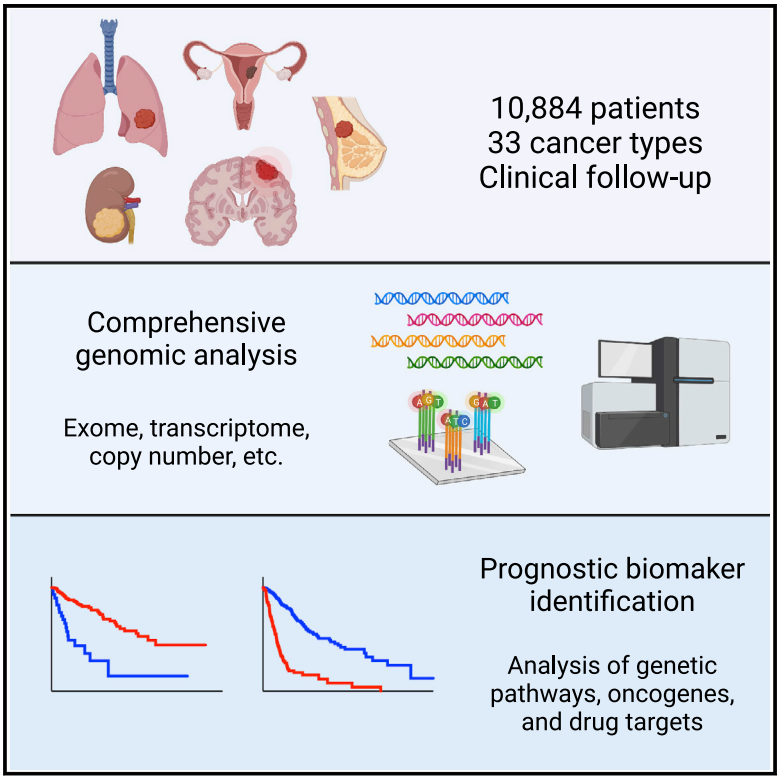
Cell Reports

Article

Genome-wide identification and analysis of prognostic features in human cancers

Graphical abstract



Highlights

- A pan-cancer analysis identifies genomic biomarkers linked with patient outcome
- Copy number alterations and transcripts are more prognostic than mutations
- Oncogenes and successful cancer drug targets are rarely prognostic
- A web portal is created to facilitate analysis of outcome-linked biomarkers

Authors

Joan C. Smith, Jason M. Sheltzer

Correspondence

jason.sheltzer@yale.edu

In brief

Smith and Sheltzer identify genomic alterations linked with outcome across 32 cancer types. They identify thousands of prognostic biomarkers and reveal a prominent association between gene copy number alterations and patient death. This comprehensive resource can be mined to uncover biomarkers that identify the patients most at risk for disease progression.

Smith & Sheltzer, 2022, Cell Reports *38*, 110569
March 29, 2022
https://doi.org/10.1016/j.celrep.2022.110569

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Genome-wide identification and analysis of prognostic features in human cancers

Joan C. Smith[1,2] and Jason M. Sheltzer[1,3,*]
[1]Yale University School of Medicine, New Haven, CT 06511, USA
[2]Google, Inc., New York, NY 10011, USA
[3]Lead contact
*Correspondence: jason.sheltzer@yale.edu
https://doi.org/10.1016/j.celrep.2022.110569

SUMMARY

Clinical decisions in cancer rely on precisely assessing patient risk. To improve our ability to identify the most aggressive malignancies, we constructed genome-wide survival models using gene expression, copy number, methylation, and mutation data from 10,884 patients. We identified more than 100,000 significant prognostic biomarkers and demonstrate that these genomic features can predict patient outcomes in clinically ambiguous situations. While adverse biomarkers are commonly believed to represent cancer driver genes and promising therapeutic targets, we show that cancer features associated with shorter survival times are not enriched for either oncogenes or for successful drug targets. Instead, the strongest adverse biomarkers represent widely expressed cell-cycle and housekeeping genes, and, correspondingly, nearly all therapies directed against these features have failed in clinical trials. In total, our analysis establishes a rich resource for prognostic biomarker analysis and clarifies the use of patient survival data in preclinical cancer research and therapeutic development.

INTRODUCTION

The ability to accurately discriminate between aggressive and indolent cancers underlies the prediction of patient risk and can guide crucial treatment decisions (Ludwig and Weinstein, 2005). For benign cancers, watchful waiting and/or surgical resection can be appropriate, while invasive cancers may require multimodal treatment with cytotoxic therapies that themselves cause substantial morbidity. Both cancer undertreatment and cancer overtreatment have been identified as significant sources of patient mortality, underscoring the urgent need to improve our ability to precisely identify patients with the most aggressive malignancies (Bouchardy et al., 2003, 2007; Dale, 2003; Esserman et al., 2013; Jegerlehner et al., 2017; Loeb et al., 2014).

Current risk prediction largely relies on histopathological and radiological assessment of disease status (Ludwig and Weinstein, 2005). The presence of features such as lymph node metastases and cellular dedifferentiation have been identified as strong predictors of patient outcome and are used to determine cancer stage and grade (Connolly et al., 2003). However, these pathological markers require subjective judgements and can exhibit low levels of interobserver agreement (Elmore et al., 2015; Evans et al., 2008; Gilks et al., 2013; Griffiths et al., 2006; Lang et al., 2005; Ozkan et al., 2016). Moreover, even perfect tumor staging cannot unambiguously predict a patient's subsequent clinical course (Bijker et al., 2013; Young, 2003; Zaniboni and Labianca, 2004).

The widespread adoption of gene expression analysis, DNA sequencing, copy number determination, and other genomic technologies in clinical settings has raised the exciting possibility that molecular markers could be developed to improve risk assessment (Berger and Mardis, 2018). For instance, a 21-gene RT-PCR panel called Oncotype DX has been demonstrated to accurately predict the likelihood of disease recurrence in ER+ breast cancer, and assigning chemotherapy to patients identified as high-risk based on Oncotype DX decreases the frequency of disease recurrence (Sparano et al., 2018). Similar gene panels for colon cancer, prostate cancer, and several other cancer types are in development (Goossens et al., 2015).

Thus far, efforts to discover prognostic biomarkers have largely sought to identify gene expression changes associated with clinical outcome (Anaya et al., 2016; Gentles et al., 2015; Tang et al., 2019; Uhlen et al., 2017). These studies have demonstrated that genes associated with cell-cycle progression correlate with aggressive disease in multiple cancer types (Cuzick et al., 2011; Dancik and Theodorescu, 2015; Gentles et al., 2015; Mosley and Keri, 2008; Venet et al., 2011). Moreover, a series of recent reports have highlighted that copy number alterations (CNAs) also convey significant prognostic information, with increasing CNA burden generally associated with disease recurrence and metastatic dissemination (Hieronymus et al., 2018; Lukow and Sheltzer, 2022; Lukow et al., 2021; Smith and Sheltzer, 2018; Stopsack et al., 2019; Vasudevan et al., 2020, 2021). However, existing studies suffer from several limitations: (1) published analyses largely focus on single cancers and/or genomic data types (e.g., gene expression or DNA methylation or genetic mutations). A comprehensive comparison of prognostic biomarkers across both cancer types

and genomic platforms has not been reported. (2) Biomarker research is potentially affected by the "file drawer" problem, in which certain results (like the discovery of a new biomarker) are more likely to be published, while negative results may end up in a file drawer rather than in the academic literature (Andre et al., 2011; Rosenthal, 1979; Shields, 2000). Unbiased genome-wide biomarker studies can potentially counteract this publication bias and provide an accurate depiction of the prognostic landscape for a cancer of interest. (3) The Cancer Genome Atlas (TCGA), a project that collected genomic and clinical information from 33 cancer types (The Cancer Genome Atlas Research Network et al., 2013), has provided a rich resource for biomarker discovery. However, existing analyses (Anaya, 2016; Gentles et al., 2015; Tang et al., 2017)—including our own previous work (Smith and Sheltzer, 2018)—have relied on preliminary survival data that were published on a cohort-by-cohort basis. A final set of updated and harmonized TCGA survival data has recently been published (Liu et al., 2018), and these existing resources have not been revised.

Beyond the potential clinical relevance of prognostic biomarkers, molecular survival analysis has also become a staple of preclinical cancer research (Chopra and Raynaud, 2020; Kaelin, 2017). With the increasing availability of genome-wide data from patient cohorts such as TCGA (Liu et al., 2018) and MSK-IMPACT (Zehir et al., 2017), it has become straightforward to assess whether a gene of interest is associated with patient outcome. These analyses typically seek to leverage survival data as clinical validation of the importance of a gene of interest in cancer biology. If the overexpression or mutation of some gene is associated with metastasis and patient death, then this is sometimes presented as evidence that that gene is a driver of cancer progression. Alternately, if the overexpression of a gene is associated with favorable prognosis, then it may be assumed that this gene has tumor-suppressive properties. Similar reasoning can underlie the prioritization of targets for therapeutic development: genes that are associated with poor prognosis are presumed to make the best drug targets due to their conjectured role as cancer drivers, while genes associated with favorable prognosis may be disregarded as non-essential for cancer progression (Kaelin, 2017).

To our knowledge, the assumptions underlying these inferences have never been directly tested. While it seems intuitive that the presence of a genetic alteration that drives cancer progression would be associated with worse outcomes, it is not currently known whether real-world data supports this link. Moreover, the prognostic correlations of successful and unsuccessful cancer drug targets have not been investigated. To gain insight into the molecular differences between aggressive and indolent human cancers, and to critically evaluate the use of prognostic data in preclinical research, we performed unbiased survival analysis from all cancer patients and all genomic data platforms included in TCGA.

