cancer (i.e. chromatin accessibility) may be measured at-scale and included in Dyscovr's framework. Similarly, as proteomic sequencing increases in availability and scope, we see exciting possibilities for harnessing Dyscovr to relate driver mutations directly to changes in protein levels—an application with clear relevance to cancer therapies that act on protein targets. Finally, the swift rise of single-cell sequencing and the concerted push to sequence multiomic data from single cells portends the compelling possibility of using a framework such as Dyscovr to study these regulatory mechanisms within tumor cellular subpopulations.

Ultimately, Dyscovr provides a powerful, integrative approach to study the molecular mechanisms of cancer cells. As cancer treatment becomes increasingly personalized and informed by genomic science, we anticipate that Dyscovr will prove a valuable tool to tease apart the ways that driver mutations reshape cellular processes.

ACKNOWLEDGEMENTS

Thanks to members of the Singh and Rabinowitz labs for their insights and comments. Thanks especially to Joshua Wetzel for helpful discussions regarding TCGA mutation and CNA data and for his review of the manuscript. The results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/, the METABRIC Consortium: https://www.nature.com/articles/nature10983, and the DepMap Consortium: https://depmap.org/portal/. This work was funded in part by NIH grant R01-CA208148 (to M.S.), Ludwig-Princeton AGMT DTD 1/1/2021 (to J.D.R. and M.S.), and NSF GRFP grant DGE-2039656 (to S.G.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, S.G. and M.S. (Dyscovr), J.B., M.M., and J.D.R. (Experimental Validation); Methodology, S.G. and M.S.; Software, S.G. (Dyscovr), N.A. and L.P. (Dyscovr Website); Validation, S.G. (Computational), J.B., M.F., and R.R. (Experimental); Resources, M.S. and J.D.R.; Writing – Original Draft, S.G.; Writing – Review & Editing, S.G., M.S., J.B., and J.D.R.; Visualization, S.G. and J.B.; Supervision, M.S. and J.D.R.; Project Administration, M.S. and J.D.R; Funding Acquisition, M.S. and J.D.R.

DECLARATIONS OF INTEREST

J.D.R. is a member of the Rutgers Cancer Institute of New Jersey (RCINJ) and the University of Pennsylvania Diabetes Research Center (U Penn DRC); a director of the U Penn DRC-Princeton inter-institutional metabolomics core and RCINJ metabolomics core; an advisor and stockholder in Colorado Research Partners, Bantam Pharmaceuticals, Barer Institute, Rafael Pharmaceuticals, Faeth Therapeutics, and Empress Therapeutics; a founder, director, and

stockholder of Farber Partners, Raze Therapeutics, and Sofro Pharmaceuticals; a founder, advisor, and stockholder in Marea Therapeutics and Fargo Biotechnologies; inventor of patents held by Princeton University; and a director of the Princeton University-PKU Shenzhen collaboration.

METHODS

Star Methods Key Resources Table

Reagent or Resource	Source	Identifier
Deposited Data		
Annotated simple nucleotide variation (SNV) data, gene-level somatic copy number variation (CNV) data, RNA-seq transcriptomic profiling data, 450K DNA methylation data, biospecimen and clinical data	The Cancer Genome Atlas (TCGA)	http://cancergenome.nih.gov/
Tumor purity estimates	Aran et al., 2015	https://doi.org/10.1038/ncomms9971
Genotypic principal components (PCs), computed using Washington University method	Carrot-Zhang et al., 2020	https://doi.org/10.1016/j.ccell.2020.04.012 Supplemental materials: https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020 File name: “WashU_PCA_ethnicity_assigned.tsv”
Mutation data, copy number alteration (CNA) data, mRNA expression microarray data, promoter methylation (RRBS) data, clinical data	Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	https://doi.org/10.1038/nature10983 https://www.cbioportal.org/study/summary?id=brca_metabric
Mutation data, RNA-seq gene expression data, drug sensitivity data (PRISM	The Broad Institute Cancer Dependency Map	https://depmap.org/portal/

repurposing primary screen), CRISPR gene dependency data (22Q4+Score, Chronos)	(DepMap)	
Protein-protein interaction networks (confidence > 0.4)	STRING Consortium, 2023	https://string-db.org/
Drug-target gene associations	DrugBank v.5.1.10	https://go.drugbank.com/releases/latest
UniProtKB	The UniProt Consortium, 2019	https://www.uniprot.org/uniprot/
Vogelstein cancer driver genes	Vogelstein et al., 2013, Tables S2A, S2B, S3A–S3C, and S4	https://doi.org/10.1126/science.1235122
Cancer Gene Census (CGC) cancer driver genes	Futreal et al., 2004	https://cancer.sanger.ac.uk/census
Curated *TP53* target genes	Fischer et al., 2017, Table 1	https://doi.org/10.1038/onc.2016.502
Reactome pathways	Jassal et al., 2020	https://reactome.org
KEGG: Kyoto Encyclopedia of Genes and Genomes	Kanehisa Laboratories	https://www.genome.jp/kegg/
TRRUST: Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining, v2	Han et al. 2018	https://www.grnpedia.org/trrust/ https://doi.org/10.1093/nar/gkx1013
hTFtarget, TF-target regulations	Zhang et al., 2020	http://bioinfo.life.hust.edu.cn/hTFtarget https://doi.org/10.1016/j.gpb.2019.09.006
DoRothEA	Garcia-Alonso et al., 2019	https://saezlab.github.io/dorothea/ https://doi.org/10.1093/bioadv/vbac016
IntAct Molecular Interaction Database v. 1.0.4	del Toro et al., 2022	https://www.ebi.ac.uk/intact/home https://doi.org/10.1093/nar/gkab1006
BioGRID v. 4.4.233 (The Biological General Repository for Interaction Datasets)	Oughtred et al., 2021	https://thebiogrid.org/ https://doi.org/10.1002/pro.3978

Software and Algorithms		
Dyscovr	This paper	https://github.com/Singh-Lab/Dyscovr
Gene ID Conversion		
biomaRt package for R (version 2.48.3)	Durinck et al., 2009	https://bioconductor.org/packages/biomaRt/
org.Hs.eg.db package for R (version 3.13.0)	Carlson, 2019	https://bioconductor.org/packages/org.Hs.eg.db/
Gene Set Enrichment Analysis		
DOSE package for R (version 3.18.3)	Yu et al., 2015	https://bioconductor.org/packages/DOSE/
ReactomePA package for R (version 1.36.0)	Yu and He, 2016	https://github.com/YuLab-SMU/ReactomePA
clusterProfiler package for R (version 4.0.5)	Wu et al., 2021	https://bioconductor.org/packages/clusterProfiler/
enrichplot package for R (version 1.12.3)	Yu et al., 2021	https://yulab-smu.top/biomedical-knowledge-mining-book/
GoSemSim package for R	Yu et al., 2020	https://bioconductor.org/packages/release/bioc/html/GOSemSim.html
Data Accession (Specific Datasets)		
TCGAbiolinks package for R (version 2.20.1)	Colaprico et al., 2016	https://github.com/BioinformaticsFMRP/TCGAbiolinks
maftools package for R (version 2.8.5)	Mayakonda et al., 2018	https://github.com/PoisonAlien/maftools
TRONCO package for R	Caravagna et al., 2023	https://bioconductor.org/packages/TRONCO
GenomicRanges package for R	Lawrence et al., 2013	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
immunedeconv package for R (version 1.30.16)	Sturm et al., 2019	https://icbi-lab.github.io/immunedeconv/
STRINGdb package for R	Szklarczyk et al., 2021	https://www.bioconductor.org/packages/release/bioc/html/STRINGdb.html

dorothea package for R	Garcia-Alonso et al., 2019	https://bioconductor.org/packages/release/data/experiment/html/dorothea.html
Data Manipulation		
data.table package for R (version 1.14.8)	Dowle et al. 2023	https://CRAN.R-project.org/package=data.table
broom package for R (version 1.0.5)	Dobinson et al., 2023	https://CRAN.R-project.org/package=broom
rlang package for R (version 1.1.1)	Henry et al., 2023	https://CRAN.R-project.org/package=rlang
tidyverse package for R	Wickham et al., 2023	https://doi.org/10.21105/joss.01686
reshape2 package for R	Wickham et al., 2007	http://www.jstatsoft.org/v21/i12/
rlist package for R (version 0.4.6.2)	Rem et al., 2021	https://CRAN.R-project.org/package=rlist
abind package for R (version 1.4-5)	Plate et al., 2016	https://CRAN.R-project.org/package=abind
fastmatch package for R (version 1.1-3)	Urbanek, 2021	https://CRAN.R-project.org/package=fastmatch
pandas package for Python	The Pandas Development Team, 2020	https://doi.org/10.5281/zenodo.3509134
Statistical Analyses		
edgeR package for R (version 3.34.1)	Robinson et al., 2010	https://bioconductor.org/packages/edgeR/
speedglm package for R (version 0.3-5)	Enea, 2023	https://CRAN.R-project.org/package=speedglm
KSgeneral package for R	Dimitrova, et al., 2020	https://doi.org/10.18637/jss.v095.i10
dqrng package for R (version 0.3.1)	Stubner et al., 2023	https://CRAN.R-project.org/package=dqrng
qvalue package for R (version 2.24.0)	Storey et al., 2021	http://github.com/jdstorey/qvalue
matrixStats package for R	Bengtsson, 2023	https://CRAN.R-

(version 1.0.0)		project.org/package=matrixStats
Hmisc package for R (version 5.1-1)	Harrell Jr, 2023	https://CRAN.R-project.org/package=Hmisc
caret package for R	Kuhn, 2008	https://doi.org/10.18637/jss.v028.i05
olsrr package for R (version 0.5.3)	Hebbali, 2020	https://CRAN.R-project.org/package=olsrr
wCorr package for R (version 1.9.8)	Bailey et al., 2023	https://CRAN.R-project.org/package=wCorr
EmpiricalBrownsMethod package for R (version 1.9)	Poole, 2023	https://bioconductor.org/packages/EmpiricalBrownsMethod
survminer package for R (version 0.4.9)	Kassambara et al., 2021	https://CRAN.R-project.org/package=survminer
survival package for R (version 3.5-5)	Therneau, 2023	https://CRAN.R-project.org/package=survival
scikit-learn package for Python	Pedregosa et al., 2011	https://api.semanticscholar.org/CorpusID:10659969
numpy package for Python	Harris et al., 2020	https://doi.org/10.1038/s41586-020-2649-2
Benchmarking and Parallelization		
tictoc package for R (version 1.2)	Izrailev, 2023	https://CRAN.R-project.org/package=tictoc
snow package for R (version 0.4-4)	Tierney, 2021	https://CRAN.R-project.org/package=snow
foreach package for R (version 1.5.2)	Microsoft, 2022	https://CRAN.R-project.org/package=foreach
doParallel package for R (version 1.0.17)	Corporation et al., 2022	https://CRAN.R-project.org/package=doParallel
Data Visualization		
igraph package for R (version 1.5.0.1)	Csardi et al., 2006	https://CRAN.R-project.org/package=igraph
gplots package for R (version 3.1.3)	Warnes et al., 2022	https://CRAN.R-project.org/package=gplots
VennDiagram package for R	Chen, 2022	https://CRAN.R-

(version 1.7.3)		project.org/package=VennDiagram
RColorBrewer package for R (version 1.1-3)	Neuwirth, 2022	https://CRAN.R-project.org/package=RColorBrewer
ggrepel package for R (version 0.9.4)	Slowikowski, 2023	https://CRAN.R-project.org/package=ggrepel
ggsci package for R (version 3.0.0)	Xiao, 2023	https://CRAN.R-project.org/package=ggsci
ggraph package for R (version 2.1.0)	Pedersen, 2022	https://CRAN.R-project.org/package=ggraph
tidygraph package for R (version 1.2.3)	Pedersen, 2022	https://CRAN.R-project.org/package=tidygraph
UpSetR package for R (version 1.4.0)	Gehlenborg, 2019	https://CRAN.R-project.org/package=UpSetR
pheatmap package for R (version 1.0.12)	Kolde, 2019	https://CRAN.R-project.org/package=pheatmap
cowplot package for R (version 1.1.1)	Wilke, 2020	https://CRAN.R-project.org/package=cowplot
Miscellaneous		
argparse package for R	Davis, 2023	https://CRAN.R-project.org/package=argparse

Quantification and Statistical Analysis

All statistical analyses were performed within the R platform for statistical computing. All analysis scripts and scripts to recreate each figure are made available at GitHub (DOIs are listed in the key resources table). Quantification methods and statistical analyses for the omics datasets are described in the respective sections of the STAR Methods. Unless otherwise stated, relevant statistical parameters are reported in the legend of each figure.

I. A Framework to Estimate Regression Coefficients for the Nonsynonymous Mutation Status of Driver Genes. We introduce a linear regression framework to estimate relationships between the

mutation status of driver genes and the expression of each putative target gene, *t*, across a set of cancer samples. In particular, for each putative target gene *t*, we consider the following model:

$$E_t \sim \sum_{d \in S}(\alpha_d U_d + \beta_d C_d + \gamma_d Y_d) + \delta U_t + \zeta C_t + \eta Y_t + \sum_{i=1}^{K} \theta_i X_i + \varepsilon_t$$

where *S* is the set of frequently mutated driver genes considered (Table S1, Methods IV.E), E_t is a continuous value representing the expression of putative target gene *t* (Methods IV.A), U_d is a binary variable indicating whether driver gene *d* possesses a nonsynonymous mutation (Methods IV.B), C_d is a continuous value representing the normalized copy number status of *d* (Methods IV.C), and Y_d is a continuous value representing the methylation status of *d* (Methods IV.D). Similarly, U_t is a binary variable indicating whether the putative target gene *t* possesses a nonsynonymous mutation, C_t is a continuous value representing the normalized copy number status of *t,* and Y_t is a continuous value representing the methylation status of *t.* In addition to these core features, we also have a number of additional covariates that correspond to various other clinical and molecular features that may have nontrivial effects on E_t, with the number *K* of such covariates dependent upon the characteristics of the set of samples being examined (Methods IV.F). These covariates include the cancer type and subtype, age, gender, genotypic background, prior malignancies, prior treatment, nonsynonymous tumor mutational burden (TMB), tumor purity, and fraction of infiltrating immune cells. The coefficients fit from the data are the $\alpha_d, \beta_d, \gamma_d$ and $\{\theta_1, \theta_2, \ldots, \theta_K\}$, as well as δ, ζ, and η.

II. Applying Regression Framework to Samples from the TCGA, Pan-Cancer and Within 19 Individual Cancer Types. We apply the multiple linear regression model from Methods I to all putative target genes *t* in the human genome (Methods IV.A), with one model per *t*. We do this both across all TCGA samples, or "pan-cancer", as well as within the 19 individual cancer types possessing ≥75 samples (Table S1). For both the pan-cancer analysis and the per-cancer analyses, each model includes U_d, C_d, and Y_d features for all driver genes *d* as annotated by

Vogelstein et al.[25] that have nonsynonymous mutations in at least 5% of available samples and at least 5 total samples; see Table S1 for the set of drivers tested in each case. In the pan-cancer case, drivers must be mutated in at least 5 samples, as well as at least 5% of all samples both pan-cancer and within at least 2 individual cancer types (to ensure signal is not driven by a single cancer type). This 5% threshold was chosen in an effort to balance statistical power with including as many potentially interesting driver genes as possible. All multiple regression models were run using the R package speedglm's *speedlm* function[87], which fit coefficients and provided associated *p*-values for all terms, including the U_d mutational terms of interest. For each driver gene, we performed multiple hypothesis correction across the set of coefficients corresponding to its mutational term to convert *p*-values across the targets to *q*-values, using the *qvalue* function from the qvalue package in R with default parameters[88]. We deemed pairings between a nonsynonymous mutation in driver gene *d* and the expression of putative target gene *t* significant if the corresponding *q*-value was less than a threshold value of 0.2; given the increased statistical power in the pan-cancer setting, in the main body of the paper, we report pairings that are significant using a threshold of 0.01 for pan-cancer analyses and 0.2 for per-cancer analyses.