RESULTS

A comprehensive analysis of TCGA patient survival data

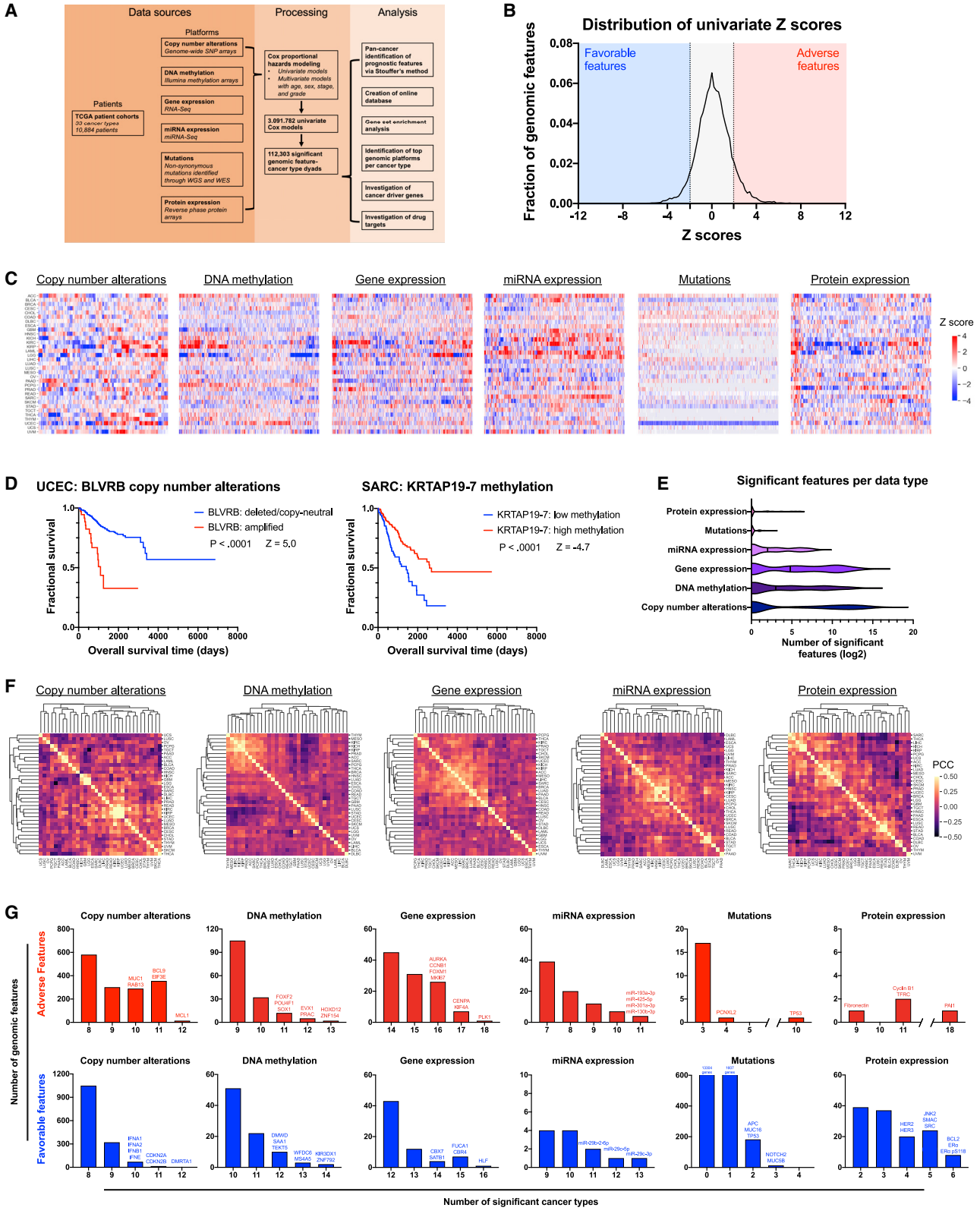
To identify the genomic features that correlate with cancer patient prognosis, we conducted a comprehensive analysis of outcome data for the 33 cancer types profiled by TCGA. Clinical endpoints were selected based on the updated data and recommendations provided in Liu et al. (2018): overall survival was used as an endpoint in 24 cancer types, while progression-free intervals were used in 9 cancer types for which few deaths were observed during the study period (Table 1). For every patient cohort, we extracted information on six different features that were measured in individual tumors: point mutations, CNAs, gene expression, microRNA expression, DNA methylation, and protein expression. For the mutational analysis, we considered only non-synonymous mutations, and in each patient cohort we excluded genes that were mutated in <2% of patients in that cohort (see STAR Methods). We then generated Cox proportional hazard models to assess the relationship between patient outcome and each individual gene for every tumor feature.

To verify the overall fidelity of the clinical and genomic data, we conducted several control analyses. First, we compared survival curves for each of the 33 cancer types that comprise TCGA with cancer survivorship data reported by NCI-SEER. Five-year survival frequencies of individual cancer types were strongly correlated between TCGA and NCI-SEER (Figure S1B; $R = 0.83$, $p < 0.0001$), suggesting that these patient cohorts are broadly representative of the general population. Secondly, we confirmed that patient age, tumor stage, and tumor grade all exhibited a cancer type-independent association with shorter survival times, consistent with the well-established relationship between these clinical variables and poor outcomes (Figure S1C–S1E) (Colonna et al., 2010, Hagberg et al., 2017). Third, we inferred chromosomal sex based on the expression of an X chromosome marker (XIST) and a Y chromosome marker (RPS4Y1), and we found that the extrapolated values exhibited >99% concordance with the annotated gender of each patient (Figure S1F) (Gentles et al., 2015). Similarly, the methylation patterns of two X chromosome genes were also >99% concordant with gender (Figure S1G). Finally, we calculated the mutation frequencies of 106 verified oncogenes and tumor suppressors, and we found that these frequencies were highly similar to a previously reported pan-cancer analysis of TCGA data (Figure S1H; $R = 0.99$, $p < 0.0001$) (Bailey et al., 2018).

After establishing the validity of our analysis pipeline, we conducted two types of Cox analysis using the processed clinical and genomic data (Figure 1A). First, we generated univariate models, in which individual genomic features were directly associated with patient outcome (Table S1). Secondly, we generated multivariate ("fully adjusted") models, in which patient age, sex, tumor stage, and tumor grade were incorporated along with the genomic data (Table S2). For each Cox model, we report the Z score, which encodes both the directionality and significance of a survival relationship. Z scores across cancer types were combined using Stouffer's method (Stouffer, 1949). A Z score >1.96 indicates that the presence or upregulation of a tumor feature is associated with shorter survival times at a $p < 0.05$ threshold, while a Z score < -1.96 indicates that the presence or upregulation of a feature is associated with longer survival times at a $p < 0.05$ threshold. In general, the Z scores produced by the univariate and fully adjusted models were highly concordant within individual data

Table 1. Summary of the patient cohorts and data types included in this study

		Patient counts						
TCGA Cohort	Cancer type	Copy number alterations	DNA methylation	Gene expression	miRNA expression	Mutations	Protein expression	Clinical endpoint
ACC	adrenocortical carcinoma	89	79	79	79	92	46	overall survival
BLCA	bladder urothelial carcinoma	404	411	404	405	406	340	overall survival
BRCA	breast invasive carcinoma	1,063	1,066	1,089	1,064	1,015	882	progression-free interval
CESC	cervical squamous cell carcinoma and endocervical adenocarcinoma	294	306	304	306	289	173	overall survival
CHOL	cholangiocarcinoma	36	36	36	36	36	30	overall survival
COAD	colon adenocarcinoma	425	449	448	418	403	357	overall survival
DLBC	lymphoid neoplasm diffuse large B cell lymphoma	48	48	48	47	37	33	progression-free interval
ESCA	esophageal carcinoma	182	183	184	182	184	126	overall survival
GBM	glioblastoma multiforme	571	428	155	0	390	232	overall survival
HNSC	head and neck squamous cell carcinoma	516	522	519	518	507	212	overall survival
KICH	kidney chromophobe	65	65	65	65	65	62	overall survival
KIRC	kidney renal clear cell carcinoma	506	521	533	499	369	478	overall survival
KIRP	kidney renal papillary cell carcinoma	282	285	289	285	280	214	overall survival
LAML	acute myeloid leukemia	179	178	161	175	130	0	overall survival
LGG	brain lower grade glioma	509	512	514	508	509	427	progression-free interval
LIHC	liver hepatocellular carcinoma	365	372	369	367	361	183	overall survival
LUAD	lung adenocarcinoma	491	512	506	499	506	357	overall survival
LUSC	lung squamous cell carcinoma	481	484	496	461	479	323	overall survival
MESO	mesothelioma	86	86	86	86	81	62	overall survival
OV	ovarian serous cystadenocarcinoma	515	531	295	440	380	402	overall survival
PAAD	pancreatic adenocarcinoma	183	183	178	177	176	123	overall survival
PCPG	pheochromocytoma and paraganglioma	161	178	179	178	178	79	progression-free interval
PRAD	prostate adenocarcinoma	489	495	497	491	494	352	progression-free interval
READ	rectum adenocarcinoma	154	155	159	152	149	130	progression-free interval
SARC	sarcoma	252	257	259	256	236	223	overall survival
SKCM	skin cutaneous melanoma	454	454	454	433	451	342	overall survival
STAD	stomach adenocarcinoma	433	435	409	428	432	351	overall survival
TGCT	testicular germ cell tumors	133	133	134	133	129	104	progression-free interval
THCA	thyroid carcinoma	497	503	505	502	492	377	progression-free interval
THYM	thymoma	122	123	119	123	122	89	progression-free interval
UCEC	uterine corpus endometrial carcinoma	518	534	531	522	529	438	overall survival
UCS	uterine carcinosarcoma	56	57	57	56	57	48	overall survival
UVM	uveal melanoma	80	80	80	80	80	12	overall survival



(legend on next page)

types (median R = 0.96) and within individual cancer types (median R = 0.95), suggesting that few prognostic markers were affected by the inclusion of additional clinical variables (Figure S2). Because of this high degree of concordance, and in order to avoid the reported ambiguities in assessing clinical features, such as stage and grade (Elmore et al., 2015; Evans et al., 2008; Gilks et al., 2013; Griffiths et al., 2006; Lang et al., 2005; Ozkan et al., 2016), we focus our analysis below on the genome-wide univariate models. In addition, we created an online resource, available at http://www.tcga-survival.com/, to facilitate community access to this biomarker dataset.

Identification of genomic features that correlate with cancer patient prognosis

In total, we generated *Z* scores for 3,091,782 univariate Cox models (Figures 1A–1C; Table S1). Across these models, we identified 112,303 genomic feature-cancer type dyads that were significantly associated with patient survival time at a Benjamini-Hochberg false-discovery rate of 1%. Two representative prognostic biomarkers discovered through this analysis are displayed in Figure 1D: CNAs affecting the gene encoding the heme metabolism enzyme BLVRB were identified as an adverse biomarker in endometrial carcinoma (UCEC), while methylation of the gene encoding the keratin-associated protein KRTAP19-7 was identified as a favorable biomarker in sarcoma (SARC). We detected a median of 2,145 significant genomic features per cancer type, out of ~93,000 individual genomic features that were measured (Table S1G). In general, gene expression, DNA methylation, and CNAs provided the most prognostic information, while mutational analysis provided the least (Figure 1E). Cancers that arose from related tissues of origin tended to display similar survival profiles (Figure 1F). For instance, *Z* scores derived from DNA methylation profiles were similar between renal clear cell carcinomas and renal papillary cell carcinomas ($R = 0.48$, $p < 0.0001$).

By randomly permuting gene labels, we discovered that prognostic biomarkers were shared across multiple cancer types significantly more often than expected by chance (Figure S3). However, no single genomic feature was prognostic across all cancer types. The most broadly prognostic features were the expression of the RNA encoding the mitotic kinase PLK1 and the protein levels of the serine protease inhibitor PAI1, both of which were significantly associated with poor outcomes in 18 of 33 cancer types (Figure 1G). Mutations in the tumor suppressor TP53 were an adverse event in 10 cancer types; no other mutation was associated with worse outcomes in more than four cancer types (Figures 1G and S3B). TP53 mutations were also found to correlate with longer survival times in GBM and lung squamous cell carcinomas (LUSCs) (Figure S3C). Mutations in TP53 have previously been recognized as a favorable prognostic biomarker in glioma while, to our knowledge, no such relationship has been observed in LUSCs (Chen et al., 2006; Schmidt et al., 2002). Including TP53 mutations as a variable in multivariate Cox analysis did not significantly affect the identification of prognostic biomarkers (Figure S3D; Table S3).

Gene-level *Z* scores generated via RNA-seq and protein-level *Z* scores generated via reverse-phase protein arrays were highly similar for features that were shared between these platforms (Figure S4). For instance, high expression of transferrin receptor mRNA and high expression of its protein product were both associated with poor outcomes in 11 different cancer types, including ACC (adrenocortical carcinoma) and LGG (Figure S4D). However, the same relationships were not apparent across all genomic platforms. *Z* scores from mutations, methylation profiling, and CNAs were only modestly correlated with *Z* scores at the transcript level (Figures S4A–S4C; $R = -0.09$ to $R = 0.20$). Methylation, mutations, and CNAs have complicated effects on downstream gene expression: for instance, inactivating mutations in TP53 can increase TP53 levels through a feedforward mechanism (Marks et al., 1991; Rodrigues et al., 1990), while certain chromosomal amplifications can fail to increase protein expression due to dosage compensation (McShane et al., 2016; Schukken and Sheltzer, 2021). These results indicate that, while survival profiles generated at the level of transcription and protein expression are comparable, other classes of alterations can provide distinct prognostic information that is not captured solely by measuring the expression of the affected gene(s).