III. Addressing Multicollinearity. In certain cases, we expect the nonsynonymous mutation status of driver gene *d*, represented as U_d, to result in correlations with other variables in our framework (e.g., mutations within *IDH1* promote hypermethylation[89], and thus *IDH1* mutation status in some cancers may be correlated with the methylation status of genes). Linear regression models assume independence between variables, and in such cases will not be able to disentangle the contributions of the mutation within the driver gene from other variables it is correlated with. As such, prior to running the regression, our framework checks for multicollinearity.

First, we check that cancer subtype covariates (Methods IV.F.h) are not correlated with the mutation status of any driver genes present in the given regression model. If they are, this would indicate that these subtypes were at least partially defined by driver mutation status. To

address this, we generate a Spearman correlation coefficient matrix and associated *p*-value matrix using the R package Hmisc, and check, for each cancer subtype variable, if it is correlated with any U_d variable with Spearman correlation >0.7 and *p*-value <1E-05. If this correlation meets both of these exclusion criteria, the subtype variable is removed from the regression model. We use Spearman correlation for subtype variables, rather than other commonly used measures of multicollinearity such as the variance inflation factor (VIF)[90], because subtypes are encoded as bucketed, binary variables (Methods IV.F.h), and thus the corresponding variables are correlated with each other and will have high variance inflation factors (VIFs), even if they are not correlated with the mutation status of the driver gene. Across targets, we find that the vast majority of variables removed using this procedure pan-cancer (~99.9%) correspond to subtypes that are defined by IDH1-mutation status in LGG, particularly the IDH1mut-non-codel subtype.

Following this, we check other variables in the model for multicollinearity using the VIF. We use the R package caret's *vif* function to calculate a VIF measure for all non-bucketed variables in the regression, since as described above, bucketed variables (Methods IV.F.f-i) by design are collinear with one another and have high VIF scores. For the remainder of these variables, we eliminate any non-driver mutation (U_d) variables whose VIF score exceeds a threshold of 5, a generally conservative but widely accepted threshold that suggests moderate collinearity[39]. To ensure that our variables of interest, U_d, are not collinear with other variables in the model, we repeat this process iteratively until U_d for all *d* in *S* are below the threshold VIF of 5. In pan-cancer analyses, 13,917 genes (84.6%) had at least one variable removed, though 13,384 of these genes (96.2% of the 13,917 cases, 81.3% of target genes overall) had only the IDH1mut-non-codel subtype variable removed. Aside from this, we find that the methylation status of the target gene is most commonly eliminated pan-cancer (for ~2.10% of tested genes), followed by gender (~1.51%). For all significant pairings, variables that were eliminated using either of the above techniques are reported in Table S6 and Fig. S5B.

IV. Data Acquisition and Processing. In the National Cancer Institute's GDC data portal, there are 6378 primary tumor samples that possess all data types of interest, including annotated somatic mutation data, transcriptomic profiling, copy number variation, methylation, and clinical data. We downloaded these files from the GDC data repository, with parameters provided in Table S7.

A. *Expression Data.* Raw read count RNA-seq files from the TCGA were converted to counts per million (CPM) using edgeR's *cpm* function[91]. Genes minimally expressed across all samples were eliminated using edgeR's *filterByExpr* function with default parameters. These include requirements that all genes are required to have a minimum overall total count of at least 15 (*min.total.count*), and a minimum CPM of 10 in at least 10 samples (*min.count*). The remaining 19,052 genes' counts were quantile normalized so that each of the 6378 samples has the same distribution of gene expression values using scikit learn's *quantile.transform* function[92], with an output distribution of 'normal.' Each gene *t*'s expression level in Dyscovr corresponds to a quantile-normalized gene expression value E_t.

B. *Mutation Data.* We imported the simple nucleotide variation (SNV) file, with mutations called using *muse*[93], into R using maftools' *read.maf* function[94]. We then subsetted this file to include only nonsynonymous mutations (as annotated by *muse*), including missense, nonsense, nonstop, and splice site mutations. We then used maftools' *mutCountMatrix* function to compute the number of nonsynonymous mutations per gene and per sample. We excluded samples with excessively high mutation rates across all genes, referred to here as 'hypermutators', which we defined according to the analyses performed in Campbell et al.[95]. In this work, they used a linear regression approach across 81,337 cancer patients to determine a reasonable threshold for hypermutation, which they recommend being ~10 mutations per Mb. Given that the human exome is ~36.8Mb, we selected 368 mutations as our threshold for hypermutation, such that any sample with greater than 368 total nonsynonymous mutations was discarded from the analysis. This

removed a total of 360 samples, leaving 6018 pan-cancer TCGA samples. The mutation status of each gene *t* in each sample, U_t , is 1 if the gene has a nonsynonymous mutation in that sample and 0 otherwise.

C. *Copy Number Alteration (CNA) Data*. We obtained absolute CNA values for each gene in each sample, as computed by the *ASCAT*[96] pipeline and made available by TCGA. In our model, for each gene in a given sample, we compute its normalized copy number value as the log_2 of the absolute CNA value divided by the mean CNA value across genes in the sample. Pseudocounts of 1 were used to adjust both the gene-level and average copy number values. The normalized copy number gives us a value that accounts for large scale ploidy differences between tumor samples. In practice, to be robust to extreme outlier CNA events, we exclude the top 10% and bottom 10% of gene-level CNA values for each sample when calculating its mean CNA value.

D. *Methylation Data*. We imported individual samples' level 3 Liftover methylation Beta (β) files into R and removed 'NA' or empty values. These files have already been processed to include gene-level annotations. For any gene with more than one reported β value, we averaged these β values to produce a single β value per gene. Subsequently, we compiled the average β value for each gene in each sample and then converted β values to *M*-values using the logit, or $log2(\beta \ / \ 1 - \beta))$. We represent methylation levels in our model using these M-values. We use the *M*-value rather than the Beta value to assess methylation due to its purported statistical rigor[97] and improved model performance in early testing.

E. *Gene Set Data*. Our list of driver genes consists of genes from Vogelstein et al. tables S2A, S2B, S3A, S3B, S3C and S4[25].

F. *Clinical Data*. Using data from the TCGA's clinical supplement, we created a composite data table with patient-level features. Using data from TCGA's biosample data supplement, a TCGA tumor purity file obtained from Aran et al.[98], genotypic principal

components from Carrot-Zhang et al.[99], and immune cell infiltration estimates from R's immunedeconv package[100] using the tool CIBERSORT Abs[101], we created a data table with sample-level features. We then combined the patient- and sample-specific files such that all patient-specific data was applied to all samples from that patient, for use in the linear regression framework. Further information on each clinical feature is described in subsections below.

a. *Age.* Normalized to take on a value approximately between 0 and 1 by dividing each patient's age (given in years) by 100.
b. *Gender.* Takes on a binary value of 1 for male and 0 for female patients.
c. *Prior malignancies.* Takes on binary value, with 1 signifying that the patient had a prior malignancy and 0 signifying that there was no prior malignancy.
d. *Prior treatment.* Encompasses two binary variables, as per the data available in the TCGA's clinical supplement: prior radiation treatment, where 0 represents no prior radiation treatment and 1 represents prior radiation treatment, and prior pharmaceutical treatment, where 0 represents no prior pharmaceutical treatment and 1 represents prior pharmaceutical treatment.
e. *Genotypic Principal Components*. Consists of three continuous covariates, each of which corresponds to the value of one of the first three genotypic principal components (PCs). These PCs are provided in the supplemental data table of Carrot-Zhang et al.[99] (see Star Methods Key Resources Table) and were calculated using the Washington University approach. Briefly, this approach involves the conversion of Birdseed genotype files to individual VCF files, which are then merged (with only variants of MAF >15% retained) prior to PCA using PLINK 1.9.[102] See Carrot-Zhang et al.[99] for more detailed methods.
f. *Nonsynonymous Mutational Burden.* Consists of three binary covariates, representing low, moderate, and high nonsynonymous tumor mutational burden

(TMB). For any given sample, one of these covariates will take on a value of 1 and the other two a value of 0, depending on the total number of nonsynonymous mutations their tumor possesses. Based on the distribution of TMB across TCGA samples, we defined low TMB as having <30 nonsynonymous mutations, moderate TMB as between 30 and 60 mutations, and high TMB as >60 mutations.

g. *Tumor Purity.* Consists of three binary covariates, representing low, moderate, and high tumor purity. For any given sample, one of these covariates will take on a value of 1 and the other two a value of 0, depending on the magnitude of the tumor purity estimate for that sample. We define purity using the combined purity estimate (CPE) from Aran et al.[98] when available, and for the samples without a provided CPE, we use the median of the available purity measures. Based on the distribution of CPE values across TCGA samples, we defined low tumor purity as having a CPE ≤ 0.5, moderate tumor purity as having a 0.5 < CPE ≤ 0.75, and high tumor purity as having a CPE > 0.75.

h. *Tumor Subtype*. Consists of a binary covariate for each subtype in the given cancer type. For any given sample, one of these covariates will take on a value of 1 and all others a value of 0, depending on which subtype it is classified as. Because subtypes are defined differently for each cancer type, the column we used to define subtype in the TCGA clinical supplement is provided in Table S8. Each cancer type will have a different number of tumor subtype variables, as this is dependent on the number of unique subtypes present in the given column. Molecular subtypes were used whenever possible; when unavailable, histological or expression-based clustering subtypes were used. In the pan-cancer analyses, a binary covariate was created for each cancer type:subtype combination (e.g. Breast Cancer:Luminal A), across all cancer types. In this case, for any given sample, one of these

combination covariates will take on a value of 1 and all others 0, depending on its combined cancer type and subtype classification.

i. *Immune Cell Infiltration*. In the case of immune cell infiltration (F), we use the absolute immune cell fractions provided by the tool CIBERSORT Abs[101], run using R's immunedeconv package[100]. To get a single value representing the level of immune cell infiltration in the given sample, we add individual immune cell type fractions, e.g. predicted fractions of B cells, T cells, etc., into a total fraction of immune cells per sample. From there, we group this fraction into one of three buckets: low immune cell infiltration (total immune cell fraction ≤0.3), medium immune cell infiltration (total immune cell fraction >0.3 and ≤0.7), and high immune cell infiltration (total immune cell fraction >0.7), again defined by the distribution of total immune cell fractions across the samples.

G. In all of the cases that involve two or more binary covariates, and in which only one of these covariates can take on a value of 1 (Methods IVFf-i), one fewer covariate is needed in the linear regression equation to represent all distinct possible classifications[103]. To limit multicollinearity, the final covariate for a given feature is removed (e.g. a dummy variable encoding, rather than a one-hot encoding). For example, in the case of nonsynonymous TMB, low TMB would be represented by a binary covariate taking on values of 1 and 0, moderate TMB would be represented by a binary covariate taking on values of 0 and 1, and high TMB would be represented by two binary covariates taking on values of 0 and 0.

V. Computational Validation via Comparison of TCGA BRCA results to METABRIC. To test whether our models are capable of generating consistent and meaningful correlations between driver gene mutations and target gene expression across independent cohorts, we also applied the Dyscovr framework to data from the Molecular Taxonomy of Breast Cancer International

Consortium (METABRIC)[38], a collection of 854 breast cancer patients of European ancestry with paired mutation, CNA, mRNA expression, methylation, and clinical data. Due to differences in available data types, we made the following modifications to our model and data processing.

A. METABRIC: Estimating Regression Coefficient for Mutation Status of Driver Genes. To ensure comparability to TCGA-BRCA, linear regression models were constructed in the same fashion as described in Methods I. Due to limitations in available germline data for the METABRIC cohort, genotypic principal components are not included in these models.

B. METABRIC: Data Acquisition and Processing. METABRIC primary tumor data files were downloaded from cBioPortal[104]. To keep processing pipelines as similar as possible to TCGA-BRCA, the same protocol was used for preprocessing mutation data, including generation of a mutation count, restriction to only nonsynonymous mutations, and removal of hypermutators (Methods IV.B). Mutations in METABRIC were called using MuTect and filtered according to the procedure given in Curtis et al.[38]. As with TCGA-BRCA, *ASCAT*[96] was used for CNA calling. The gene expression data available for the METABRIC cohort is Illumina HT-12 v3 microarray data, which requires distinct preprocessing procedures to TCGA-BRCA's RNA-sequencing data. We filtered genes with greater than 50% missing values and with mean expression <5 or standard deviation <0.3 across samples, as in Liao et al.[105], but otherwise used the provided quantile-normalized *log2*-intensity values as input to our models. The METABRIC methylation data is also distinct from TCGA-BRCA in that only promoter methylation bisulfite sequencing (RRBS) is available. Files were already processed using the gpatterns package such that each file contains a [0,1] gene-level value of CpG methylation. Models were run using both these [0,1] CpG values, as well as the logit of these values, though models performed comparably and the [0,1] CpG values were ultimately used in figure creation. Due to differences in labeling and availability of clinical data types in METABRIC as compared to TCGA-BRCA, different

column names were often used; any features with notable differences are described below.

a. *Prior malignancies (Methods IV.F.c)*. The "RFS_STATUS" column was used.

b. *Prior treatment (Methods IV.F.d).* The "RADIO_THERAPY" column was used to create a binary representation of prior radiation treatment, while the "CHEMOTHERAPY" and "HORMONE_THERAPY" columns were combined to create a binary representation of prior pharmaceutical treatment. In the latter case, the sample received a 1 if they had received either chemotherapy or hormone therapy, and 0 otherwise.

c. *Nonsynonymous tumor mutational burden (TMB) (Methods IV.F.f).* For the nonsynonymous TMB, the *log2* of the "TMB_NONSYNONYMOUS" column with a pseudocount of 1 was used, with 3 binary covariates for representing low (TMB ≤ 2.5), moderate (2.5 < TMB ≤ 3.5) and high (TMB > 3.5) tumor mutational burden.

d. *Tumor purity (Methods IV.F.g)*. METABRIC's clinical supplement provides a "CELLULARITY" column that is an estimate of tumor purity from the tool MCP-counter[106]. The values this column can take include "Low", "Moderate", and "High". As with TCGA-BRCA, we represented this in our model as three binary covariates, with a sample taking on a value of 1 in one of these covariates and a 0 in the two others.

C. <u>METABRIC: Comparison of Betas to TCGA-BRCA.</u> For each driver gene mutated at ≥5% frequency in both TCGA-BRCA and METABRIC cohorts, which includes *TP53* and *PIK3CA*, we computed the Spearman correlation between the fit coefficients for all target genes tested in both TCGA-BRCA and METABRIC (Fig. 3C, Fig. S3C).

<u>VII. Gene Set Enrichment Analysis</u>.

We compute enrichment in curated gene sets (Fig. 1B-D; Table S3) using a one-sided Kolomogorov-Smirnov (K-S) test with a uniform distribution (H_A = greater). We use gene sets from the DoRothEA network[26] (Fig. 2B), a curated set of *TP53* targets from Fischer et al.[27] (Table 1, Fig. 2C) and Reactome[31] (Fig. 2C), and from the CGC[32] (Fig. 2D). The DoRothEA network was accessed via the dorothea R package using the *dorothea_hs* function. Genes at all confidence levels were used. Additional gene sets used for *TP53*-specific gene set enrichment analysis were downloaded directly from TRRUST[28], hTFtarget[29], and KEGG[30] websites (Table S3, STAR Methods).