Identification of gene sets and pathways broadly associated with cancer patient outcome

We next sought to understand the biological pathways that were differentially regulated in deadly cancers. We performed gene ontology (GO) enrichment analysis on the genes that were identified by RNA-seq as cross-cancer prognostic biomarkers at a Benjamini-Hochberg false-discovery rate of 1%. Consistent with previous results (Baak et al., 2009; Gentles et al., 2015; Venet et al., 2011), we observed that transcripts overexpressed in aggressive tumors were highly enriched for genes associated

Figure 1. Pan-cancer, cross-platform identification of genomic features associated with patient outcome

(A) Schematic outline of the data processing and analysis performed for this work.

(B) A density plot showing the distribution of genomic feature *Z* scores combined across all six platforms (CNAs, methylation, mutation, gene expression, miRNA expression, and protein expression). The dotted line at $Z = -1.96$ corresponds to $p < 0.05$ for a favorable feature, while the dotted line at $Z = 1.96$ corresponds to $p < 0.05$ for an adverse feature.

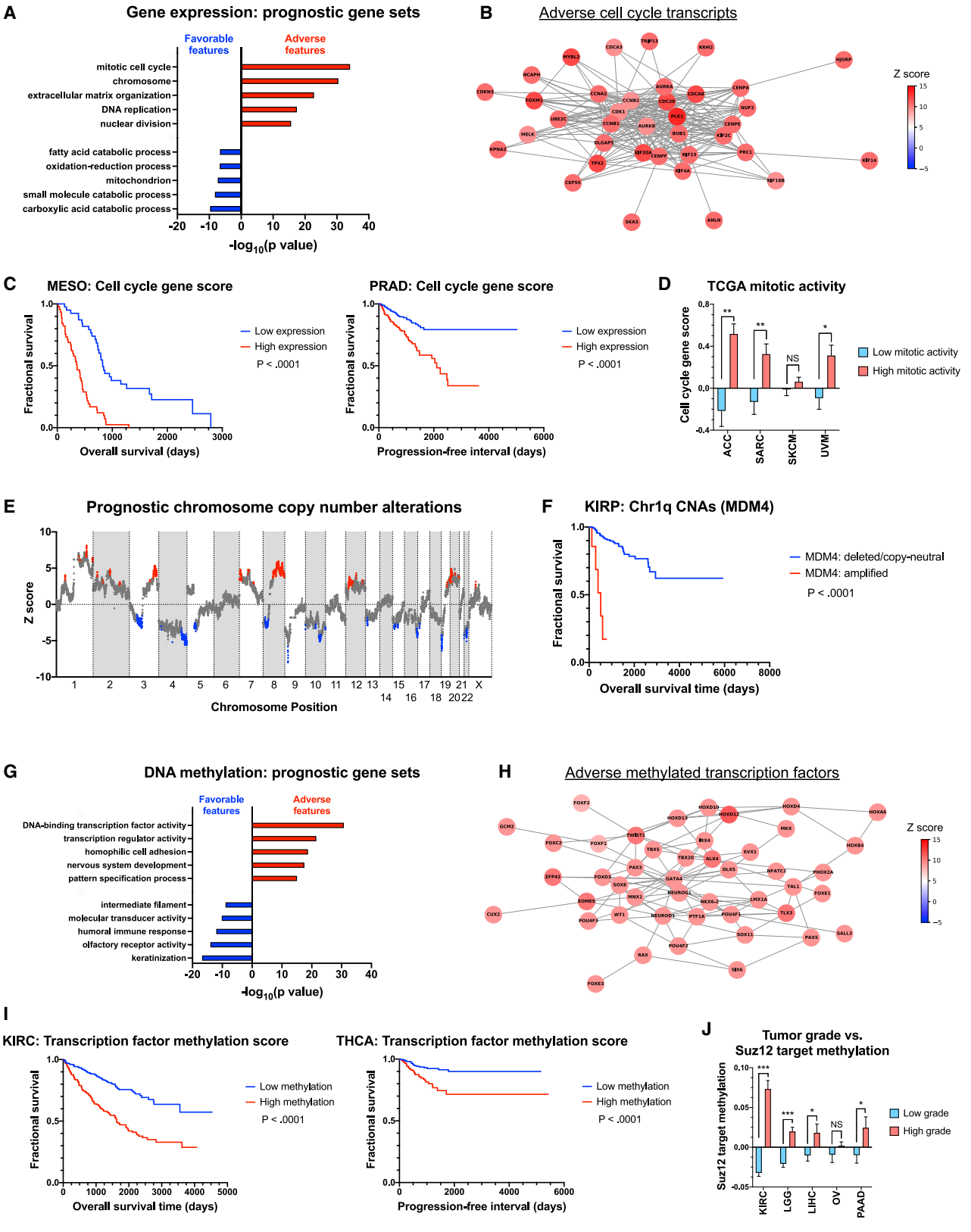
(C) Heatmaps showing the distribution of *Z* scores within each of the six genomic platforms. Each row corresponds to a cancer type and each column corresponds to a gene or genomic feature. The complete set of *Z* scores are included in Table S1.

(D) Kaplan-Meier plots displaying two representative prognostic biomarkers identified from our genome-wide Cox modeling. Copy number amplification of BLVRB is associated with shorter survival times in UCEC (left). Methylation of KRTAP19-7 is associated with longer survival times in SARC (right).

(E) Violin plots showing the distribution of significant prognostic features ($|Z| > 1.96$) per cancer type for each genomic platform.

(F) Cluster plots of the correlation of *Z* scores for each genomic platform across cancer types. The scale indicates the strength of the Pearson correlation coefficient between *Z* score vectors.

(G) Histograms displaying the number of shared prognostic biomarkers for each genomic platform across cancer types.



(legend on next page)

with chromosome segregation, DNA replication, and the mitotic cell cycle (Figures 2A–2C; Tables S1C and S5A). This gene set included several known proliferation markers (MCM2, MKI67, and PCNA) (Whitfield et al., 2006), and promoter analysis revealed that many adverse genes were controlled by the cell-cycle-regulated E2F family of transcription factors (Figures 2B, S5A, and S5B) (Black and Azizkhan-Clifford, 1999). Within the protein expression dataset, cell-cycle-associated proteins, including Cyclin B1 and MSH6, were also among the top-scoring adverse features (Figure S5C; Table S1F). Finally, we observed that the expression of the adverse transcripts was tightly correlated with both cancer cell line doubling times measured *in vitro* and a direct analysis of mitotic activity in tumor specimens (Figures 2D and S5D) (Sheltzer, 2013). In total, these results suggest that the expression of these adverse biomarkers reflect a tumor's proliferative rate. In contrast, favorable expression markers included ACAD11, INPP5K, and CHP, and were enriched for genes whose products localize to mitochondria and are involved in catabolic processes (Tables S1C and S5B).

To find genomic loci where CNAs were associated with patient outcome, we generated a profile of *Z* scores by chromosomal coordinates (Figure 2E). We applied a peak-finding algorithm to identify genomic "peaks," which correspond to loci where copy number gains are associated with poor outcomes, and "valleys," which correspond to loci where deletions are associated with poor outcomes. GO term enrichment analysis revealed that very few gene sets were significantly enriched in either peaks or valleys (Tables S4C and S4D). This indicates that CNAs affecting many different cellular pathways are enriched in deadly tumors. However, manual inspection revealed that a number of known oncogenes and tumor suppressors are encoded in these regions. For instance, the highest peak is found on the q arm of chromosome 1 and encompasses the known cancer driver gene MDM4 (Figure 2F). The deepest valley is found on Chr9p and is centered on the cell-cycle inhibitor CDKN2A, deletion of which has previously been associated with deadly cancers (Smith and Sheltzer, 2018; Zhao et al., 2016). Finally, we determined prognostic peaks and valleys from CNA data for each of the 33 individual cancer types (Table S5). We speculate that some of these regions may harbor genes with uncharacterized roles in cancer biology.

GO analysis of methylation events in tumors with grim prognosis revealed a striking enrichment of transcription factors and genes involved in embryonic development, including *NKX6-1*, *HOXD12*, and *FOXE1* (Figures 2G–2I; Tables S1B and S4E). Favorable methylation events were more diverse and included genes encoding intermediate filaments, olfactory receptors, and keratin-associated proteins (Table S4F). Certain cancers exhibit *de novo* methylation of developmental genes that are silenced by the chromatin-modifying Polycomb complex during embryogenesis (Bracken and Helin, 2009; Schlesinger et al., 2007). These Polycomb targets include lineage-defining transcription factors that are activated or repressed to specify tissue identity. Our finding that developmental transcription factors were methylated in aggressive tumors led us to investigate whether these adverse features were also linked with Polycomb activity. Indeed, we observed a highly significant enrichment of Polycomb component Suz12 binding sites among the genes where high levels of methylation were associated with shorter patient survival (Figure S5E) (Lee et al., 2006). In contrast, Suz12 sites were under-represented among genes where high levels of methylation were favorable features. Expression of EZH2, which encodes the catalytic subunit of the Polycomb complex, and the DNA methyltransferases DNMT1, DNMT3A, and DNMT3B, which cooperate with Polycomb to silence target loci (Viré et al., 2006), were all identified as significant pan-cancer adverse features (Figure S5F and S5G; Table S1A). While the genome-wide correlation between methylation- and transcription-associated survival profiles was minimal (Figure S5A), we found that methylation of these Polycomb-associated loci in particular was associated with decreased gene expression (Figure S5H). Finally, we observed that methylation of these adverse biomarkers was frequently observed in high-grade (dedifferentiated) malignancies (Figure 2J). These data suggest that cancers methylate and silence lineage-defining transcription factors, which promotes the loss of cell identity and is associated with aggressive disease.

Cross-platform identification of the most informative prognostic biomarkers per cancer type

Given the ability to interrogate any gene on any genomic platform in a primary tumor, what measurements confer the most prognostic information? To address this question, we identified the 100 features in each cancer type that exhibit the strongest

Figure 2. Identification of prognostic gene sets across genomic platforms

(A) GO terms enriched among adverse and favorable gene expression biomarkers. The complete set of GO terms are included in Tables S4A and S4B.

(B) Network of interactions among cell-cycle genes, colored according to Stouffer's *Z* from the combined gene expression Cox models.

(C) Kaplan-Meier plots displaying survival in MESO (left) and PRAD (right) based on the mean expression of a set of transcripts associated with the gene ontology term "mitotic cell cycle".

(D) Bar graph showing cell-cycle scores based on pathologically observed mitotic activity in different TCGA cohorts. *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$ (Student's t test).

(E) A plot displaying Stouffer's Z by chromosomal coordinate. Red dots indicate loci where genomic amplifications are associated with worse outcomes and blue dots indicate loci where genomic deletions are associated with worse outcomes. The complete list of genes found within these regions is included in Table S5.

(F) Kaplan-Meier plot displaying survival in KIRP split based on the copy number status of the Chr1q gene MDM4.

(G) GO terms enriched among adverse and favorable methylation biomarkers. The complete set of GO terms are included in Tables S4E and S4F.

(H) Network of interactions among developmental transcription factors colored according to Stouffer's *Z* from the combined methylation Cox models.

(I) Kaplan-Meier plots displaying survival in KIRC (left) and THCA (right) based on the methylation of a collection of developmental transcription factors.

(J) Bar graph showing the average methylation of Suz12 targets based on tumor grade in different TCGA cohorts. *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$ (Student's t test).