We compute pathway-level gene set enrichment analysis (Fig. 4D; Fig. S1C; Table S4) using the package ReactomePA[107] in R, specifically using the *gseGO, gseKEGG,* and *gseMKEGG* functions applied to the -*log*(*q*-values) multiplied by the directionality of the associated mutation coefficient (1 for coefficient > 0, -1 for coefficient < 0) produced by Dyscovr. In cases with more than five significant GO pathways, functionally similar GO pathways[33] were consolidated using the Wang et al. method[108]; pathways with a similarity metric greater than 0.7 were merged, retaining the name of the more statistically significant pathway (Fig. S1C). Full sets of enriched GO pathways, without merging, can be found in Table S4.

VIII. Narrowing of Experimental Candidates Using the Cancer Dependency Map (DepMap)[24]. We use a combination of CRISPRi gene dependency data, mutation data, and gene expression data from DepMap public v.23Q2 to narrow down putative candidates from Dyscovr with those that are potentially synthetic lethal with the corresponding driver gene. For a given cancer driver gene *d,* we first limit *d*'s hits to those that are statistically significant both pan-cancer (q < 0.2) and within at least one individual cancer type (q < 0.2). For each of these remaining targets *t* in set *T*, we use the following regression framework to relate the cell viability upon CRISPRi knockdown of *t* (V_t) to the mutation status (U_d) and expression (E_d) of *d* across a set *K* of cancer types (X_i) (see Methods VIII.A.d):

$$V_t \sim \alpha U_d + \beta E_d + \sum_{i=1}^{K} \gamma_i X_i + \varepsilon_t$$

Overall, *T* regression models are fit for a given *d.* From here, we evaluate the significance of the combined effect of U_d and E_d–what we refer to as the 'activity' of *d*–on V_t using Empirical Brown's method (specifically, the *empiricalBrownsMethod* function from the EmpiricalBrownsMethod R package). This method is an adaptation of Fisher's method[109] that allows for interdependency of the terms. We use it with default parameters for the *p*-values associated with α and β, obtaining a single *p*-value describing the effect of the driver *d* on V_t. We obtain multiple hypothesis corrected *q*-values from these *p*-values using the *qvalue* function from the qvalue package in R with default parameters[88].

For each driver gene *d*, we next try to identify which of its predicted targets are most likely to be in synthetic lethality with it, as these relationships can be exploited in the clinical setting. For these target genes *t*, cell viability values upon knockout of *t* should be smaller when the driver gene is less active and larger when the driver gene is more active. In the case of an oncogene (*KRAS* and *PIK3CA*), we expect both a nonsynonymous mutation or high expression to result in cancer-promoting overactivity, and thus a promising synthetic lethal candidate would display a positive relationship between activity of *d* and cell viability upon knockout of *t* ($\alpha > 0$, $\beta > 0$). For tumor suppressor genes (*TP53*), a driver nonsynonymous mutation results in its underactivity. In this case, we expect a negative relationship between the mutation status of *d* and cell viability upon knockout of *t* ($\alpha < 0$, $\beta > 0$) (Fig. 4A). We report the set of targets *t* for which the recombined *q*-value is less than 0.2 and the coefficients α and β from the pre-recombined analysis align with the above schema.

A. DepMap: Data Acquisition and Processing. CRISPRi dependency, mutation, gene expression, and cell line metadata for cancer cell lines were downloaded from the DepMap online portal (see STAR methods table). Models were run using the 693 cancer cell lines with all data types available and the set of 3 TCGA pan-cancer drivers (*TP53, PIK3CA,*

and *KRAS*) that have nonsynonymous mutations in at least 15 cell lines across cell lines in the set of cancer types in which that driver is recurrently mutated (Table S1).

a. *CRISPRi Dependency Data*. Each target gene *t*'s dependency score upon genetic interference corresponds to the CRISPRi dependency value V_t. Dependency scores were processed using Chronos[110] (see Table 1), which normalizes the dependency score by copy number, eliminating the need to include target gene CNA status as a regression term. Positive dependency values indicate high cell viability upon knockout, while negative dependency values indicate reduced cell viability upon knockout.

b. *Expression Data*. *log2(*TPM + 1)-normalized RNA-seq files from DepMap were quantile normalized so that each of the cell lines has the same distribution of gene expression values using scikit learn's *quantile.transform* function[92], with an output distribution of 'normal.' Each driver gene *d*'s expression level corresponds to a quantile-normalized gene expression value E_d.

c. *Mutation Data*. The DepMap mutation file was subsetted to include only nonsynonymous mutations, selecting for 'VariantInfo' column annotations that include 'MISSENSE', 'NONSENSE', 'NONSTOP', and 'SPLICE_SITE'. The mutation status of each driver gene *d* in each sample, U_d , is 1 if the driver gene has a nonsynonymous mutation in that sample and 0 otherwise.

d. *Cancer Type*. Consists of a binary covariate for each cancer type. For any given cell line, one of these covariates will take on a value of 1 and all others a value of 0, depending on which cancer type it is classified as in the cell line metadata supplement column 'primary_disease'. Cell line data from a given cancer type is only considered if *d* is mutated in at least 5% of samples from that cancer type in the TCGA cohort (see Table S1).

B. Evaluation of Candidate Synthetic Lethals via Correlation to Drug Sensitivity. For putative synthetic lethal targets *t* of clinical interest (i.e. *KBTBD2*), we further used DepMap's repository of drug sensitivity data to compute a Spearman's correlation between the viability of the cell line on CRISPR interference of *t* (V_t) and the sensitivity to each of 4659 drugs. Associated *p*-values were corrected using the qvalue package in R with default parameters[88]. Significant correlation to sensitivity of drugs (q < 0.2) were used to infer potential mechanisms-of-action of *t*.

IX. Survival Analysis for *PIK3CA* Mutation and *KBTBD2* Expression. To evaluate whether the expression level of putative target gene *KBTBD2* (E_K) has an effect on patient survival (S), taking into consideration the binary nonsynonymous mutation status of cancer driver gene *PIK3CA* (U_P), we used the following Cox proportional hazards model:

$$S \sim \alpha E_K + \beta U_P + \sum_{i=1}^{K} \gamma_i X_i$$

As in the main Dyscovr framework, we also have a number of additional covariates that correspond to various other clinical and molecular features that may have nontrivial effects on S, with the number *K* of such covariates dependent upon the characteristics of the set of samples being examined (i.e. pan-cancer or breast cancer, Methods IV.F). These covariates include the cancer type and subtype, age, gender, genotypic background, prior malignancies, prior treatment, nonsynonymous tumor mutational burden (TMB), tumor purity, and fraction of infiltrating immune cells. This model was fit using the *coxph* function in the survminer package in R version 0.4.9[111] with default parameters. Hazard ratio estimates were computed as e^{α}, where α is the fit coefficient for the E_K term (above).

A. Data Acquisition and Representation. Survival, mutation, and gene expression data were obtained for all 785 pan-cancer patients (see Methods IV.A and IV.B for details about mutation and gene expression data acquisition and preprocessing) from the National

Cancer Institute's GDC Data Portal with reported information about survival status and days until death or last follow-up (see Methods IX.A.a) and have high or low *KBTBD2* expression (see Methods IX.A.c).

a. *Survival Data.* Survival data was extracted from the TCGA clinical supplement (see STAR Methods Table), including a) binary survival status (alive or dead), and b) time in days, defined as time until death if patient has died, or time until last follow-up if patient is alive. Survival S was represented as a *Surv* object from the survminer package for modeling.

b. *Mutation Data.* The mutation status of *PIK3CA* in each patient, U_P, is 1 if *PIK3CA* has a nonsynonymous mutation in that patients' sample and 0 otherwise.

c. *Expression Data*. The quantile-normalized expression matrix was further *z*-score normalized per patient column, such that the resulting matrix took on a mean (μ) of ~0 and a standard deviation (σ) of ~1. This formulation allowed us to select patients whose normalized expression value for *KBTBD2* is less than $\mu - \sigma$ (which we define as having "low" expression) or is greater than $\mu + \sigma$ (which we define as having "high" expression). Patients who do not fall into either of these categories are excluded from further analysis. In the Cox proportional hazards model, the expression status of *KBTBD2* in each patient, E_K, is 1 if the patient has "high" *KBTBD2* expression and 0 if the patient has "low" *KBTBD2* expression, as previously defined.

B. *Survival Visualization.* Adjusted patient survival curves were visualized using the *ggadjustedcurves* function from the survminer package using the default 'single' (average for population) approach[111]. Briefly, this approach displays expected survival curves calculated based on the Cox model fit.

X. Cell Lines. MCF7 (HTB-22) was purchased from American Type Culture Collection (ATCC) and maintained in DMEM supplemented with 10% Fetal Bovine Serum (FBS) and 1% penicillin and streptomycin.

XI. Immunoblotting. Cells were collected in ice cold PBS and lysed with RIPA lysis buffer (Pierce #89901) supplemented with Halt protease and phosphatase inhibitors (Pierce Chemical). Lysates were centrifuged at 20,000 × g for 5 minutes at 4°C. The supernatant was collected, and protein concentration was determined using the BCA kit (Pierce) per manufacturer's instructions. Equal amounts of protein (20µg) in cell lysates were separated by SDS–PAGE, transferred to nitrocellulose membranes (GE healthcare), immunoblotted with specific primary and secondary antibodies and detected by chemiluminescence with the ECL detection reagents from Thermo Fisher or Millipore.

XII. Antibodies. pAKT S473 (Cell Signaling Technology (CST) 4060), pS6 S235/S236 (CST 4858), p4EBP1 S65 (CST 9451), Cyclin D1 (CST 55506), p85 alpha (CST 4292), Beta Actin (CST 4967). Secondary Goat anti-Rabbit IgG (H+L) Secondary Antibody HRP (Thermo Fisher 65-6120).

XIII. siRNA Knockdown. Cells were transfected for 48hr with Dharmacon SMARTpool nontargeting or siRNA designed against human KBTBD2. Transfection was aided by preincubation of siRNA with lipofectamine RNAiMAX (Thermo Fisher Scientific) and used according to the manufacturer's instructions.

XIV. mRNA extraction and RT-qPCR. mRNA was isolated using RNeasy kit (Qiagen) with Qiaschredder and eluted in 50uL. cDNA was synthesized using SuperScript First-Strand Synthesis System for RT-PCR kit (ThermoFisher #11904018). Synthesis was performed using random primers. Probes and primers were obtained from ThermoFisher as follows: KBTBD2

(Assay ID: Hs01556149_m1, FAM-MGB and GAPDH (Assay ID: Hs02786624_g1, VIC-MGB). qPCR was performed in 20uL reaction volumes with *KBTBD2* assay together with *GAPDH* and amplified using iTaq universal probes supermix (Bio-Rad Laboratories, #1725134). Relative quantification was performed using (2–ΔΔCt).

XV. Quantification of Cell Growth and Viability. MCF7 cells were seeded into 24-well plates at 50,000 cells per well and transfected as indicated above. Cell growth was quantified using the sulforhodamine B assay. For each condition at least 3 replicates were measured. Growth at day 0 was subtracted from all Day 4 values and growth was normalized to Veh. treated samples.

Resource Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Mona Singh (mona@cs.princeton.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Additional Resources

Dyscovr model results from TCGA patients are publicly searchable and downloadable at: dyscovr.princeton.edu.

Supplemental Information

Table S1: Vogelstein Drivers Mutated at ≥5% Frequency Per TCGA Cancer Type.

Table S2: *PIK3CA, KRAS,* and *IDH1* Pan-Cancer Enrichments in Effector TF Targets.

Table S3: *TP53* Pan-Cancer Enrichments in Gold-Standard Target Sets.

Table S4: GO and KEGG per-Driver Pathway Enrichment, Unmerged.

Table S5: Putative SL Pan-Cancer Targets for *TP53, PIK3CA, KRAS.*

Table S6: Variables Removed Due to Multicollinearity, Pan-Cancer ($q < 0.01$).

Table S7: Parameters for GDC Data Portal File Downloads.

Table S8: Per-Cancer TCGA Subtype Information.

References

1. Huang, L., Guo, Z., Wang, F. & Fu, L. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct. Target. Ther.* **6**, 386 (2021).
2. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24**, 869–884 (2014).
3. Jiang, P., Freedman, M. L., Liu, J. S. & Liu, X. S. Inference of transcriptional regulation in cancers. *Proc. Natl. Acad. Sci.* **112**, 7731–7736 (2015).
4. Basha, O., Mauer, O., Simonovsky, E., Shpringer, R. & Yeger-Lotem, E. ResponseNet v.3: revealing signaling and regulatory pathways connecting your proteins and genes across human tissues. *Nucleic Acids Res.* **47**, W242–W247 (2019).
5. Logsdon, B. A. *et al.* Sparse expression bases in cancer reveal tumor drivers. *Nucleic Acids Res.* **43**, 1332–1344 (2015).
6. Grechkin, M., Logsdon, B. A., Gentles, A. J. & Lee, S.-I. Identifying Network Perturbation in Cancer. *PLOS Comput. Biol.* **12**, e1004888 (2016).
7. Knaack, S. A., Siahpirani, A. F. & Roy, S. A Pan-Cancer Modular Regulatory Network Analysis to Identify Common and Cancer-Specific Network Components. *Cancer Inform.* **13s5**, CIN.S14058 (2014).
8. Kurup, J. T., Kim, S. & Kidder, B. L. Identifying Cancer Type-Specific Transcriptional Programs through Network Analysis. *Cancers* **15**, 4167 (2023).
9. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **13**, R124 (2012).
10. Bertrand, D. *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* **43**, e44–e44 (2015).
11. Wang, Z. *et al.* Cancer driver mutation prediction through Bayesian integration of multi-omic data. *PLOS ONE* **13**, e0196939 (2018).
12. Cutigi, J. F., Evangelista, A. F., Reis, R. M. & Simao, A. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. *Sci. Rep.* **11**, 23551 (2021).
13. Frost, H. R. & Amos, C. I. A multi-omics approach for identifying important pathways and genes in human cancer. *BMC Bioinformatics* **19**, 479 (2018).
14. Khalighi, S. *et al.* SYSMut: decoding the functional significance of rare somatic mutations in cancer. *Brief. Bioinform.* **23**, bbac280 (2022).
15. Ding, J. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.* **6**, 8554 (2015).
16. Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013).
17. Hou, J. P. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome Med.* **6**, 56 (2014).
18. Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640–i646 (2012).
19. Chen, Y. *et al.* Identification of Druggable Cancer Driver Genes Amplified across TCGA Datasets. *PLoS ONE* **9**, e98293 (2014).