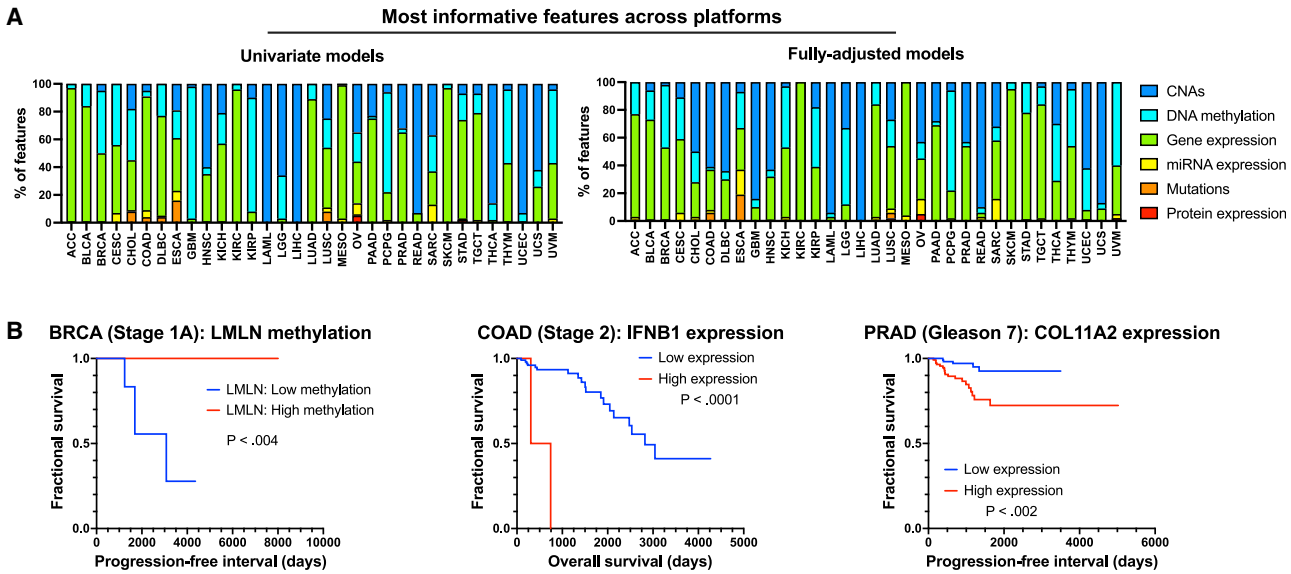


Figure 3. Cross-platform analysis reveals that cancer gene expression measurements harbor the most prognostic information

(A) The 100 genomic features that exhibit the strongest associations with patient outcomes in univariate or fully adjusted Cox models are displayed.

(B) Kaplan-Meier plots displaying patient survival in stage 1A breast cancer, stage 2 colon cancer, and Gleason 7 prostate cancer, split based on the indicated biomarkers.

overall correlations with patient outcome in both univariate and fully adjusted models (Figure 3A). We found that, for the average cancer type, 46% of top-scoring features were gene expression measurements, 22% were methylation events, 30% were CNAs, and only 1% were mutations. Some cancers diverged from this overall trend: in GBM, 95 of the top 100 univariate biomarkers were favorable methylation events, which likely reflects the CpG island methylator phenotype that has been linked with long-term survival in brain cancers (Shinawi et al., 2013). Interestingly, several top prognostic features within individual cancer types are poorly characterized. For instance, in stomach adenocarcinoma, across 96,730 Cox models that we generated, the genomic feature that exhibited the strongest association with patient outcome was expression of the uncharacterized lncRNA FLJ16779/LOC100192386. Finally, we found that classifying patients based on single features identified through this analysis was sufficient to distinguish outcomes in ambiguous clinical situations where patients are at risk of undertreatment or overtreatment. This includes stage 1A breast cancer (Elias, 2012), stage 2 colon cancer (Booth et al., 2017; Lee et al., 2019), and Gleason 7 prostate cancer (Srigley et al., 2019; Stark et al., 2009) (Figure 3B).

Overexpression or mutation of verified cancer driver genes is not widely associated with poor prognosis

When the expression or mutation of a gene is found to be associated with poor patient prognosis, this is typically presented as evidence that that gene is an important driver of disease progression (Kaelin, 2017). However, we were surprised to find that very few established oncogenes or tumor suppressors were identified as significant adverse features in our genome-wide analyses described above. Commonly mutated cancer driver genes, including KRAS, PIK3CA, CTNNB1, RB1, and APC, were not recovered as prominent biomarkers. Instead, the strongest biomarkers in our expression analysis tended to be housekeeping genes with roles in the cell cycle. In our sequencing analysis, TP53 mutations were identified as an adverse feature in 10 of 33 cancer types, but no other gene was significant in more than four cancer types. These findings challenged the notion that we should expect important cancer driver genes to be associated with patient outcomes. We therefore decided to study these unexpected results more closely.

To systematically examine the prognostic significance of mutations in cancer driver genes, we assessed two collections of verified oncogenes. First, we considered a set of 31 genes that exhibit pan-cancer oncogenic activity (BRAF, EGFR, KRAS, etc.) (Bailey et al., 2018), and we calculated Z scores for these genes in each of the 33 TCGA cancer types. Secondly, we considered an expanded set of 81 oncogenes, but we only calculated Z scores for these genes in cancer types in which that gene is recurrently activated (e.g., FLT3 in LAML, EGFR in LUAD, ERBB2 in BRCA, etc.) (Bailey et al., 2018). We found that, considered as a group, the Z scores for these oncogenes were not enriched for prognostic features relative to randomly permuted gene sets of the same size (Figures 4A and 4B). The mean Z score for the pan-cancer oncogene set was −0.06, and the mean Z score for the tissue-limited oncogene set was −0.08. Out of 1,023 possible cancer type-oncogene pairs, we found that mutations in a pan-cancer oncogene were associated with worse prognosis in <1% of instances, which was not significantly different from the background rate of prognostic mutations across all genes (Figure 4C). Analyzing individual Kaplan-Meier curves supported these findings. For example, EGFR is a known driver of glioblastoma (Frederick et al., 2000), but EGFR mutations were not associated with poor prognosis in

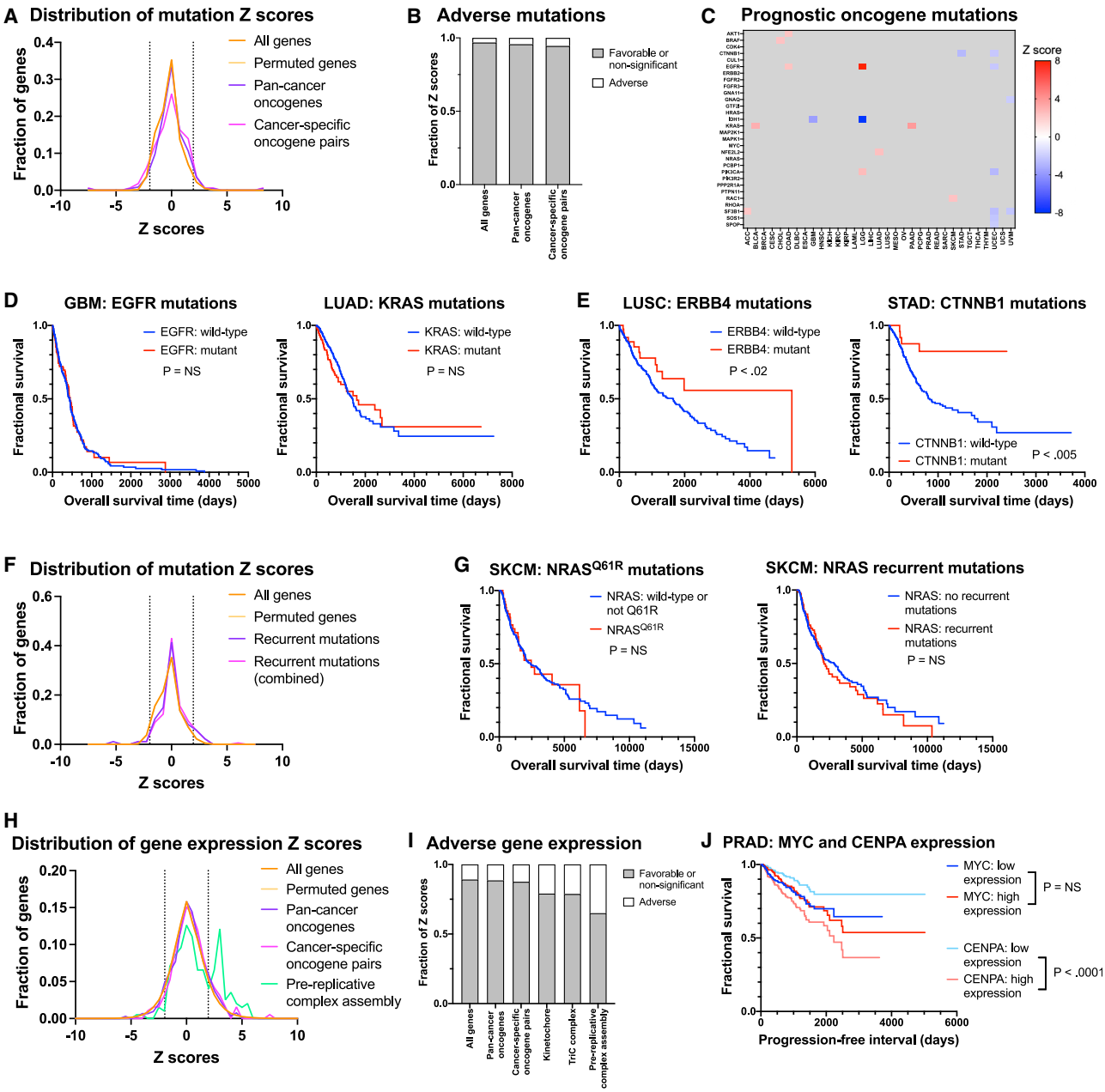


Figure 4. Oncogene mutation or overexpression is not widely associated with patient outcome

(A) A density plot showing the distribution of mutation *Z* scores for the indicated gene sets. The dotted line at $Z = -1.96$ corresponds to $p < 0.05$ for a favorable mutation, while the dotted line at $Z = 1.96$ corresponds to $p < 0.05$ for an adverse mutation.

(B) Stacked bar graph showing the fraction of mutations that are associated with adverse outcomes for the indicated gene sets.

(C) A heatmap showing significant ($|Z| > 1.96$) survival associations for oncogene mutations. Each row represents a pan-cancer oncogene identified in Bailey et al. (2018) and each column represents a patient cohort from TCGA.

(D) Kaplan-Meier plots showing that mutations in the established GBM oncogene EGFR (left) and the established LUAD oncogene KRAS (right) are not associated with shorter survival times.

(E) Kaplan-Meier plots showing that mutations in the LUSC oncogene ERBB4 (left) and the STAD oncogene CTNNB1 (right) are associated with longer survival times.

(F) A density plot showing the distribution of mutation *Z* scores for the indicated gene sets, including "hotspot" mutations that affect specific recurrently mutated codons. The complete list of mutation *Z* scores are included in Table S6.

(G) Kaplan-Meier plots demonstrating that the most common NRAS driver mutation—Q61R—or a combination of all common NRAS driver mutations—G12D, G12S, G13D, G13R, K16N, Q61H, Q61K, Q61R, and E62K—are not associated with shorter survival times in SKCM.

(legend continued on next page)

the TCGA GBM cohort (Figure 4D). Indeed, in several cases, we observed that mutations in established driver oncogenes, such as CTNNB1 and ERBB4, were associated with favorable outcomes relative to cancers that lacked mutations in these genes (Figure 4E).

In the above analysis, all cancers that harbored a non-synonymous mutation in a gene of interest were classified as "mutant" for that gene. While we recognize that different mutations may have different functional effects, manual inspection of the data revealed that most of these mutations were likely to exhibit oncogenic activity. For instance, >95% of non-synonymous mutations in KRAS were found in codons 12, 13, or 61, which have all been identified as sites of driver mutations (Muñoz-Maldonado et al., 2019). However, we considered the possibility that our above analysis could be obscuring certain specific mutations with prognostic significance. Accordingly, we conducted two further analyses: (1) we analyzed individual recurrent mutations separately (e.g., we calculated *Z* scores separately for tumors with $KRAS^{G12V}$ mutations, $KRAS^{G12R}$ mutations, $KRAS^{G13D}$ mutations, etc.) and (2) we combined patients with any recurrently observed mutation within a gene but excluded patients with non-synonymous mutations that were not found in a commonly altered codon. However, these analyses were largely consistent with our whole-gene analysis and revealed few prognostic mutations (Figures 4F and S6; Table S6). Many of the most common cancer driver mutations, including $PIK3CA^{E545K}$, $KRAS^{G13D}$, $IDH1^{R132H}$, and $FBXW7^{R465H}$, were not associated with worse outcomes in any of the 33 TCGA cohorts (Figure S6). For instance, while NRAS mutations are an established driver of melanoma, neither the most common NRAS alteration (Q61R) nor a combination of all common NRAS alterations predicted poor survival (Figure 4G) (Hodis et al., 2012). In total, these results indicate that mutations in verified cancer driver genes are not widely associated with adverse outcomes.

Next, we investigated whether the overexpression of verified oncogenes was associated with shorter survival times. We calculated *Z* scores for the two collections of driver oncogenes described above and compared them with randomly permuted gene sets of the same size. Consistent with our mutational analysis, we found that the expression of verified oncogenes was no more likely to be an adverse prognostic feature than the expression of a randomly selected gene (Figure 4H). Indeed, we found that oncogenes harbored less prognostic power than sets of genes encoding the kinetochore, the TriC complex, or the pre-replication complex, which are not known to harbor oncogenic activity but are associated with cell-cycle progression (Figure 4I). For instance, in prostate cancer, we observed that high expression of the driver oncogene MYC was not associated with adverse outcomes, but high expression of the kinetochore component CENPA was strongly associated with disease progression (Figure 4J). Taken together, these results demonstrate that the expression and/or mutation of key cancer driver genes is not a robust predictor of poor patient outcomes.

Successful cancer drugs generally do not target adverse prognostic genes

Many papers characterizing a novel drug or drug target in cancer biology present evidence that the overexpression or mutation of that drug target is associated with aggressive disease (Kaelin, 2017). The assumption underlying this line of evidence is that genes that are associated with shorter survival times make the best targets for therapeutic development. However, to our knowledge, this assumption has never been directly tested. We therefore set out to explore whether successful cancer drugs that currently exist are likely to target genes that are associated with poor prognosis.

We generated a list of FDA-approved cancer drugs matched with each drug's reported target(s) (Table S7A). We then calculated *Z* scores for each drug target in the cancer type(s) for which that drug has received FDA approval. Surprisingly, we found that successful drug targets were not generally enriched for adverse prognostic factors (Figures 5A–5D; Tables S7B and S7C). Out of 212 target-cancer type pairs, mutation of <2% of targets was associated with worse outcomes. Among gene expression biomarkers, drug target *Z* scores were not significantly greater than the *Z* scores of randomly permuted gene sets. In fact, we observed that FDA-approved drugs were as likely to target a gene whose expression correlated with favorable prognosis as they were to target a gene whose expression correlated with poor prognosis (12% versus 17%, p = NS; Figure 5D).

Closer inspections of individual genes and patient cohorts revealed multiple factors that contribute to the minimal overlap between adverse biomarkers and successful drug targets (Figures 5E–5J). First, we observed that some genes were strongly upregulated in certain cancer types, but within those cancer types, expression or mutation of that gene was non-prognostic. For instance, the FDA-approved LAML therapy gemtuzumab ozogamicin consists of a DNA damaging agent conjugated to an antibody targeting the CD33 antigen. CD33 expression is strongly upregulated in myeloid cells (Laszlo et al., 2014), which confers specificity to this agent; but, within LAML, variation in CD33 levels do not correlate with aggressive disease (Figure 5H). Secondly, many targetable driver mutations are mutually exclusive and serve to activate the same signaling pathway (Gainor et al., 2013; Mack et al., 2020; Unni et al., 2015). For instance, lung cancers can harbor targetable mutations in ALK, BRAF, EGFR, MEK1, MET, RET, or ROS1, each of which activates the MAPK signaling pathway. There is no *prima facie* reason to believe that any one of these genes would be associated with worse prognosis than all others and, in the LUAD dataset, none of these mutations were correlated with outcome (Figure 5J; Table S1E).

(H) A density plot showing the distribution of gene expression *Z* scores for the indicated gene sets. Note that while the oncogene sets largely overlap with the control gene sets, the pre-replicative complex assembly gene set displays a significant peak to the right of the control gene sets.
(I) Stacked bar graph showing the fraction of gene expression biomarkers that are associated with adverse outcomes for the indicated gene sets.
(J) Kaplan-Meier plot showing survival in PRAD split based on the expression of a key driver oncogene (MYC) or split based on the expression of a non-oncogenic cell-cycle gene (CENPA).

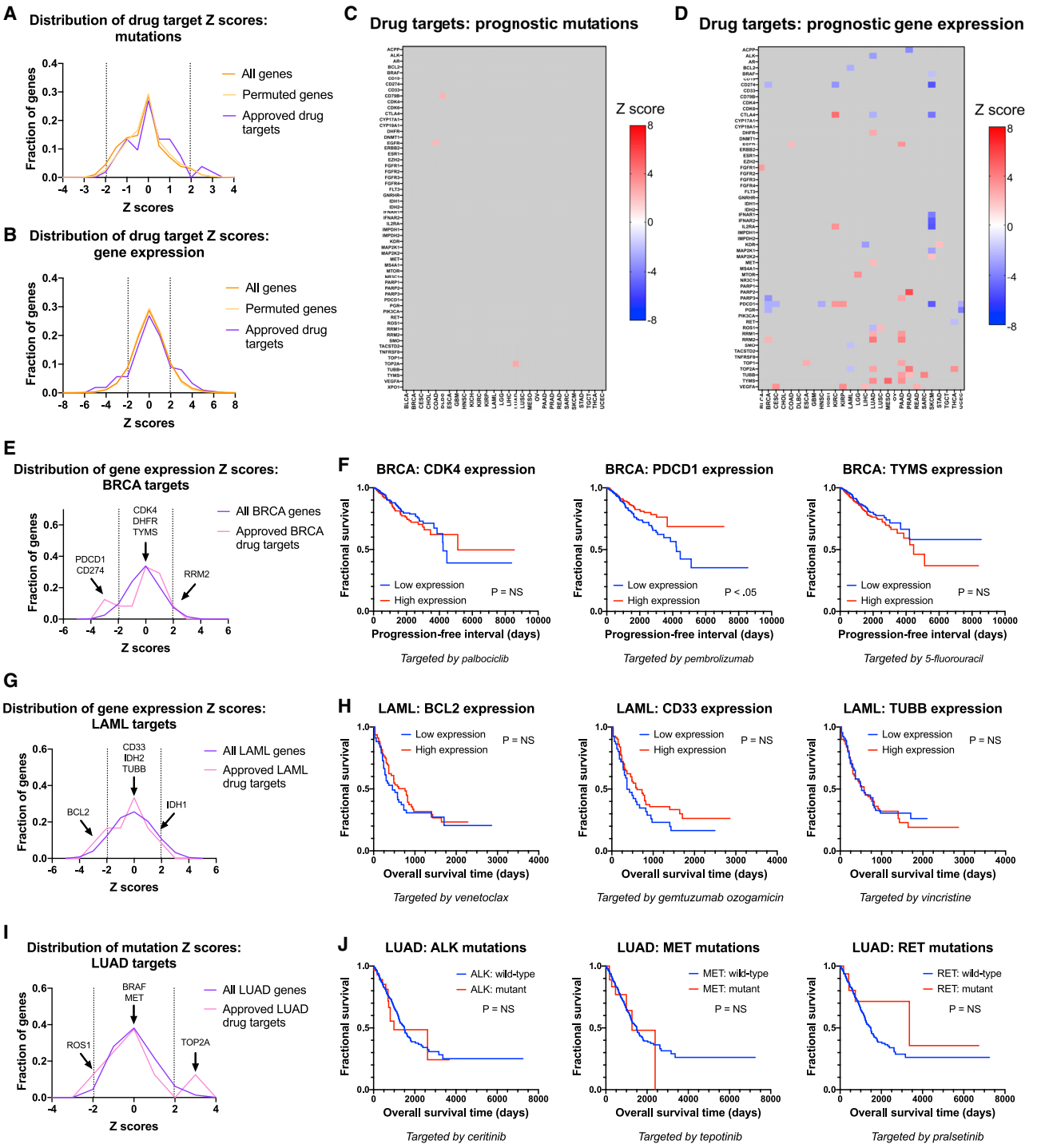


Figure 5. The genomic targets of FDA-approved cancer drugs are not strong prognostic biomarkers

(A) A density plot showing the distribution of mutation Z scores for the indicated gene sets. The dotted line at $Z = -1.96$ corresponds to $p < 0.05$ for a favorable mutation, while the dotted line at $Z = 1.96$ corresponds to $p < 0.05$ for an adverse mutation.

(B) A density plot showing the distribution of gene expression Z scores for the indicated gene sets.

(C) A heatmap showing significant ($|Z| > 1.96$) survival associations for mutations in the targets of FDA-approved drugs. Each row represents a drug target and each column represents a cancer patient cohort from TCGA.

(D) A heatmap showing significant ($|Z| > 1.96$) survival associations for expression changes in the targets of FDA-approved drugs. Each row represents a drug target and each column represents a cancer patient cohort from TCGA.

(legend continued on next page)

Third, many cancer therapies exhibit significant cancer cell non-autonomous effects. For instance, breast tumors with high expression levels of PDCD1 (PD1) have superior outcomes relative to breast tumors with low PDCD1 expression (Figure 5F). Based on this survival correlation, one might assume that an ideal cancer therapy would upregulate PD1 expression and, correspondingly, inhibiting PD1 should decrease patient survival. However, antibodies such as pembrolizumab that inhibit PD1 have a pronounced benefit in breast cancer and several other cancer types (Darvin et al., 2018; Singh et al., 2021; Sun et al., 2020). In this case, PD1 is expressed by immune cells (Ahmadzadeh et al., 2009; Francisco et al., 2010), and high PD1 expression is evidence of tumor-controlling immune infiltration (Ali et al., 2014).

Finally, some targetable genes play key roles in cancer biology, even though their expression levels are uncorrelated with disease severity. For instance, thymidylate synthetase (TYMS) is required for DNA replication, and TYMS inhibitors, such as 5-fluorouracil, are effective at blocking DNA replication in several cancer types (Jarmula, 2010; Peters et al., 2002). TYMS inhibitors can thereby prolong patient survival, even though TYMS is not known to be an oncogene and TYMS upregulation does not drive disease progression (Figure 5F). In total, these and other factors may contribute to our observation that a large majority of successful cancer drugs do not target genes that are associated with poor patient outcomes.