20. Osmanbeyoglu, H. U., Toska, E., Chan, C., Baselga, J. & Leslie, C. S. Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat. Commun.* **8**, 14249 (2017).
21. Sousa, A. *et al.* Pan-Cancer landscape of protein activities identifies drivers of signalling dysregulation and patient survival. *Mol. Syst. Biol.* **19**, e10631 (2023).
22. Tao, Y. *et al.* Interpretable deep learning for chromatin-informed inference of transcriptional programs driven by somatic alterations across cancers. *Nucleic Acids Res.* **50**, 10869–10881 (2022).
23. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
24. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).
25. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
26. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
27. Fischer, M. Census and evaluation of p53 target genes. *Oncogene* **36**, 3943–3956 (2017).
28. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
29. Zhang, Q. *et al.* hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. *Genomics Proteomics Bioinformatics* **18**, 120–128 (2020).
30. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
31. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).
32. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
33. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
34. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
35. Subhasree, N., Jiangjiang, Q., Kalkunte, S., Minghai, W. & Ruiwen, Z. The MDM2-p53 pathway revisited. *J. Biomed. Res.* **27**, 254 (2013).
36. Cagnol, S. & Rivard, N. Oncogenic KRAS and BRAF activation of the MEK/ERK signaling pathway promotes expression of dual-specificity phosphatase 4 (DUSP4/MKP2) resulting in nuclear ERK1/2 inhibition. *Oncogene* **32**, 564–576 (2013).
37. Li, R.-Z. *et al.* The key role of sphingolipid metabolism in cancer: New therapeutic targets, diagnostic and prognostic values, and anti-tumor immunotherapy resistance. *Front. Oncol.* **12**, 941643 (2022).
38. METABRIC Group *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

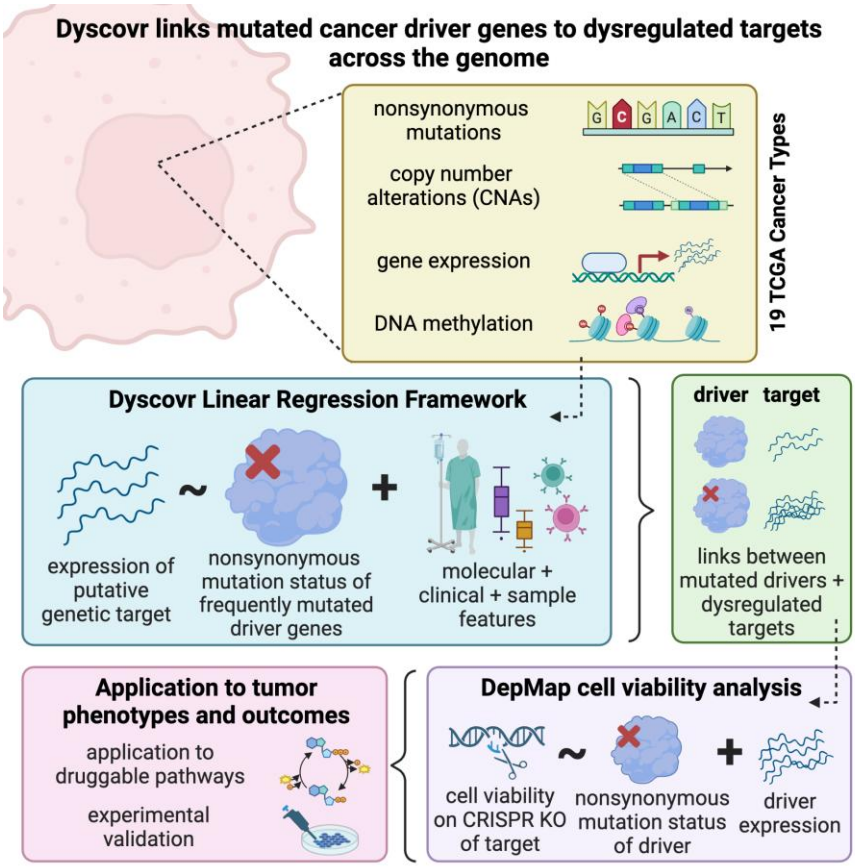
39. Yu, K. *et al.* PIK3CA variants selectively initiate brain hyperactivity during gliomagenesis. *Nature* **578**, 166–171 (2020).
40. Sun, X. *et al.* FYCO1 regulates migration, invasion, and invadopodia formation in HeLa cells through CDC42/N-WASP/Arp2/3 signaling pathway. *Biochem. Cell Biol.* **100**, 458–472 (2022).
41. Jerby-Arnon, L. *et al.* Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality. *Cell* **158**, 1199–1209 (2014).
42. Wang, J. *et al.* SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database* **2022**, baac030 (2022).
43. del Toro, N. *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* **50**, D648–D653 (2022).
44. Manchado, E. *et al.* A combinatorial strategy for treating KRAS-mutant lung cancer. *Nature* **534**, 647–651 (2016).
45. Kong, Y. *et al.* Mutant *KRAS* Mediates circARFGEF2 Biogenesis to Promote Lymphatic Metastasis of Pancreatic Ductal Adenocarcinoma. *Cancer Res.* **83**, 3077–3094 (2023).
46. Relógio, A. *et al.* Ras-Mediated Deregulation of the Circadian Clock in Cancer. *PLoS Genet.* **10**, e1004338 (2014).
47. Khan, T. *et al.* CUB Domain-Containing Protein 1 (CDCP1) is a rational target for the development of imaging tracers and antibody-drug conjugates for cancer detection and therapy. *Theranostics* **12**, 6915–6930 (2022).
48. Khan, T., Kryza, T., Lyons, N. J., He, Y. & Hooper, J. D. The CDCP1 Signaling Hub: A Target for Cancer Detection and Therapeutic Intervention. *Cancer Res.* **81**, 2259–2269 (2021).
49. Cai, Z. *et al.* The Skp2 Pathway: A Critical Target for Cancer Therapy. *Semin. Cancer Biol.* **67**, 16–33 (2020).
50. Prieto-Garcia, C., Tomašković, I., Shah, V. J., Dikic, I. & Diefenbacher, M. USP28: Oncogene or Tumor Suppressor? A Unifying Paradigm for Squamous Cell Carcinoma. *Cells* **10**, 2652 (2021).
51. Chen, Y. *et al.* CHEK2 knockout is a therapeutic target for TP53-mutated hepatocellular carcinoma. *Cell Death Discov.* **10**, 37 (2024).
52. Chrisanthar, R. *et al.* CHEK2 Mutations Affecting Kinase Activity Together With Mutations in TP53 Indicate a Functional Pathway Associated with Resistance to Epirubicin in Primary Breast Cancer. *PLoS ONE* **3**, e3062 (2008).
53. Choi, H. & Kang, C. Living Beyond Restriction: LBR promotes cellular immortalization by suppressing genomic instability and senescence. *FEBS J.* **291**, 2091–2093 (2024).
54. Evangelisti, C. *et al.* The wide and growing range of lamin B-related diseases: from laminopathies to cancer. *Cell. Mol. Life Sci.* **79**, 126 (2022).
55. Funauchi, Y. *et al.* Regulation of iron homeostasis by the p53-ISCU pathway. *Sci. Rep.* **5**, 16497 (2015).
56. Zhang, J. & Chen, X. p53 tumor suppressor and iron homeostasis. *FEBS J.* **286**, 620–629 (2019).
57. Li, Z. & Xiong, Y. Cytoplasmic E3 ubiquitin ligase CUL9 controls cell proliferation, senescence, apoptosis and genome integrity through p53. *Oncogene* **36**, 5212–5218 (2017).
58. Huang, X. *et al. NCDN* is a Potential Biomarker and Therapeutic Target for Glioblastoma. *J. Cancer* **15**, 1067–1076 (2024).

59. Ge, S. *et al.* PSME4 Activates mTOR Signaling and Promotes the Malignant Progression of Hepatocellular Carcinoma. *Int. J. Gen. Med.* **Volume 15**, 885–895 (2022).
60. Yazgili, A. S., Ebstein, F. & Meiners, S. The Proteasome Activator PA200/PSME4: An Emerging New Player in Health and Disease. *Biomolecules* **12**, 1150 (2022).
61. Ma, W., Zhang, X., Ma, C. & Liu, P. Highly expressed FAM189B predicts poor prognosis in hepatocellular carcinoma. *Pathol. Oncol. Res.* **28**, 1610674 (2022).
62. Wang, P. & He, X. Oncogenic and prognostic role of CKAP2L in hepatocellular carcinoma. *Int. J. Clin. Exp. Pathol.* **13**, 923–933 (2020).
63. Dos Santos, A., Ouellete, G., Diorio, C., Elowe, S. & Durocher, F. Knockdown of CKAP2 Inhibits Proliferation, Migration, and Aggregate Formation in Aggressive Breast Cancer. *Cancers* **14**, 3759 (2022).
64. Basu, R. & Kopchick, J. J. The effects of growth hormone on therapy resistance in cancer. *Cancer Drug Resist.* (2019) doi:10.20517/cdr.2019.27.
65. Kong, R. *et al.* Myo9b is a key player in SLIT/ROBO-mediated lung tumor suppression. *J. Clin. Invest.* **125**, 4407–4420 (2015).
66. Fu, F. *et al.* Role of Transmembrane 4 L Six Family 1 in the Development and Progression of Cancer. *Front. Mol. Biosci.* **7**, 202 (2020).
67. Xia, S. *et al.* SLC7A2 deficiency promotes hepatocellular carcinoma progression by enhancing recruitment of myeloid-derived suppressors cells. *Cell Death Dis.* **12**, 570 (2021).
68. Wenzel, J. *et al.* Loss of the nuclear Wnt pathway effector TCF7L2 promotes migration and invasion of human colorectal cancer cells. *Oncogene* **39**, 3893–3909 (2020).
69. Cizkova, M. *et al.* Gene Expression Profiling Reveals New Aspects of PIK3CA Mutation in ERalpha-Positive Breast Cancer: Major Implication of the Wnt Signaling Pathway. *PLoS ONE* **5**, e15647 (2010).
70. Lee, J.-S. *et al.* The insulin and IGF signaling pathway sustains breast cancer stem cells by IRS2/PI3K-mediated regulation of MYC. *Cell Rep.* **41**, 111759 (2022).
71. Day, E. *et al. IRS2* is a candidate driver oncogene on 13q34 in colorectal cancer. *Int. J. Exp. Pathol.* **94**, 203–211 (2013).
72. Savage, S. L. *et al.* Activating Mutations of Insulin Receptor Substrate 2 (IRS2) in Patients with Tyrosine Kinase Inhibitor-Refractory Chronic Myeloid Leukemia. *Blood* **126**, 2461–2461 (2015).
73. Alfaro-Arnedo, E. *et al.* IGF1R acts as a cancer-promoting factor in the tumor microenvironment facilitating lung metastasis implantation and progression. *Oncogene* **41**, 3625–3639 (2022).
74. Wang, P., Mak, V. Cy. & Cheung, L. Wt. Drugging IGF-1R in cancer: New insights and emerging opportunities. *Genes Dis.* **10**, 199–211 (2023).
75. Hu, Y. *et al.* Dynamic molecular architecture and substrate recruitment of cullin3–RING E3 ligase CRL3KBTBD2. *Nat. Struct. Mol. Biol.* **31**, 336–350 (2024).
76. Werner, A. *et al.* Cell-fate determination by ubiquitin-dependent regulation of translation. *Nature* **525**, 523–527 (2015).
77. Zhang, Z. *et al.* Insulin resistance and diabetes caused by genetic or diet-induced KBTBD2 deficiency in mice. *Proc. Natl. Acad. Sci.* **113**, (2016).

78. Juric, D. *et al.* Phosphatidylinositol 3-Kinase α–Selective Inhibition With Alpelisib (BYL719) in *PIK3CA* -Altered Solid Tumors: Results From the First-in-Human Study. *J. Clin. Oncol.* **36**, 1291–1299 (2018).
79. Elkabets, M. *et al.* mTORC1 Inhibition Is Required for Sensitivity to PI3K p110α Inhibitors in *PIK3CA* -Mutant Breast Cancer. *Sci. Transl. Med.* **5**, (2013).
80. Zou, Z., Tao, T., Li, H. & Zhu, X. mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. *Cell Biosci.* **10**, 31 (2020).
81. Averous, J., Fonseca, B. D. & Proud, C. G. Regulation of cyclin D1 expression by mTORC1 signaling requires eukaryotic initiation factor 4E-binding protein 1. *Oncogene* **27**, 1106–1113 (2008).
82. Varkaris, A. *et al.* Discovery and Clinical Proof-of-Concept of RLY-2608, a First-in-Class Mutant-Selective Allosteric PI3Kα Inhibitor That Decouples Antitumor Activity from Hyperinsulinemia. *Cancer Discov.* **14**, 240–257 (2024).
83. Geeleher, P. *et al.* Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biol.* **19**, 130 (2018).
84. Saito, M., Momma, T. & Kono, K. Targeted therapy according to next generation sequencing-based panel sequencing. *FUKUSHIMA J. Med. Sci.* **64**, 9–14 (2018).
85. Ahlmann-Eltze, C., Huber, W. & Anders, S. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods. Preprint at https://doi.org/10.1101/2024.09.16.613342 (2024).
86. Guerrero, S. *et al.* Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies. *Sci. Rep.* **8**, 13978 (2018).
87. Enea, M. speedglm: Fitting linear and generalized linear models to large data sets. (2023).
88. Storey, J. D. qvalue: Q-value estimation for false discovery rate control.
89. Bledea, R. *et al.* Functional and topographic effects on DNA methylation in IDH1/2 mutant cancers. *Sci. Rep.* **9**, 16830 (2019).
90. Marcoulides, K. M. & Raykov, T. Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods. *Educ. Psychol. Meas.* **79**, 874–882 (2019).
91. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
92. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. (2012) doi:10.48550/ARXIV.1201.0490.
93. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
94. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
95. Campbell, B. B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **171**, 1042-1056.e10 (2017).

96. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).
97. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
98. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
99. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell* **37**, 639-654.e6 (2020).
100. Sturm, G., Finotello, F. & List, M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. in *Bioinformatics for Cancer Immunotherapy* (ed. Boegel, S.) vol. 2120 223–232 (Springer US, New York, NY, 2020).
101. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. in *Cancer Systems Biology* (ed. Von Stechow, L.) vol. 1711 243–259 (Springer New York, New York, NY, 2018).
102. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
103. Brownlee, J. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. (Machine Learning Mastery, 2020).
104. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).
105. Liao, S. *et al.* The molecular landscape of premenopausal breast cancer. *Breast Cancer Res.* **17**, 104 (2015).
106. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
107. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
108. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
109. Mosteller, F. Questions and Answers. *Am. Stat.* **2**, 30–31 (1948).
110. Dempster, J. M. *et al.* Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol.* **22**, 343 (2021).
111. Kassambara, A. survminer: Drawing Survival Curves using 'ggplot2'.
112. Hoxhaj, G. & Manning, B. D. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. *Nat. Rev. Cancer* **20**, 74–88 (2020).
113. Huang, L., Guo, Z., Wang, F. & Fu, L. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct. Target. Ther.* **6**, 386 (2021).
114. Zhou, Y. *et al.* Impact of KRAS mutation on the tumor microenvironment in colorectal cancer. *Int. J. Biol. Sci.* **20**, 1947–1964 (2024).
115. Kolch, W., Berta, D. & Rosta, E. Dynamic regulation of RAS and RAS signaling. *Biochem. J.* **480**, 1–23 (2023).

FIGURES



Graphical abstract.