We also considered an alternate explanation for these results: tumors harboring a mutation or overexpression of a drug target may be treated with that drug, and so the lack of association between these genes and aggressive disease could be a reflection of the treatment received rather than underlying differences in cancer biology. To investigate this possibility, we conducted an additional analysis using only drug/indication pairs that received FDA approval in 2018 or later. As TCGA patient follow-up stopped in 2015/2016, we expect that extremely few patients in these cohorts would have received these targeted therapies. However, our findings with this subset of drugs were consistent with our analysis of the complete dataset and revealed very few prognostic correlations among drug targets (Figures S7A and S7B). For instance, the FGFR inhibitor erdafitinib received FDA approval for use in FGFR3-mutant bladder cancer in 2019 (Markham, 2019), but neither FGFR3 mutations nor FGFR3 overexpression were prognostic in the TCGA BLCA cohort collected prior to this time (Figures S7C and S7D). In total, these results demonstrate that successful cancer drugs generally do not target biomarkers associated with aggressive disease.

Many therapies targeting adverse prognostic factors have failed in clinical trials

To further evaluate the relative importance of targeting genetic features that are associated with aggressive tumors, we focused on the 50 genes whose expression exhibits the strongest correlation with adverse outcomes across cancer types (Table S1C). Using http://www.clinicaltrials.gov and other related resources, we identified therapeutic agents designed to target these top-scoring genes that have been tested in patients (Figures 6A and 6B). We found that 16 of the top 50 genes have been targeted in clinical trials, but therapies against 15 of these genes have failed to receive FDA approval. For instance, the top-scoring prognostic factor in our analysis is the mitotic kinase PLK1, and small-molecule compounds designed to block PLK1 activity have been tested in patients with multiple cancer types (Gutteridge et al., 2016). However, PLK1 inhibitors were found to cause severe and sometimes fatal side effects, thwarting their clinical utility (Green and Konig, 2020). Other inhibitors against top-scoring genes, including CDK1, AURKA, AURKB, and CENPE, were similarly found to exhibit unacceptable toxicities or insufficient activity (Figure 6A). Peptides derived from several top-scoring genes have also been used in immunotherapy vaccines, though their therapeutic efficacy has not been demonstrated in randomized trials. Of the 50 genes that exhibit the strongest correlations with aggressive disease, only a single gene, RRM2, is targeted by an FDA-approved compound (Aye et al., 2015).

In our GO analysis, we noted that many top-scoring prognostic factors were widely expressed housekeeping genes with crucial roles in cell-cycle progression. We hypothesized that the frequent failure of these top-scoring genes as cancer drug targets could result from the fact that many of them represent broadly essential genes, potentially leading to significant side effects when their proteins are inhibited. To investigate this hypothesis, we analyzed cancer dependency scores from whole-genome CRISPR screening data across several hundred cancer cell lines (https://doi.org/10.6084/m9.figshare.6931364.v1; Meyers et al., 2017). Each score measures the fitness effects of ablating the gene in question, with larger negative scores indicating more significant fitness defects upon gene loss. We observed that the dependency score distribution for FDA-approved cancer targets was very similar to the distribution of scores across all genes (Figures 6C and 6D). While some approved drugs inhibit pan-cancer dependencies (e.g., TUBB, TOP1, and TOP2), a majority of targeted genes exhibit more selective effects (e.g., ESR1, PARP1, and ALK). In contrast, the top-scoring prognostic factors exhibit essentiality patterns that are significantly different from the essentiality patterns of approved drugs. We observed that 50% of top-scoring prognostic factors are essential across all cell types, compared with only 15% of genes targeted by approved drugs (Figure 6E). This limited cell-type selectivity could contribute to the toxicity and high failure rate of drugs designed to target these top-scoring genes. In total, these analyses suggest that prioritizing targets for therapeutic

(E) A density plot showing the distribution of gene expression *Z* scores for the indicated gene sets in BRCA.
(F) Kaplan-Meier plots displaying survival times in BRCA based on the expression levels of the BRCA drug targets CDK4, PCDC1, and TYMS.
(G) A density plot showing the distribution of gene expression *Z* scores for the indicated gene sets in LAML.
(H) Kaplan-Meier plots displaying survival times in LAML based on the expression levels of the LAML drug targets BCL2, CD33, and TUBB.
(I) A density plot showing the distribution of mutation *Z* scores for the indicated gene sets in LUAD.
(J) Kaplan-Meier plots displaying survival times in LUAD based on the mutations in the indicated LUAD drug targets ALK, MET, and RET.

A Clinical trials targeting the top 50 prognostic factors

Gene and prognostic ranking	Target	FDA approval?	Highest clinical phase	Therapeutic agents	NCT # of highest trial	Notes	Citations
(1) PLK1	Polo-like kinase 1	No	Phase 3	Volasertib, BI2536, GSK461364	NCT01721876	Early clinical trials suggested activity for volasertib, but significant toxicities, including fatal infections, were also reported.	PMID: 30104712, PMID: 27330107
(3) FOXM1	FOXM1-expressing cells	No	Phase 2	OTSGC-A24	NCT01227772	FOXM1 peptides have been used as immunotherapy vaccines. No therapeutic responses were seen in 20 patients with gastric cancer.	PMID: 29587677
(4) KIF20A	KIF20A-expressing cells	No	Phase 2	KIF20A peptides with adjuvant	NCT01950156	KIF20A peptides have been used as immunotherapy vaccines. No randomized trials have been conducted.	PMID: 31571332
(13) AURKA	Aurora kinase A	No	Phase 3	Alisertib, Tozasertib, ENMD-2076, MK-5108	NCT00952588	Alisertib had no significant benefit in a randomized phase 3 trial in lymphona. Other trials are ongoing.	PMID: 27466629
(18) HJURP	HJURP-expressing cells	No	Phase 2	HJURP peptides	Not listed	HJURP peptides have been used as immunotherapy vaccines. Minimal evidence of therapeutic efficacy.	PMID: 25335716
(24) CCNB1, (25) CCNA2, (45) CCNB2, (46) CDK1	Cyclin/CDK1 and Cyclin/CDK2 complexes	No	Phase 3	Flavopiridol, dinaciclib, AT7519, Selicicib	NCT01580228	Pan-CDK inhibitors have been tested in >50 clinical trials. Dose-limiting toxicities have been observed in several studies, and the single Phase 3 study that has been initiated was terminated eary.	PMID: 25633797, PMID: 30257348
(28) RRM2	Ribonucleotide reductase	Yes	FDA approved	Hydroxyurea, gemcitabine	-	Ribonucleotide reductase inhibitors have received FDA approval for use in several cancer types, including breast and lung cancers.	PMID: 24083455
(32) NUF2	NUF2-expressing cells	No	Phase 2	NUF2 peptides	NCT01267578	NUF2 (CDCA1) peptides have been used as immunotherapy vaccines. Minimal evidence of therapeutic efficacy.	PMID: 28498618
(36) IGF2BP3	IGF2BP3-expressing cells	No	Phase 2	IGF2BP3 peptides	NCT00681577	IGF2BP3 (KOC1) peptides have been used as immunotherapy vaccines. Minimal evidence of therapeutic efficacy.	PMID: 27072896
(38) CENPE	CENPE kinesin	No	Phase 1	GSK923295	NCT00504790	A therapeutic response to GSK923295 was observed in 1 of 39 treated patients. Study completed in 2012, no evidence that future trials are planned.	PMID: 22020315
(40) DEPDC1	DEPDC1-expressing cells	No	Phase 2	DEPDC1 peptides	NCT00633204	DEPDC1 peptides have been used as immunotherapy vaccines. Minimal evidence of therapeutic efficacy.	PMID: 30791546
(42) AURKB	Aurora kinase B	No	Phase 3	AZD1152, GSK1070916, AT9283	NCT00952588	Along with Aurora A, Aurora B is a key driver of mitosis. However, significant toxicities and limited efficacy have stymied trial success.	PMID: 28918096
(48) MELK	MELK kinase	No	Phase 2	OTS167	NCT02795520	While OTS167 was entered into clinical trials as a MELK inhibitor, subsequent research questioned its mechanism-of-action. Clinical trial results have not been reported.	PMID: 28337968, PMID: 29417929

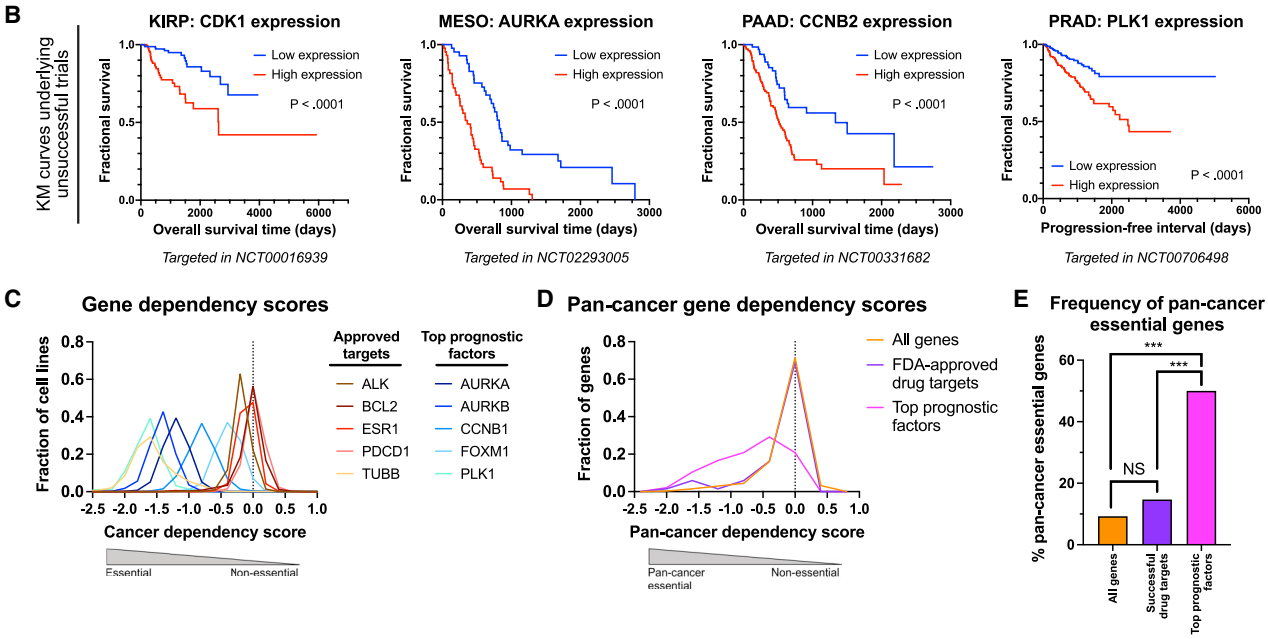


Figure 6. Therapies targeting top prognostic genes have failed in clinical trials

(A) A table displaying the genes among the 50 prognostic factors that exhibit the strongest correlations with cancer patient outcomes that have been targeted in cancer clinical trials.

(B) Kaplan-Meier plots showing patient survival in the indicated cancer cohorts. Each graph displays a gene that has been targeted in clinical trials in that cancer type.

(C) A density plot showing the distribution of cancer dependency scores for the indicated genes, split according to whether the gene is the target of an FDA-approved cancer therapy or whether the gene is a top-scoring prognostic factor.

(D) A density plot showing the distribution of pan-cancer cancer dependency scores for the indicated gene sets.

(E) A bar graph showing the percent of genes that are essential across cancer types in the indicated gene sets. ***$p < 0.0005$ (hypergeometric test).

development based on prognostic correlations may be counterproductive, as a large fraction of these correlated factors represent ubiquitously expressed housekeeping genes rather than cancer-specific dependencies.

DISCUSSION

Genomic analysis has the potential to shed unprecedented insight into the molecular architecture of human cancers. In light of studies demonstrating both the pervasive overtreatment and undertreatment of cancer patients, the use of genomic technologies to discover and validate prognostic biomarkers could greatly enhance risk prediction and clinical treatment decisions. In this work, we generated a rich dataset of more than 3,000,000 individual Cox proportional hazards models and identified more than 100,000 significant prognostic biomarkers across 33 cancer types. These data have also been shared via a web portal at http://www.tcga-survival.com to facilitate further analysis.

Our study illustrates the unexpected prognostic potential of different classes of genomic data. For instance, while there has been substantial attention devoted toward developing routine whole-exome and targeted sequencing panels for clinical use (Berger and Mardis, 2018; Conway et al., 2019), our findings demonstrate that relatively few point mutations are significantly associated with cancer patient outcome. Aside from mutations in TP53, which were prognostic in 12 of 33 patient cohorts, mutations in established cancer genes, such as CDKN2A, EGFR, KRAS, PIK3CA, PTEN, RB1, and many others, had extremely limited prognostic power. Sequencing oncogenes and tumor suppressors may be useful in order to assign patients to specific targeted therapy regimens, and larger patient cohorts sequenced at greater depths may identify prognostic relationships not found in this study. Nonetheless, considered as a whole, this work suggests that routine sequencing of patient tumors will not yield significant improvements in risk prediction relative to other potential genomic platforms.

In contrast to the paucity of prognostic mutations uncovered through this work, we identified several hundred genes whose methylation was associated with outcome across cancer types. GO analysis demonstrated that these genes were enriched for developmental transcription factors, and that these genes were typically downregulated in high-grade tumors. Polycomb activity may thereby facilitate cancer cell reprogramming and a loss of cellular identity, returning cancers to a stem cell-like state that can rapidly progress (Bracken and Helin, 2009). Similarly, the most penetrant gene expression biomarkers were cell-cycle-associated transcripts that were upregulated in cancers with high mitotic activity. Importantly, cellular proliferation has a profound influence on gene expression genome wide, and so many genes with diverse functions may still be prognostic by indirectly capturing cell-cycle activity (Venet et al., 2011). In future work, the construction of multivariate Cox models incorporating proliferation markers, such as MKI67 and PCNA, may help differentiate between cell-cycle-dependent and cell-cycle-independent prognostic features.

Our findings also have significant implications for the analysis of cancer survival data in a preclinical or therapeutic-discovery setting. Using multiple datasets of verified oncogenes, we unambiguously demonstrate that genes that drive tumorigenesis are not significantly enriched among adverse biomarkers within cohorts of cancer patients. Similarly, while genes associated with metastasis and patient death are sometimes presumed to encode the most promising targets for therapeutic development, we show that successful cancer drugs generally do not target adverse biomarkers. Correspondingly, a large majority of experimental drugs that do target adverse biomarkers have failed in clinical trials. We believe that these results underscore a crucial distinction between causation and correlation in clinical observations. To illustrate, among a random group of adults, individuals receiving kidney dialysis are more likely to die than individuals who are not receiving dialysis. Based strictly on this correlative observation, one could assume that kidney dialysis kills people. Yet, we know that people receiving dialysis are likely to be older and have several medical comorbidities, and dialysis saves their lives (Kaelin, 2017; Henrich and Burkat, 2021).

In general, we caution that deducing functional relationships and prioritizing drug targets based on cancer survival data may be inappropriate, and that such relationships may be fraught with confounding variables and spurious correlations. From our analysis, one could incorrectly infer that KRAS mutations are not important in lung cancer (Figure 4D), or that kinetochore gene expression is a more important driver of prostate cancer than MYC expression (Figure 4J). Moreover, we observed that strongly prognostic genes tend to be widely essential across cell types, which could explain why so many therapies designed against these genes have exhibited dangerous side effects in human patients. Consequently, leveraging survival analysis to select targets for therapeutic development could inadvertently prioritize targets that are unlikely to succeed in clinical testing. Alternative genomic approaches and functional studies in which causative relationships can be interrogated are necessary to rigorously identify potential drug targets and genes that drive cancer progression. We suggest that, in general, the use of survival data to identify prognostic biomarkers should be decoupled from the use of survival data to infer gene function in cancer biology.

Limitations of the study

In an ideal biomarker discovery study, the patients within each cohort would receive uniform treatment, thereby minimizing one potential source of inter-patient variability. Patients analyzed as part of the TCGA received heterogeneous treatments, which may confound the identification of prognostic biomarkers. In addition, certain cancer types within the TCGA include fewer than 100 patients, which may leave them underpowered for comprehensive biomarker identification (Table 1). Despite these limitations, we note that our analysis correctly recapitulates many established prognostic features, including patient age, tumor grade, tumor stage, TP53 mutations, cell-cycle gene expression, and more (Figure 2A, 2B, S1C–S1E, and S3B). Finally, each cancer type within this study is represented by a single patient cohort, and any individual biomarkers of interest should be validated in multiple independent cohorts prior to clinical application.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Survival analysis in TCGA cohorts
  - Selection of analysis methodology
  - Overall analysis strategy
  - Kaplan-Meier analysis
  - Gene ontology analysis
  - Additional tools and resources
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110569.

ACKNOWLEDGMENTS

Research in the Sheltzer Lab is supported by an NIH Early Independence award (1DP5OD021385), NIH grant R01CA237652, Department of Defense grant W81XWH-20-1-068, a Damon Runyon-Rachleff Innovation award, an American Cancer Society Research Scholar Grant, and a grant from the New York Community Trust.

AUTHOR CONTRIBUTIONS

J.C.S. and J.M.S. conceived, designed, and performed the analysis described in this work. J.C.S. and J.M.S. wrote the manuscript and prepared the figures.

DECLARATION OF INTERESTS

J.C.S. is a co-founder of Meliora Therapeutics, a member of the advisory board of Surface Ventures, and an employee of Google, Inc. This work was performed outside of her affiliation with Google and used no proprietary knowledge or materials from Google. J.M.S. has received consulting fees from Ono Pharmaceuticals and Merck, is a member of the advisory board of Tyra Biosciences, and is a co-founder of Meliora Therapeutics.

Received: October 19, 2021
Revised: January 30, 2022
Accepted: March 3, 2022
Published: March 29, 2022

REFERENCES

Ahmadzadeh, M., Johnson, L.A., Heemskerk, B., Wunderlich, J.R., Dudley, M.E., White, D.E., and Rosenberg, S.A. (2009). Tumor antigen–specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. Blood *114*, 1537–1544.

Ali, H.R., Provenzano, E., Dawson, S.-J., Blows, F.M., Liu, B., Shah, M., Earl, H.M., Poole, C.J., Hiller, L., Dunn, J.A., et al. (2014). Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. Ann. Oncol. *25*, 1536–1543.

Amar, D., Izraeli, S., and Shamir, R. (2017). Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. Oncogene *36*, 3375–3383.

Anaya, J. (2016). OncoRank: a pan-cancer method of combining survival correlations and its application to mRNAs, miRNAs, and lncRNAs. PeerJ Preprint. *4*, e2574v1.

Anaya, J., Reon, B., Chen, W.-M., Bekiranov, S., and Dutta, A. (2016). A pan-cancer analysis of prognostic genes. PeerJ *3*, e1499.

Andre, F., McShane, L.M., Michiels, S., Ransohoff, D.F., Altman, D.G., Reis-Filho, J.S., Hayes, D.F., and Pusztai, L. (2011). Biomarker studies: a call for a comprehensive biomarker study registry. Nat. Rev. Clin. Oncol. *8*, 171–176.

Aye, Y., Li, M., Long, M.J.C., and Weiss, R.S. (2015). Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. Oncogene *34*, 2011–2021.

Baak, J.P.A., Gudlaugsson, E., Skaland, I., Guo, L.H.R., Klos, J., Lende, T.H., Søiland, H., Janssen, E.A.M., and Zur Hausen, A. (2009). Proliferation is the strongest prognosticator in node-negative breast cancer: significance, error sources, alternatives and comparison with molecular prognostic markers. Breast Cancer Res. Treat. *115*, 241–254.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Berger, M.F., and Mardis, E.R. (2018). The emerging clinical relevance of genomics in cancer medicine. Nat. Rev. Clin. Oncol. *15*, 353–365.

Bijker, N., Donker, M., Wesseling, J., Heeten, G.J., and Rutgers, E.J.T. (2013). Is DCIS breast cancer, and how do I treat it? Curr. Treat. Options Oncol. *14*, 75–87.

Black, A.R., and Azizkhan-Clifford, J. (1999). Regulation of E2F: a family of transcription factors involved in proliferation control. Gene *237*, 281–302.

Booth, C.M., Nanji, S., Wei, X., Peng, Y., Biagi, J.J., Hanna, T.P., Krzyzanowska, M.K., and Mackillop, W.J. (2017). Adjuvant chemotherapy for stage II colon cancer: practice patterns and effectiveness in the general population. Clin. Oncol. *29*, e29–e38.

Bouchardy, C., Rapiti, E., Fioretta, G., Laissue, P., Neyroud-Caspar, I., Schäfer, P., Kurtz, J., Sappino, A.-P., and Vlastos, G. (2003). Undertreatment strongly decreases prognosis of breast cancer in elderly women. J. Clin. Oncol. *21*, 3580–3587.

Bouchardy, C., Rapiti, E., Blagojevic, S., Vlastos, A.-T., and Vlastos, G. (2007). Older female cancer patients: importance, causes, and consequences of undertreatment. J. Clin. Oncol. *25*, 1858–1869.

Bracken, A.P., and Helin, K. (2009). Polycomb group proteins: navigators of lineage pathways led astray in cancer. Nat. Rev. Cancer *9*, 773–784.

Chen, Y.-J., Hakin-Smith, V., Teo, M., Xinarianos, G.E., Jellinek, D.A., Carroll, T., McDowell, D., MacFarlane, M.R., Boet, R., Baguley, B.C., et al. (2006). Association of mutant TP53 with alternative lengthening of telomeres and favorable prognosis in glioma. Cancer Res. *66*, 6473–6476.

Chopra, R., and Raynaud, F.I. (2020). Preclinical studies to enable first in human clinical trials. In Phase I Oncology Drug Development, T.A. Yap, J. Rodon, and D.S. Hong, eds. (Springer International Publishing), pp. 45–69.

Colonna, M., Bossard, N., Remontet, L., and Grosclaude, P. (2010). Changes in the risk of death from cancer up to five years after diagnosis in elderly patients: a study of five common cancers. Int. J. Cancer *127*, 924–931.

Connolly, J.L., Schnitt, S.J., Wang, H.H., Longtine, J.A., Dvorak, A., and Dvorak, H.F. (2003). Principles of Cancer Pathology. In Holland-Frei Cancer Medicine, D.W. Kufe, R.E. Pollock, R.R. Weichselbaum, R.C. Bast, T.S. Gansler, J.F. Holland, and E. Frei, III, eds. (Hamilton, ON: BC Decker).

Conway, J.R., Warner, J.L., Rubinstein, W.S., and Miller, R.S. (2019). Next-generation sequencing and the clinical oncology workflow: data challenges, proposed solutions, and a call to action. JCO Precis. Oncol. *3*, 1–10.

Corsello, S.M., Nagari, R.T., Spangler, R.D., Rossen, J., Kocak, M., Bryan, J.G., Humeidi, R., Peck, D., Wu, X., Tang, A.A., et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nat. Cancer *1*, 235–248.

Cuzick, J., Swanson, G.P., Fisher, G., Brothman, A.R., Berney, D.M., Reid, J.E., Mesher, D., Speights, V., Stankiewicz, E., Foster, C.S., et al. (2011). Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol. *12*, 245–255.

Dale, D.C. (2003). Poor prognosis in elderly patients with cancer: the role of bias and undertreatment. J. Support Oncol. *1*, 11–17.

Dancik, G.M., and Theodorescu, D. (2015). The prognostic value of cell cycle gene expression signatures in muscle invasive, high-grade bladder cancer. Bladder Cancer *1*, 45–63.