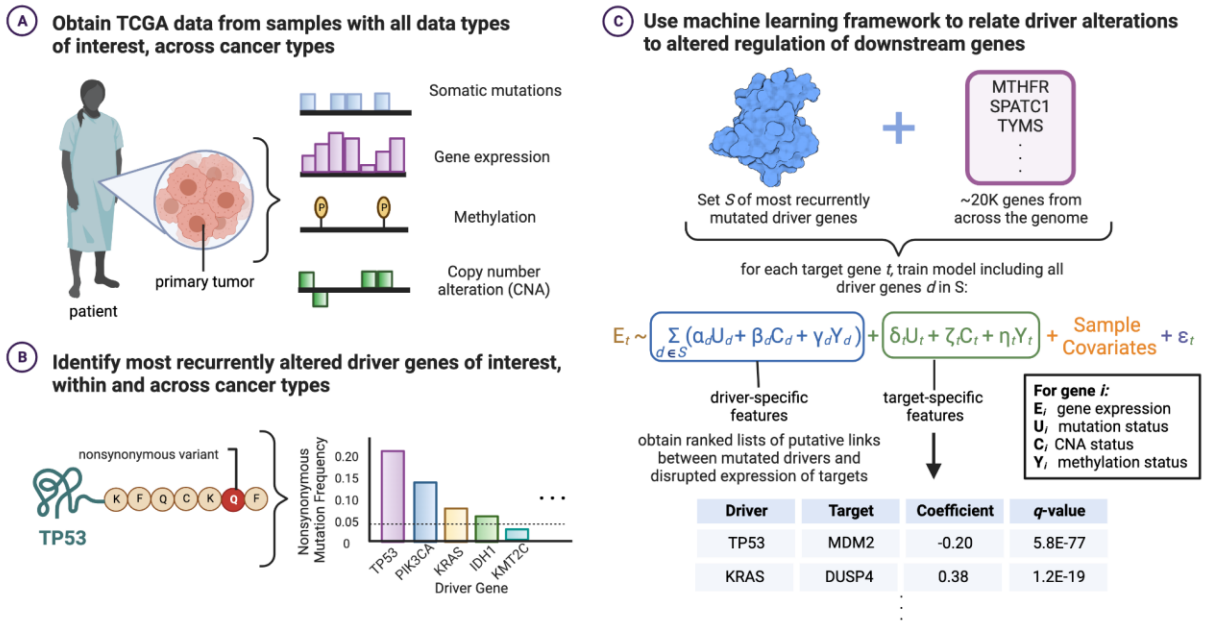


Figure 1. Methodological overview of Dyscovr pipeline. A. Dyscovr framework leverages matched somatic mutation, gene expression, methylation, and copy number alteration (CNA) data from primary patient tumors, such as those from the TCGA. B. Across all cancer types, and within individual cancer types, Dyscovr considers a set *S* of the most recurrently mutated cancer driver genes. Nonsynonymous mutations, including missense, nonsense, nonstop, and splice site mutations, are considered. C. Within each sample population, Dyscovr trains a linear regression model for each putative target gene *t* and tests for a relationship between the expression of *t* and the nonsynonymous mutation status of driver genes, while considering other driver- and target-specific features such as CNA and methylation status along with various clinical features. The expression of gene *t* is correlated to the nonsynonymous mutation status of a driver gene *d* if the fit coefficient α_d is significantly different from 0.

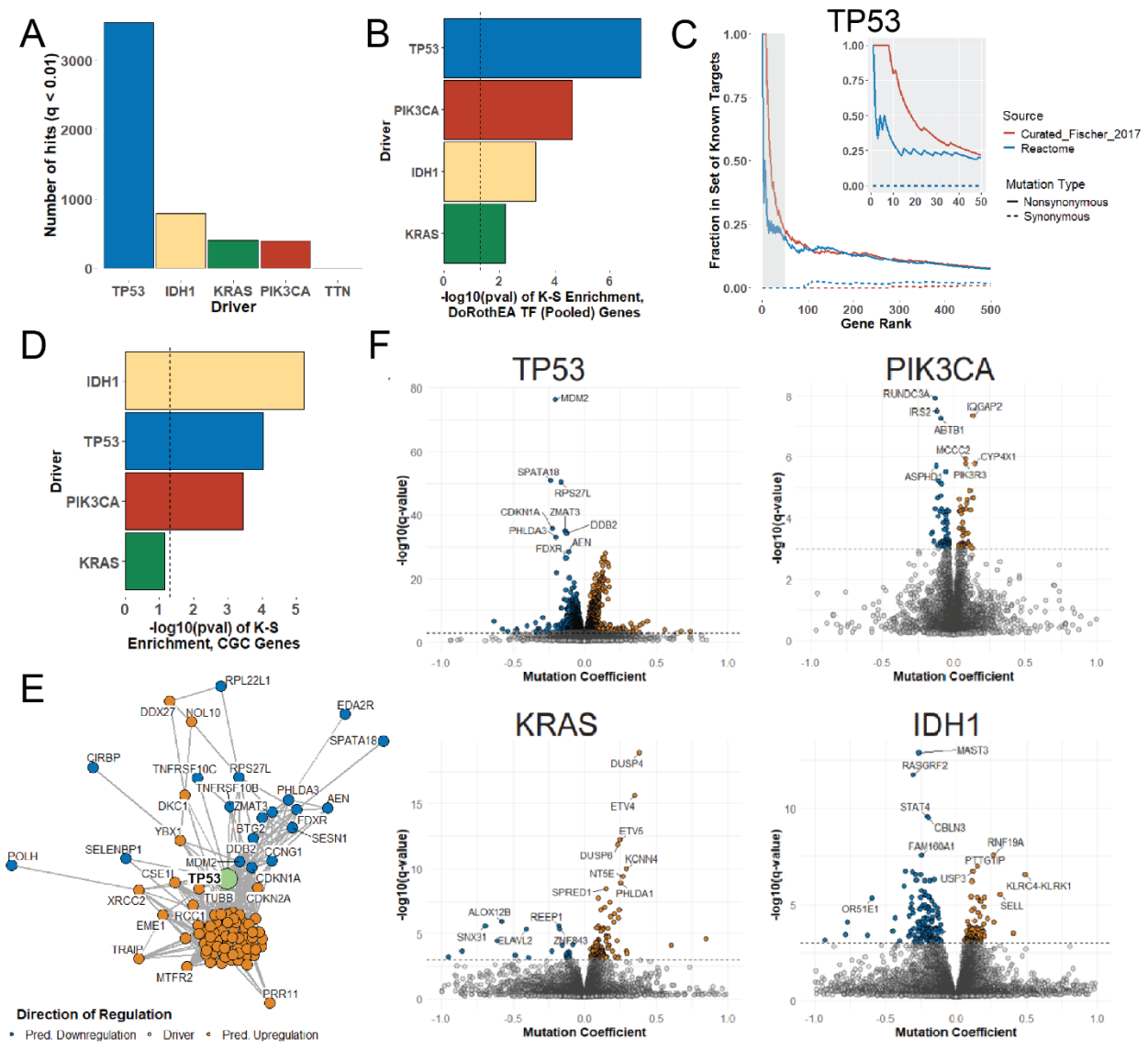


Figure 2. Dyscovr uncovers relationships between mutated pan-cancer drivers and target gene expression. A. The number of $q < 0.01$ hits obtained for *TP53*, *PIK3CA*, *KRAS*, and *IDH1*, the four cancer driver genes considered in pan-cancer modeling. Each hit corresponds to a putative relationship between a nonsynonymous mutation in that driver and an expression change in a given target gene. Dyscovr uncovered hundreds of significant hits at $q < 0.01$ for all recurrently mutated driver genes, while *TTN* (a recurrently mutated gene not annotated to be cancer-relevant[25,32]) has three hits at this threshold. B. Bar chart showing the enrichment of effector TF targets from DoRothEA[26], either *TP53*-specific targets for *TP53* (N = 248) or pooled targets across

multiple, literature-supported effector TFs for *PIK3CA* (N = 1411)*, KRAS* (N = 2246)*, and IDH1* (N = 2519) (Table S2)*.* The magnitude of bars is the *-log10*(*p*-value), with dotted line denoting *p* = 0.05*.* C. The cumulative fraction of curated *TP53* targets (Fischer et al.[27], blue) and *TP53* signaling pathway members (Reactome[31] R-HSA-3700989, red) among an increasing fraction of *q-value* ranked *TP53* hits from a model using nonsynonymous mutations (solid), or with silent mutations (dashed) shown as a control. Inlay (top right) shows the top 50 *TP53* hits. D. Bar chart showing the enrichment (one-sided Kolmogorov-Smirnov (K-S) test) of known cancer genes from the Cancer Gene Census (CGC)[32] in each driver gene's *q*-value ranked targets. The magnitude of bars is the *-log10*(*p*-value), with dotted line denoting *p* = 0.05. E. Overlay of *TP53* model results on the STRING functional network[34], with *TP53* shown in green. Top 100 *TP53* hits at $q < 0.01$ that are connected to *TP53* either directly or by means of another hit with STRING confidence >0.4 are displayed, with color corresponding to direction of predicted regulation (upregulation in orange, downregulation in blue). F. Volcano plots of hits for each driver gene, with a selection of statistically significant hits labeled by name. The *x*-axis is the fit coefficient for the driver mutation status term, thresholded at [-1.5, 1.5] for visual clarity, with predicted significant upregulation in orange and downregulation in blue at $q < 0.01$ (dotted line). The *y*-axis is the *-log10* of the *q*-value produced by Dyscovr for the driver mutation status term, with larger values having greater predicted significance.

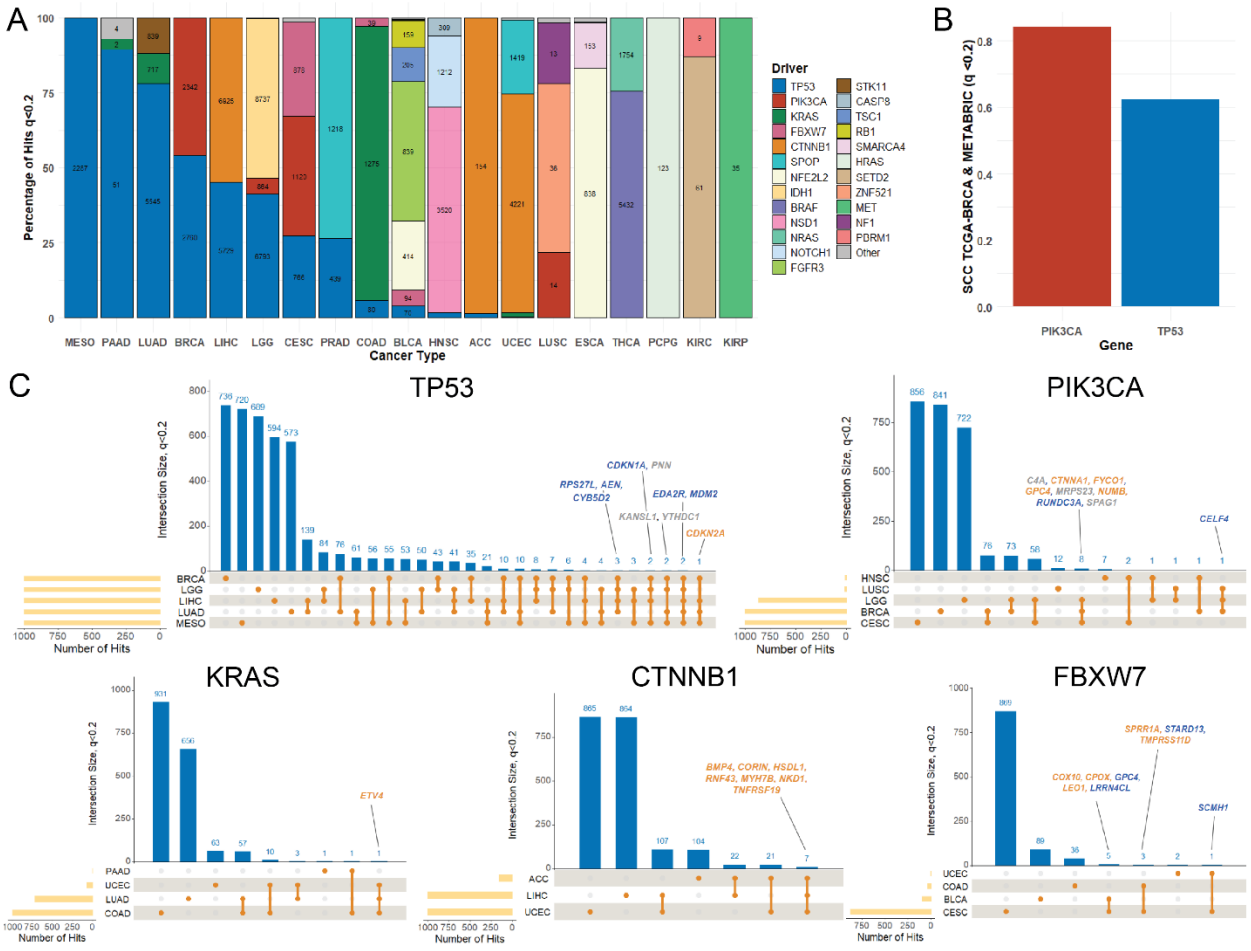


Figure 3. Driver mutation-target gene expression relationships across individual TCGA cancer types. A. In each of 19 TCGA cancer types, Dyscovr was run for every putative target gene, with each model including variables for all driver genes mutated at ≥5% frequency across samples of that cancer type. Segments of the stacked bars represent the percentage of significant hits ($q < 0.2$) in a given cancer type (*x*-axis) that belong to a given driver gene. Segments of the stacked bars are labeled with the absolute number of dysregulated targets for that given driver in the given cancer type at $q < 0.2$. See Table S1 for full-length cancer type names. B. Bar chart of the Spearman's rank correlation coefficient (*y*-axis) between the nonsynonymous mutation status coefficient from Dyscovr of the given driver (*x*-axis, *PIK3CA* in red, *TP53* in blue) of all gene targets that are significant at $q < 0.2$ in two breast cancer datasets, TCGA-BRCA and METABRIC. Pairwise Spearman's rank coefficient values for significant target genes at $q < 0.2$ are 0.84 for

PIK3CA and 0.62 for *TP53*. C. UpSet plots for the five driver genes with the most overall hits, spread across at least 3 TCGA cancer types: *TP53, PIK3CA*, *KRAS, CTNNB1,* and *FBXW7*. Plots show the absolute number of hits at $q < 0.2$, capped at 1000 for visual clarity, in each of the up-to-five cancer types with the largest number of hits for that driver (bottom left, yellow) as well as the hits that do and do not overlap between the various cancer types (blue). Significant hits ($q < 0.2$) for a given driver that are common to the most shown cancer types are labeled and colored by direction of regulation (upregulation in orange, downregulation in blue, and variable direction in gray).

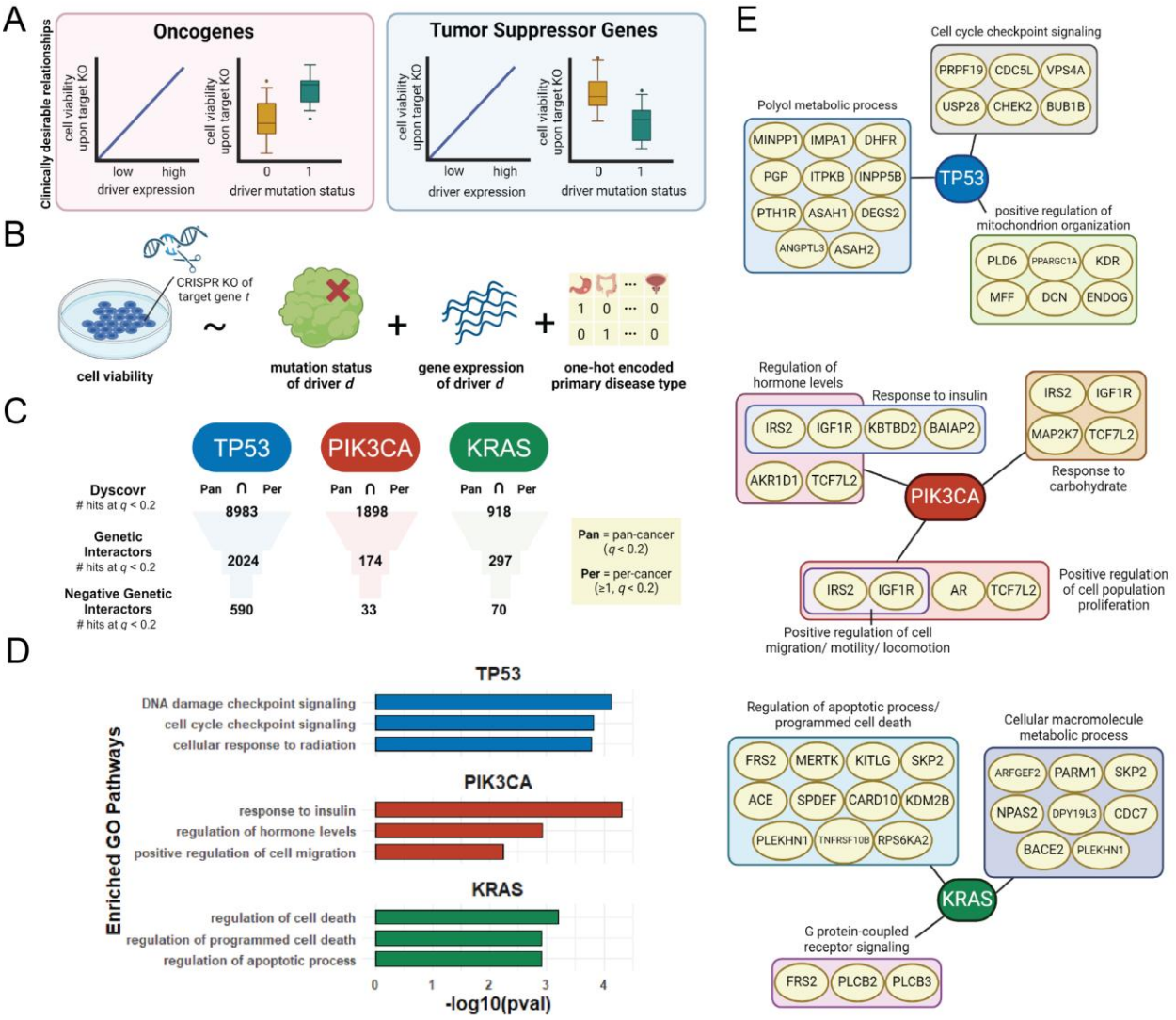


Figure 4. Cell viability analysis using DepMap data reveals Dyscovr targets that are putative synthetic lethals with nonsynonymous mutations in corresponding driver genes.

A. Graphical overview of selecting for negative genetic interactions, which are clinically desirable.

B. Schematic of cell viability analysis regression framework, which relates the nonsynonymous mutation status and expression of a putative driver gene *d* (accounting for primary disease type) to cell viability upon CRISPRi knockdown of putative target gene *t* from Dyscovr (see Methods VIII). To limit our pool of candidate genetic interactors to putative negative genetic interactors, we restrict by the directionality of the driver mutation (U_d) and driver expression (E_d) coefficients. In the case of oncogenes (i.e. *PIK3CA, KRAS*), we limit to cases where both $E_d > 0$ and $U_d > 0$, to

capture situations where application of a mutant oncogene inhibitor in combination with a target gene inhibitor would be predicted to result in synergistically reduced cell viability. Conversely, in the case of tumor suppressor genes (i.e. *TP53*), we limit to cases where both $E_d > 0$ and $U_d < 0$, to capture situations where application of a target gene inhibitor in tumors with mutant copies of the given tumor suppressor would be predicted to result in synergistically reduced cell viability. C. Schematic of using the Cancer Dependency Map (DepMap)[24] data to refine Dyscovr's hits to those that are most likely to be SL with nonsynonymous mutations in their corresponding driver. For pan-cancer driver genes with sufficient cell line data (*TP53*, *PIK3CA*, and *KRAS),* we report the number of significant hits from Dyscovr ($q < 0.2$) that are found both pan-cancer ("Pan") and in at least one individual cancer type ("Per") at top. In the middle, we report the number of significant targets from the cell viability analysis pipeline that were identified as genetic interactors with the associated driver, i.e. cell viability upon knockout of that target was found to be related to driver activity. At the bottom, we report the number of significant targets that we classified as negative genetic interactors with the associated driver (see part A). D. Top three significantly ($q < 0.2$) enriched GO pathways, ranked by -log10(*p*-value) as determined by GSEA (*x*-axis, Methods VII), among negative genetic interaction hits ($q < 0.2$) of *TP53, PIK3CA,* and *KRAS*. E. Visualization of a subset of significant negative genetic interaction hits ($q < 0.2$) for each of *TP53, PIK3CA,* and *KRAS,* annotated with pathways from GO pathway enrichment analysis.

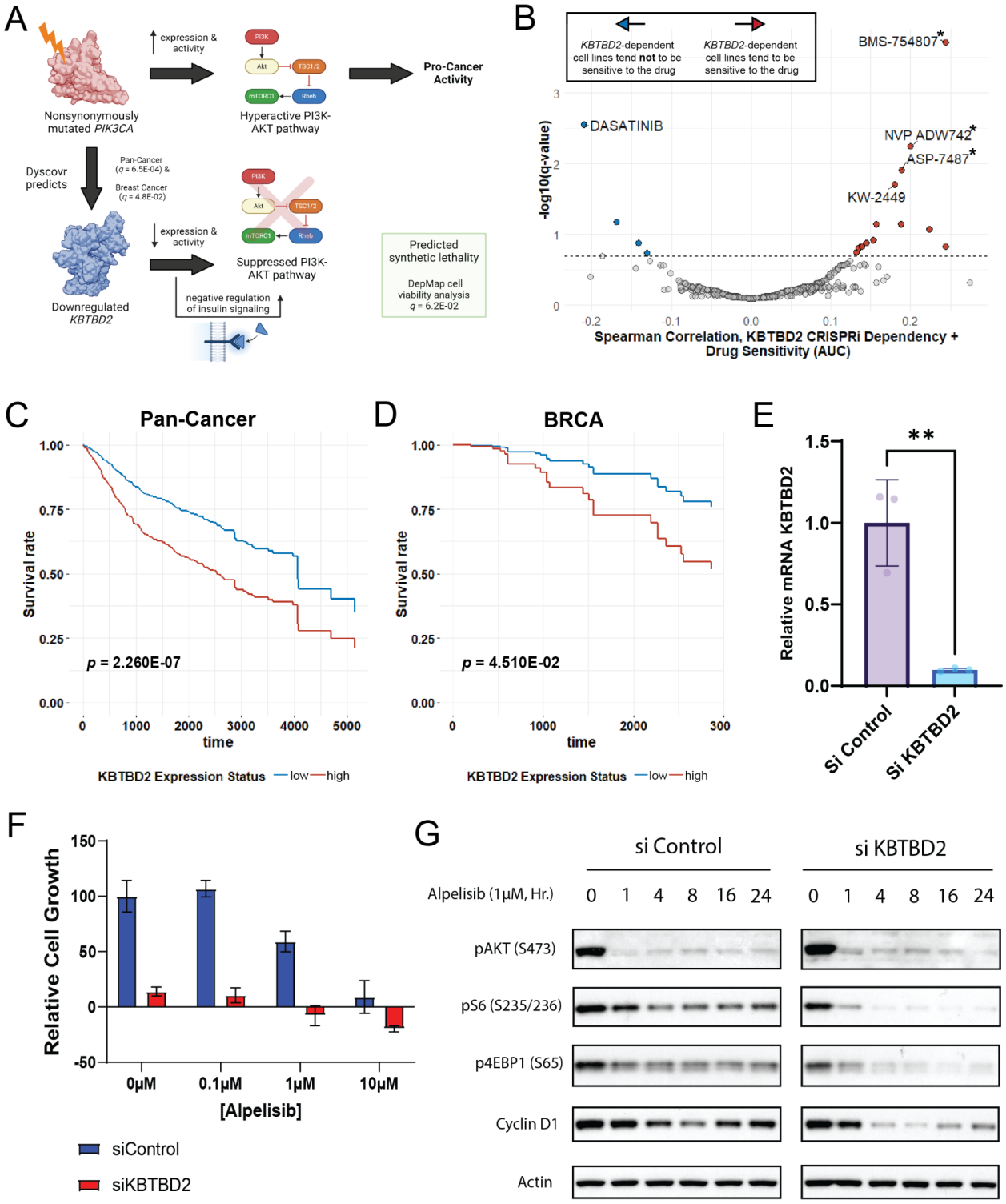
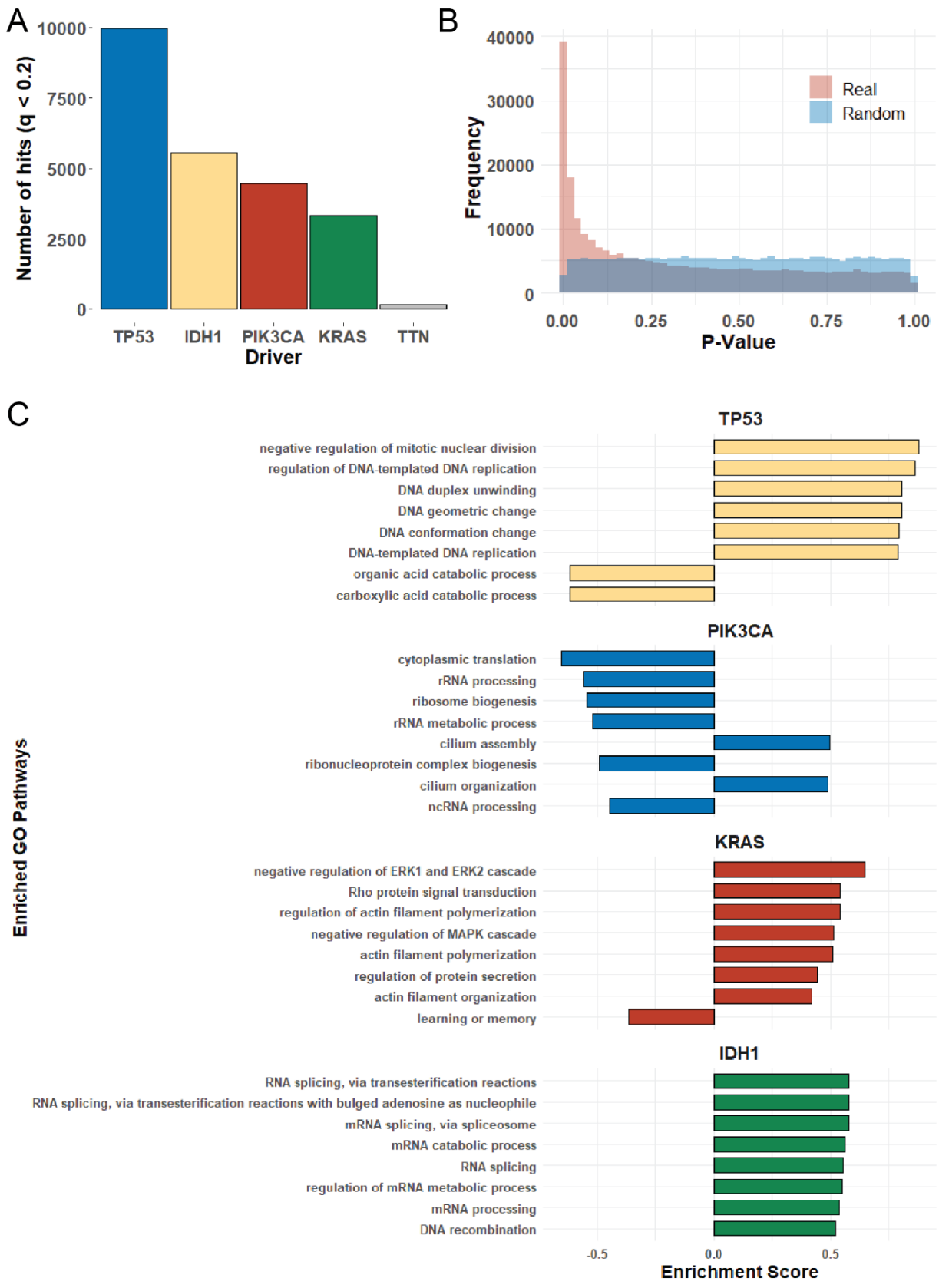


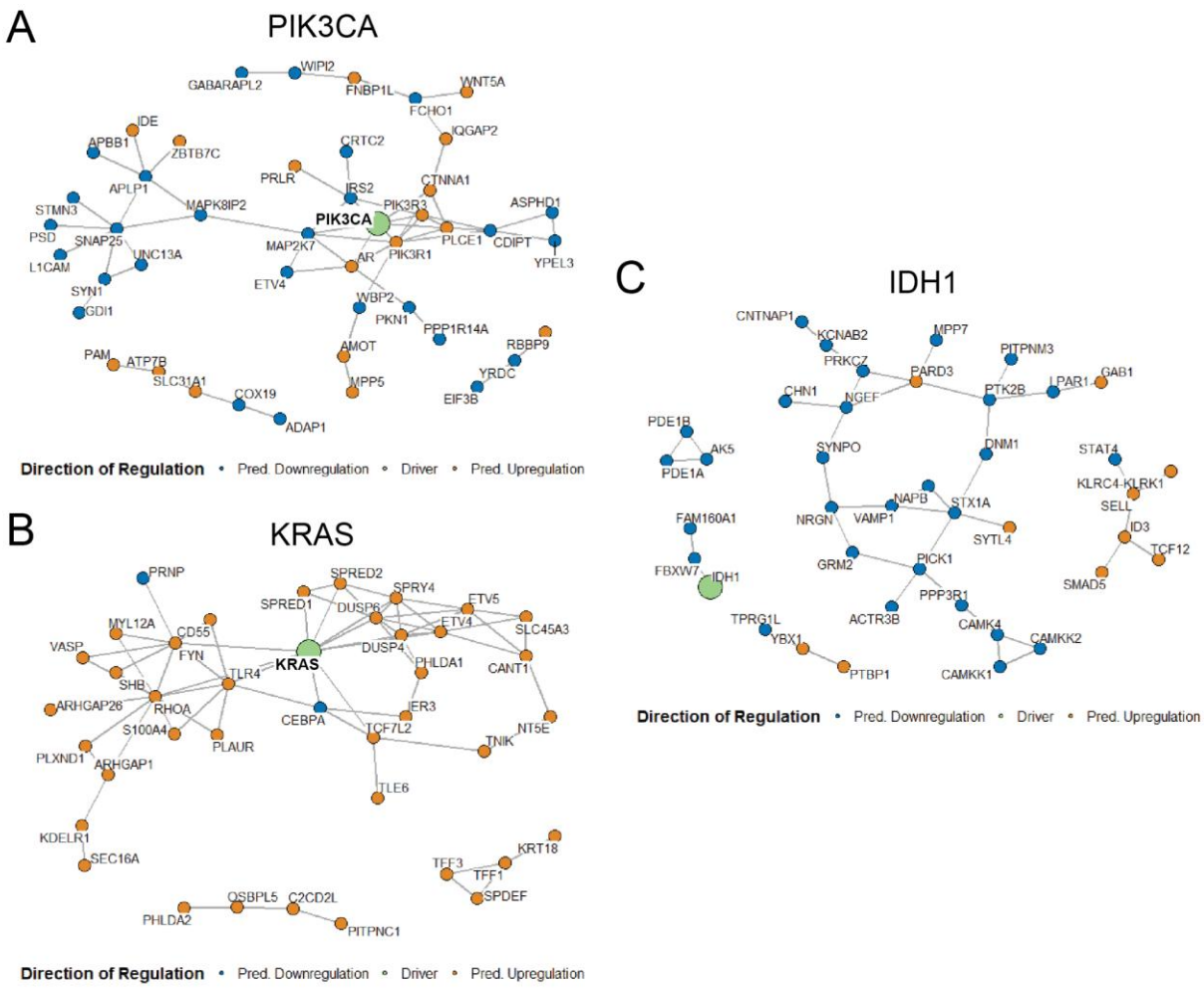
Figure 5. Mutant *PIK3CA* and expression of *KBTBD2* display synthetic lethality. A. Schematic of proposed mechanism-of-action of synthetic lethality between mutated *PIK3CA* and dysregulated expression of *KBTBD2* kinase. Dyscovr identifies a correlation between *PIK3CA*

nonsynonymous mutation status and *KBTBD2* downregulation, both pan-cancer (q = 6.53E-04) and in breast cancer (q = 4.80E-02). This pair is also identified as candidate synthetic lethal via cell viability analysis in DepMap cell lines (q = 6.17E-02, see Methods VIII). *KBTBD2* has been previously identified as being a regulator of insulin signaling in mouse models via its regulation of p85α protein abundance, with low levels of *KBTBD2* corresponding to suppressed PI3K-AKT signaling[77]. This proposed system lends itself to the hypothesis that inhibition of mutant *PIK3CA* would result in recovery of *KBTBD2* expression; therefore, joint inhibition of mutant *PIK3CA* and *KBTBD2* should result in synergistic effects on cell growth. B. The per-drug Spearman's rank correlation between reported drug sensitivity score from DepMap (AUC) and cell viability upon CRISPR knockout of *KBTBD2* (x-axis) against the significance, i.e. -log10(q-value), of that correlation. Correlations were computed for pan-cancer cell lines and each of 4659 drugs from DepMap. Drugs with statistically significant correlations to cell viability upon KBTBD2 knockout (q < 0.2) are colored, with positive correlations in red and negative correlations in blue. Labels with an asterisk (*) indicate drugs that target insulin like growth factor 1 receptor, *IGF1R*, suggesting that those cell lines most sensitive to *KBTBD2* knockout also tend to be most sensitive to inhibitors of insulin signaling. C-D. Survival of pan-cancer (C) or breast cancer (D) TCGA patients significantly differs when stratified by "low" or "high" *KBTBD2* expression (p = 2.26E-07 and hazard ratio of 2.61 pan-cancer, p = 4.51E-02 and hazard ratio of 2.99 in BRCA). Survival rate (y-axis) by time in days (x-axis) was calculated using a Cox proportional hazards model that accounts for the nonsynonymous mutation status of *PIK3CA* and other clinical and molecular confounders (see Methods IX). E. qPCR-based quantification of *KBTBD2* mRNA expression in MCF7 cells treated for 48hr with an siRNA control or siRNA against *KBTBD2*. Data are presented as mean ± SD, and statistical significance is indicated (**p < 0.01). F. Relative cell growth in response to alpelisib treatment in siControl and siKBTBD2 treated cells. MCF7 cells were transfected with siControl or siKBTBD2 for 48hr. followed by alpelisib treatment at indicated doses for an additional 2 days. Cell growth was quantified by SRB staining. Data are presented as mean

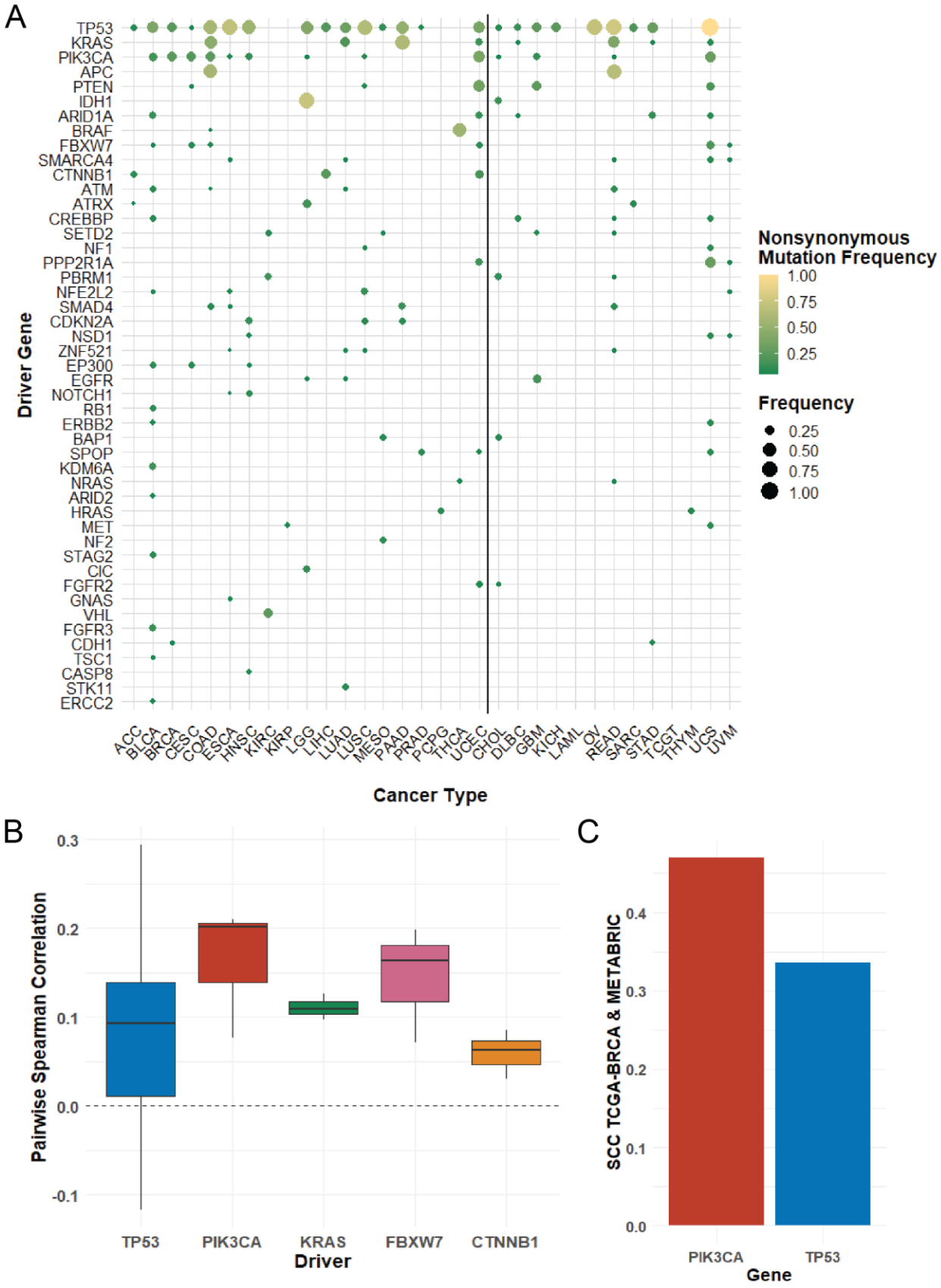
± SD. G. Western blot analysis of PI3K signal transduction following alpelisib treatment and *KBTBD2* knockdown. MCF7 cells were transfected with siControl or siKBTBD2 for 48Hr. followed by treatment with 1 µM alpelisib for time t (hrs).



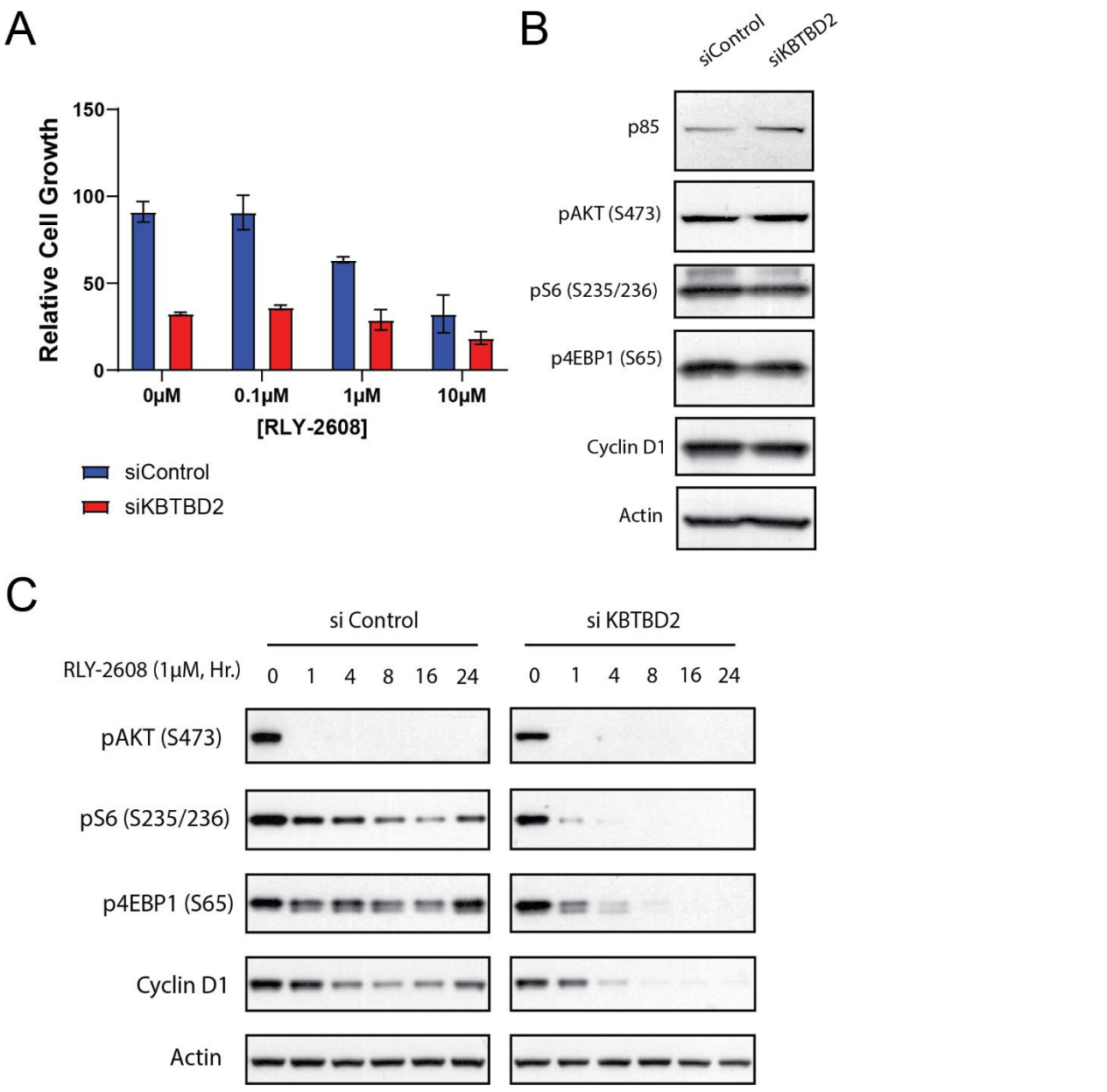
Supplemental Figure 1. Pan-cancer quantitative validations and gene set enrichment analyses. A. The number of q < 0.2 hits obtained for *TP53*, *PIK3CA*, *KRAS*, and *IDH1* pan-cancer. Each hit corresponds to a putative relationship between a mutation in that driver and an expression change in a given target gene. Dyscovr uncovered thousands of significant hits at q < 0.2 for all recurrently mutated driver genes, while *TTN* (a recurrently mutated non-Vogelstein[25] or CGC[32] gene) has 160 hits at this threshold. B. The distribution of p-values produced by Dyscovr for all driver mutation terms (e.g. all $U_d \in S$, where S is the set of pan-cancer driver genes *TP53*, *PIK3CA*, *KRAS*, and *IDH1*) when applied pan-cancer to all putative human gene targets (pink). Overlaid is the distribution of p-values produced by Dyscovr when, for each putative human gene target, expression values (e.g. E_t) are randomized across all pan-cancer samples (blue). C. Bar charts of Gene Ontology (GO)[33] gene set enrichment analysis (GSEA) results for each driver gene. Enriched up- and down-regulated pathways were generated using ReactomePA[107]. Pathways were restricted to those enriched with a q-value < 0.05, then ranked by -log10(q-value), with ties broken by the descending leading-edge percentage. The top eight pathways from this analysis are reported in each bar plot, with the size of the bar corresponding to the enrichment score.



Supplemental Figure 2. Overlay of per-driver, pan-cancer Hits (q < 0.01, Top 100) from Dyscovr on the STRING functional network[34]**.** A-C. Overlay of *PIK3CA, KRAS,* and *IDH1* (respectively) Dyscovr model results on the STRING functional protein-protein interaction network, with respective driver genes shown in green. Top 100 hits at q < 0.01 that are connected to that driver either directly or by means of another hit with STRING confidence >0.4 are displayed, with color corresponding to direction of predicted regulation (upregulation in orange, downregulation in blue). Disconnected components of size 2 or smaller are removed from the visualization.

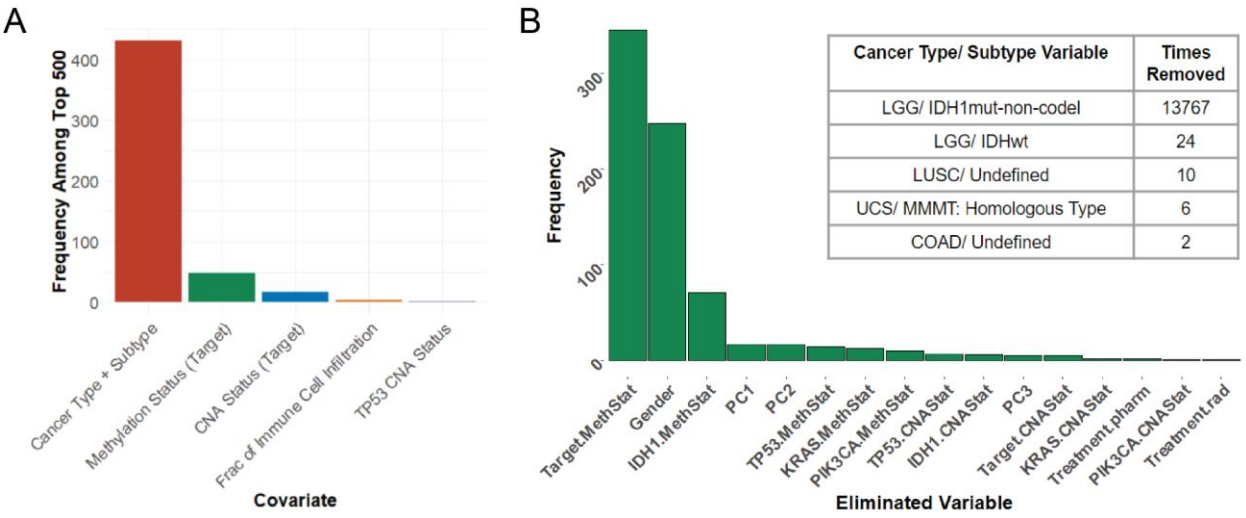


Supplemental Figure 3. Driver mutation-target gene expression relationships across individual TCGA cancer types. A. Cowplot of nonsynonymous mutation frequency for recurrently mutated (≥5%) driver genes, across TCGA cancer types. Each cancer driver gene, as annotated by Vogelstein[25], mutated at ≥5% frequency in at least one of the 19 TCGA cancer types with ≥75 samples matching inclusion criteria (see Methods II) is shown (*y*-axis, left of vertical line). A dot is present if that driver gene is mutated in ≥5% of the samples for the given cancer type (*x*-axis). Both the size and the color of the dot represent the nonsynonymous mutation frequency of that driver gene in that cancer type. Frequencies of these drivers in cancer types with fewer than 75 samples are shown to the right of the vertical black line. B. Pairwise Spearman's rank correlations of estimated mutation coefficients for *TP53, PIK3CA*, *KRAS, CTNNB1,* and *FBXW7*, across all putative target genes, for the set of cancer types in which that driver is frequently mutated (≥5% of samples) and has at least 15 significant hits for the given driver ($q < 0.2$). The average Spearman's rank correlation for all drivers is greater than 0 (dashed line), indicating that driver mutational effects on genome-wide transcription share similarities between cancer types. C. Bar charts of the Spearman's rank correlation coefficient between the nonsynonymous mutation status coefficient of the given driver (*PIK3CA* in red, *TP53* in blue) across all target genes in TCGA-BRCA and METABRIC. Pairwise Spearman's rank correlation coefficient values are 0.47 for *PIK3CA* and 0.34 for *TP53*, both with *p*-values < 1E-200.



Supplemental Figure 4. Characterization of *KBTBD2* in the context of mutant PI3K alpha inhibition. A. Relative cell growth in response to RLY-2608 treatment in siControl and siKBTBD2 treated cells. MCF7 cells were transfected with siControl or siKBTBD2 for 48hr. followed by RLY-2608 treatment at indicated doses for an additional 2 days. Cell growth was quantified by SRB staining. Data are presented as mean ± SD. B. Knockdown of *KBTBD2* and western blot analysis of PI3K pathway components. MCF7 were transfected with control or

siRNA targeting KBTBD2 for 48Hr. C. Western blot analysis of PI3K signal transduction following RLY-2608 treatment and *KBTBD2* knockdown. MCF7 cells were transfected with siControl or siKBTBD2 for 48Hr. followed by treatment with 1 µM RLY-2608 for time *t* (hrs).



Cancer Type/ Subtype Variable	Times Removed
LGG/ IDH1mut-non-codel	13767
LGG/ IDHwt	24
LUSC/ Undefined	10
UCS/ MMMT: Homologous Type	6
COAD/ Undefined	2

Supplemental Figure 5. Pan-cancer top significant Dyscovr covariates and covariates removed due to multicollinearity. A. When Dyscovr is applied pan-cancer across all putative gene targets, we aggregate all terms from these models together and rank them by *q*-value. Among the top 500 most significant terms from this ranking, we show the five most represented term categories (when counting bucketed variables together, see Methods IV.F) on the *x*-axis, with the frequency of each among the top 500 on the *y*-axis. Cancer type and subtype terms are most represented among the top 500 most significant terms, followed by the methylation status of the target gene, the CNA status of the target gene, fraction of immune cell infiltration, and *TP53* CNA status. B. Barplot (left) displays the absolute number of times a given non-cancer type/ subtype covariate was removed from the regression framework due to a multicollinearity violation when applied pan-cancer to all genes (16,447 total regressions performed). All covariates in barplot were tested for collinearity using the VIF metric and were removed if they had a VIF value that exceeded 5. 'MethStat' is an abbreviation for methylation status, while 'CNAStat' is an

abbreviation for CNA status. PC1-3 refer to the germline principal components. The table (right) displays the absolute number of times a given cancer type/ subtype covariate was removed from the regression framework when applied pan-cancer to all genes. All cancer type/ subtype covariates in the table were tested for collinearity using the Spearman correlation coefficient and were removed if they had a Spearman correlation coefficient >0.7 with a significance p<1E-05 (see Methods III).

TABLES

Driver Gene	10 Most Significant Associated CGC[5] Genes
TP53	*MDM2*, *CDKN1A*, *DDB2*, *BUB1B*, *CDKN2A*, *FANCD2*, *STIL*, *CHEK2*, *XPC*, *KNSTRN*
PIK3CA	*PIK3R1*, *ETV4*, *AR*, *TET2*, *BAZ1A*, *IDH2*, *SBDS*, *VTI1A*, *SOX2*, *NOTCH2*
KRAS	*ETV4*, *ETV5*, *TCF7L2*, *CANT1*, *RHOA*, *SLC45A3*, *CREB3L1*, *ARHGAP26*, *CEBPA*, *PPP2R1A*
IDH1	*TCF12*, *FBXW7*, *PRKCB*, *ID3*, *GAS7*, *MAML2*, *CHIC2*, *PRDM2*, *HIF1A*, *TET2*

Table 1. Per-driver, pan-cancer enrichment in cancer genes from the Cancer Gene Census (CGC)[32]. Ten most statistically significant (ranked by q-value) CGC target genes for each driver gene are shown, with those genes that are upregulated in relation to driver gene mutation in orange and those that are downregulated in relation to driver gene mutation in blue.

Cancer Type Abbrev.	Cancer Type Name	Vogelstein Driver Genes at ≥5% Frequency	Sample Size
ACC	Adrenocortical carcinoma	*CTNNB1, TP53*	76
BLCA	Bladder Urothelial Carcinoma	*ARID1A, ARID2, ATM, CREBBP, EP300, ERBB2, ERCC2, FBXW7, FGFR3, KDM6A, NFE2L2, PIK3CA, RB1, STAG2, TP53, TSC1*	349
BRCA	Breast invasive carcinoma	*CDH1, PIK3CA, TP53*	732
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	*EP300, FBXW7, PIK3CA, PTEN, TP53*	276
CHOL	Cholangiocarcinoma	*NA*	33
COAD	Colon adenocarcinoma	*APC, FBXW7, KRAS, PIK3CA, SMAD4, TP53*	236
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	*CARD11, MYD88, P2RY8, PIM1*	36
ESCA	Esophageal carcinoma	*GNAS, NFE2L2, PIK3CA, SMAD4, SMARCA4, TP53*	158
GBM	Glioblastoma multiforme	*EGFR, PIK3CA, PTEN, TP53*	58
HNSC	Head and Neck squamous cell carcinoma	*CASP8, CDKN2A, EP300, NOTCH1, NSD1, PIK3CA, TP53*	465
KICH	Kidney Chromophobe	*TP53*	65
KIRC	Kidney renal clear cell carcinoma	*PBRM1, SETD2, VHL*	141
KIRP	Kidney renal papillary cell carcinoma	*MET*	224
LAML	Acute Myeloid Leukemia	*NA*	6
LGG	Brain Lower Grade Glioma	*ATRX, CIC, EGFR, IDH1, PIK3CA, TP53*	518
LIHC	Liver hepatocellular carcinoma	*CTNNB1, TP53*	359

LUAD	Lung adenocarcinoma	*KRAS, STK11, TP53*	361
LUSC	Lung squamous cell carcinoma	*CDKN2A, NF1, NFE2L2, PIK3CA, PTEN, TP53, ZNF521*	300
MESO	Mesothelioma	*BAP1, NF2, TP53*	81
OV	Ovarian serous cystadenocarcinoma	*NA*	7
PAAD	Pancreatic adenocarcinoma	*CDKN2A, KRAS, SMAD4, TP53*	170
PCPG	Pheochromocytoma and Paraganglioma	*HRAS*	167
PRAD	Prostate adenocarcinoma	*SPOP, TP53*	419
READ	Rectum adenocarcinoma	*APC, KRAS, TP53*	18
SARC	Sarcoma	*TP53*	58
STAD	Stomach adenocarcinoma	*ARID1A, TP53*	50
TGCT	Testicular Germ Cell Tumors	*KIT*	29
THCA	Thyroid carcinoma	*BRAF, NRAS*	396
THYM	Thymoma	*NA*	23
UCEC	Uterine Corpus Endometrial Carcinoma	*ARID1A, CTNNB1, FBXW7, FGFR2, KRAS, PIK3CA, PPP2R1A, PTEN, SPOP, TP53*	295
UCS	Uterine Carcinosarcoma	*TP53*	10
UVM	Uveal Melanoma	*GNA11, GNAQ*	17
PC	Pan-Cancer	*TP53, PIK3CA, KRAS, IDH1*	6135

Supplemental Table 1. The Vogelstein et al. cancer driver genes[25] that are mutated in at least 5% of samples in each TCGA cancer type. The driver genes mutated in at least 5% of pan-cancer samples are shown in the bottommost row.

Driver Gene	Curated Effector TFs	# of Targets	KS One-Sided Enrichment Test, *q*-value
PIK3CA[112]	*FOXO1/3/4/6*	907	**4.05E-03**
	MYC	386	9.18E-02
	HIF1A	148	5.22E-02
	SREBF1/2	31	**4.05E-03**
	ATF4	16	7.67E-01
	NFR2/NFE2L2	11	8.19E-01
KRAS[113–115]	*CREB1/3/5*	1779	5.22E-02
	MYC	386	7.67E-01
	ETS1/2	154	5.22E-02
	FOXO1	43	6.25E-02
	ELK1	31	5.22E-02
	ETV1	29	**4.36E-02**
	AP1 (absent from DoRothEA)	N/A	N/A
IDH1[89]	*SOX8*	802	1.35E-01
	HEY1	422	3.77E-01
	JUN/JUNB/JUND	176	2.02E-01
	ATF3	50	**4.05E-03**
	NR2F2	18	8.19E-01
	MYCN	13	**4.36E-02**

Supplemental Table 2. Pan-cancer enrichment in transcriptional targets for effector TFs of *PIK3CA, KRAS*, and *IDH1*. For each pan-cancer driver gene that is not itself a TF, we curated TFs shown in the literature to be effectors of that driver and tested for enrichment of each of those

TFs' targets from DoRothEA[26]. The *q*-value from a multiple hypothesis corrected one-sided KS test for enrichment is shown for each effector TF, with *q*-values < 0.05 displayed in bold.

***TP53* Target List**	**# of Genes in Set**	**One-sided Kolmogorov-Smirnov (K-S) Test, *p*-value**
Curated Targets and Databases		
Curated, Fischer et al.[27]	116	7.48E-14
DoRothEA[26]	248	7.68E-08
TRRUST[28]	195	2.23E-10
hTFtarget[29]	72	2.45E-02
Gene Pathways		
KEGG Pathway[30] "P53 Signaling Pathway", hsa04115	69	2.10E-07
Reactome Pathway[31] "Transcriptional Regulation by TP53", R-HSA-3700989	363	3.70E-08
Reactome Pathway[31] "TP53 Regulates Metabolic Genes", R-CFA-5628897	88	2.57E-03
Known Cancer Genes		
CGC[32] Cancer Genes	714	9.42E-05
Vogelstein[25] Cancer Genes	379	1.43E-03

Supplemental Table 3. *TP53* pan-cancer enrichments in known or inferred targets from curated datasets, TF-target databases, and driver-specific genetic pathways. Number of genes in each set and enrichments from a one-sided Kolmogorov-Smirnov (K-S) test are shown.

Data Category	Data Type	Experimental Strategy	Workflow Type	Data Format	Platform
Simple nucleotide variation	Aggregated somatic mutation	WXS	MuSE Variant Aggregation and Masking	maf	N/A
Copy number variation	Gene level copy number	Genotyping array	ASCAT2	txt	affymetrix snp 6.0
Transcriptome profiling	Gene expression quantification	RNA-seq	HTSeq - Counts	txt	N/A
DNA methylation	Methylation Beta value	Methylation Array	Liftover	txt	illumina human methylation 450
Biospecimen	Biospecimen supplement	N/A	N/A	bcr xml	N/A
Clinical	Clinical supplement	N/A	N/A	tsv	N/A

Supplemental Table 7. Parameters for GDC Data Portal file downloads. Each column represents a facet of the data that can be selected on the GDC Data Portal website when downloading files. For each Data Category (column 1), specific data facet selections used in these analyses are provided for replicability.

Cancer Type Abbrev.	Cancer Type Name	Column name in TCGAbiolinks clinical supplement	Subtypes
ACC	Adrenocortical carcinoma	COC	COC1, COC2, COC3
BLCA	Bladder Urothelial Carcinoma	mRNA cluster	Luminal_infiltrated, Luminal_papillary, Luminal, Basal_squamous, Neuronal, ND
BRCA	Breast invasive carcinoma	BRCA_Subtype_PAM50	LumA, LumB, Her2, Basal, Normal
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	SAMP:CIMP_call	CIMP-low, CIMP-intermediate, CIMP-high
COAD	Colon adenocarcinoma	MSI_status	MSI-H, MSI-L, MSS, Not Evaluable
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	N/A	N/A
ESCA	Esophageal carcinoma	MSI status	MSI-H, MSI-L, MSS, NA
GBM	Glioblastoma multiforme	Original.Subtype	Classical, G-CIMP, IDHmut-codel, IDHmut-non-codel, IDHwt, Mesenchymal, Neural, Proneural
HNSC	Head and Neck squamous cell carcinoma	RNA	Atypical, Basal, Classical, Mesenchymal
KICH	Kidney Chromophobe	Histological.Subtype	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma	N/A	N/A
KIRP	Kidney renal papillary cell carcinoma	tumor_type.KIRP.path.	Type 1 Papillary RCC, Type 2 Papillary RCC, Unclassified Papillary RCC
LGG	Brain Lower Grade Glioma	Original.Subtype	Classical, G-CIMP, IDHmut-codel, IDHmut-non-codel, IDHwt, Mesenchymal, Neural, Proneural
LIHC	Liver hepatocellular carcinoma	iCluster clusters (k=3, Ronglai Shen)	iCluster:1, iCluster:2, iCluster:3
LUAD	Lung adenocarcinoma	iCluster.Group	1,2,3,4,5,6
LUSC	Lung squamous cell carcinoma	Expression.Subtype	basal, classical, primitive, secretory

MESO	Mesothelioma	N/A	N/A
PAAD	Pancreatic adenocarcinoma	Histological type by RHH	Ductal adenocarcinoma, Adenosquamous, Other, Colloid (mucinous noncystic)
PCPG	Pheochromocytoma and Paraganglioma	mRNA Subtype Clusters	Kinase signaling, Wnt-altered, Pseudohypoxia, Cortical admixture, NA
PRAD	Prostate adenocarcinoma	Subtype	1-ERG, 2-ETV1, 3-ETV1, 4-FLI1, 5-SPOP, 6-FOXA1, 7-IDH1 8-other
READ	Rectum adenocarcinoma	MSI_subtypes	N/A
THCA	Thyroid carcinoma	mRNA_Cluster_number	1,2,3,4,5,NA
THYM	Thymoma	N/A	N/A
UCEC	Uterine Corpus Endometrial Carcinoma	msi	Indeterminant, MSI-H, MSI-L, MSS
UCS	Uterine Carcinosarcoma	histologic subtype	Uterine Carcinosarcoma/ MMMT: Heterologous Type, Uterine Carcinosarcoma/ MMMT: Homologous Type, Uterine Carcinosarcoma/ Malignant Mixed Mullerian Tumor (MMMT): NOS
UVM	Uveal Melanoma	mRNA Cluster No.	1,2,3,4

Supplemental Table 8. Per-cancer TCGA subtype information. For each TCGA cancer type, displays the TCGAbiolinks clinical supplement column name that contains the cancer subtype information used in Dyscovr (column 2) and the range of subtype values found in that column (column 3). In all cases, molecular subtypes were used, if available. If not, histological subtypes were preferentially used, followed by expression-based subtypes, such as from mRNA clusters.