Darvin, P., Toor, S.M., Sasidharan Nair, V., and Elkord, E. (2018). Immune checkpoint inhibitors: recent progress and potential biomarkers. Exp. Mol. Med. *50*, 1–11.

Elias, A.D. (2012). Management of small t1a/b N0 breast cancers. Am. Soc. Clin. Oncol. Educ. Book, 10–19.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N.A., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA *313*, 1122–1132.

Esserman, L.J., Thompson, I.M., and Reid, B. (2013). Overdiagnosis and overtreatment in cancer: an opportunity for improvement. JAMA *310*, 797–798.

Evans, A.J., Henry, P.C., Van der Kwast, T.H., Tkachuk, D.C., Watson, K., Lockwood, G.A., Fleshner, N.E., Cheung, C., Belanger, E.C., Amin, M.B., et al. (2008). Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens. Am. J. Surg. Pathol. *32*, 1503–1512.

Francisco, L.M., Sage, P.T., and Sharpe, A.H. (2010). The PD-1 pathway in tolerance and autoimmunity. Immunol. Rev. *236*, 219–242.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Frederick, L., Wang, X.-Y., Eley, G., and James, C.D. (2000). Diversity and frequency of epidermal growth factor receptor mutations in human glioblastomas. Cancer Res. *60*, 1383–1387.

Fukuoka, M., Wu, Y.-L., Thongprasert, S., Sunpaweravong, P., Leong, S.-S., Sriuranpong, V., Chao, T.-Y., Nakagawa, K., Chu, D.-T., Saijo, N., et al. (2011). Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non–small-cell lung cancer in asia (IPASS). J. Clin. Oncol. *29*, 2866–2874.

Gainor, J.F., Varghese, A.M., Ou, S.-H.I., Kabraji, S., Awad, M.M., Katayama, R., Pawlak, A., Mino-Kenudson, M., Yeap, B.Y., Riely, G.J., et al. (2013). ALK rearrangements are mutually exclusive with mutations in EGFR or KRAS: an analysis of 1,683 patients with non-small cell lung cancer. Clin. Cancer Res. *19*, 4273–4281.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Gentile, C., Martorana, A., Lauria, A., and Bonsignore, R. (2017). Kinase inhibitors in multitargeted cancer therapy. Curr. Med. Chem. *24*, 1671–1686.

Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat. Med. *21*, 938–945.

Gilks, C.B., Oliva, E., and Soslow, R.A. (2013). Poor interobserver reproducibility in the diagnosis of high-grade endometrial carcinoma. Am. J. Surg. Pathol. *37*, 874–881.

Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. Transl. Cancer Res. *4*, 256–269.

Green, S.D., and Konig, H. (2020). Treatment of acute myeloid leukemia in the era of genomics—achievements and persisting challenges. Front. Genet. *11*, 480.

Griffiths, D.F.R., Melia, J., McWilliam, L.J., Ball, R.Y., Grigor, K., Harnden, P., Jarmulowicz, M., Montironi, R., Moseley, R., Waller, M., et al. (2006). A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. Histopathology *48*, 655–662.

Gutteridge, R.E.A., Ndiaye, M.A., Liu, X., and Ahmad, N. (2016). Plk1 inhibitors in cancer therapy: from laboratory to clinics. Mol. Cancer Ther. *15*, 1427–1435.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2017). Exploring network structure, dynamics, and function using networkX. Proc. Python in Sci Conf. (SciPy). http://aric.hagberg.org/papers/hagberg-2008-exploring.pdf.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

Henrich, W.L., and Burkat, J.M. (2021). Patient Survival and Maintenance Dialysis. https://www.uptodate.com/contents/patient-survival-and-maintenance-dialysis.

Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., Berney, D., Scher, H., Carver, B., Scardino, P., et al. (2018). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. ELife *7*, e37294.

Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.-P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. Cell *150*, 251–263.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Jarmula, A. (2010). Antifolate inhibitors of thymidylate synthase as anticancer drugs. Mini Rev. Med. Chem. *10*, 1211–1222.

Jegerlehner, S., Bulliard, J.-L., Aujesky, D., Rodondi, N., Germann, S., Konzelmann, I., Chiolero, A., and Group, N.W. (2017). Overdiagnosis and overtreatment of thyroid cancer: a population-based temporal trend study. PLoS One *12*, e0179387.

Jordan, E.J., Kim, H.R., Arcila, M.E., Barron, D., Chakravarty, D., Gao, J., Chang, M.T., Ni, A., Kundra, R., Jonsson, P., et al. (2017). Prospective comprehensive molecular characterization of lung adenocarcinomas for efficient patient matching to approved and emerging therapies. Cancer Discov. *7*, 596–609.

Kaelin, W.G. (2017). Common pitfalls in preclinical cancer target validation. Nat. Rev. Cancer *17*, 425–440.

Kleinbaum, D.G., and Klein, M. (2012). Survival Analysis: A Self-Learning Text, Third Edition (Springer-Verlag).

Lang, H., Lindner, V., de Fromont, M., Molinié, V., Letourneux, H., Meyer, N., Martin, M., and Jacqmin, D. (2005). Multicenter determination of optimal interobserver agreement using the Fuhrman grading system for renal cell carcinoma: assessment of 241 patients with > 15-year follow-up. Cancer *103*, 625–629.

Laszlo, G.S., Estey, E.H., and Walter, R.B. (2014). The past and future of CD33 as therapeutic target in acute myeloid leukemia. Blood Rev. *28*, 143–153.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301–313.

Lee, K., Park, J.W., Lee, K., Cho, S., Kwon, Y.-H., Kim, M.J., Ryoo, S.-B., Jeong, S.-Y., and Park, K.J. (2019). Adjuvant chemotherapy does not provide survival benefits to elderly patients with stage II colon cancer. Sci. Rep. *9*, 11846.

Lin, A., and Sheltzer, J.M. (2020). Discovering and validating cancer genetic dependencies: approaches and pitfalls. Nat. Rev. Genet. *21*, 671–682.

Lin, A., Giuliano, C.J., Palladino, A., John, K.M., Abramowicz, C., Yuan, M.L., Sausville, E.L., Lukow, D.A., Liu, L., Chait, A.R., et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci. Transl. Med. *11*, eaaw8412.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell *173*, 400–416.e11.

Loeb, S., Bjurlin, M.A., Nicholson, J., Tammela, T.L., Penson, D.F., Carter, H.B., Carroll, P., and Etzioni, R. (2014). Overdiagnosis and overtreatment of prostate cancer. Eur. Urol. *65*, 1046–1055.

Looijenga, L.H., Gillis, A.J., van Gurp, R.J., Verkerk, A.J., and Oosterhuis, J.W. (1997). X inactivation in human testicular tumors. XIST expression and androgen receptor methylation status. Am. J. Pathol. *151*, 581–590.

Ludwig, J.A., and Weinstein, J.N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. Nat. Rev. Cancer *5*, 845–856.

Lukow, D.A., and Sheltzer, J.M. (2022). Chromosomal instability and aneuploidy as causes of cancer drug resistance. Trends Cancer *8*, 43–53.

Lukow, D.A., Sausville, E.L., Suri, P., Chunduri, N.K., Wieland, A., Leu, J., Smith, J.C., Girish, V., Kumar, A.A., Kendall, J., et al. (2021). Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. Dev. Cell *56*, 2427–2439.e4.

Mack, P.C., Banks, K.C., Espenschied, C.R., Burich, R.A., Zill, O.A., Lee, C.E., Riess, J.W., Mortimer, S.A., Talasaz, A., Lanman, R.B., et al. (2020). Spectrum of driver mutations and clinical impact of circulating tumor DNA analysis in

non–small cell lung cancer: analysis of over 8000 cases. Cancer *126*, 3219–3228.

Markham, A. (2019). Erdafitinib: first global approval. Drugs *79*, 1017–1021.

Marks, J.R., Davidoff, A.M., Kerns, B.J., Humphrey, P.A., Pence, J.C., Dodge, R.K., Clarke-Pearson, D.L., Iglehart, J.D., Bast, R.C., and Berchuck, A. (1991). Overexpression and mutation of p53 in epithelial ovarian cancer. Cancer Res. *51*, 2979–2984.

McShane, E., Sin, C., Zauber, H., Wells, J.N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J.A., et al. (2016). Kinetic analysis of protein stability reveals age-dependent degradation. Cell *167*, 803–815.e21.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Mosley, J.D., and Keri, R.A. (2008). Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. BMC Med. Genom. *1*, 11.

Muñoz-Maldonado, C., Zimmer, Y., and Medová, M. (2019). A comparative analysis of individual RAS mutations in cancer biology. Front. Oncol. *9*, 1088.

Ozkan, T.A., Eruyar, A.T., Cebeci, O.O., Memik, O., Ozcan, L., and Kuskonmaz, I. (2016). Interobserver variability in Gleason histological grading of prostate cancer. Scand. J. Urol. *50*, 420–424.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Peters, G.J., Backus, H.H.J., Freemantle, S., van Triest, B., Codacci-Pisanelli, G., van der Wilt, C.L., Smid, K., Lunec, J., Calvert, A.H., Marsh, S., et al. (2002). Induction of thymidylate synthase as a 5-fluorouracil resistance mechanism. Biochim. Biophys. Acta *1587*, 194–205.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198.

Reback, J., McKinney, W., jbrockmendel, J., Van den, B., Augspurger, T., Cloud, P., Gfyoung, S., Adam, K., Roeschke, M., et al. (2020). Pandas-Dev/Pandas: Pandas 1.0.3 (Zenodo). https://doi.org/10.5281/zenodo.3715232.

Reinhold, W.C., Sunshine, M., Liu, H., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. *72*, 3499–3511.

Robles, A.I., and Harris, C.C. (2010). Clinical outcomes and correlates of TP53 mutations and cancer. Cold Spring Harb. Perspect. Biol. *2*, a001016.

Rodrigues, N.R., Rowan, A., Smith, M.E., Kerr, I.B., Bodmer, W.F., Gannon, J.V., and Lane, D.P. (1990). p53 mutations in colorectal cancer. Proc. Natl. Acad. Sci. U S A *87*, 7555–7559.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychol. Bull. *86*, 638–641.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., et al. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat. Genet. *39*, 232–236.

Schmidt, M.C., Antweiler, S., Urban, N., Mueller, W., Kuklik, A., Meyer-Puttlitz, B., Wiestler, O.D., Louis, D.N., Fimmers, R., and von Deimling, A. (2002). Impact of genotype and morphology on the prognosis of glioblastoma. J. Neuropathol. Exp. Neurol. *61*, 321–328.

Schukken, K.M., and Sheltzer, J.M. (2021). Extensive protein dosage compensation in aneuploid human cancers. Preprint at bioRxiv. https://doi.org/10.1101/2021.06.18.449005.

Sheltzer, J.M. (2013). A transcriptional and metabolic signature of primary aneuploidy is present in chromosomally-unstable cancer cells and informs clinical prognosis. Cancer Res. *73*, 6401–6412.

Shields, P.G. (2000). Publication bias is a scientific problem with adverse ethical outcomes: the case for a section for null results. Cancer Epidemiol. Prev. Biomark. *9*, 771–772.

Shinawi, T., Hill, V.K., Krex, D., Schackert, G., Gentle, D., Morris, M.R., Wei, W., Cruickshank, G., Maher, E.R., and Latif, F. (2013). DNA methylation profiles of long- and short-term glioblastoma survivors. Epigenetics *8*, 149–156.

Singh, S., Numan, A., Maddiboyina, B., Arora, S., Riadi, Y., Md, S., Alhakamy, N.A., and Kesharwani, P. (2021). The emerging role of immune checkpoint inhibitors in the treatment of triple-negative breast cancer. Drug Discov. Today *26*, 1721–1727.

Smith, J.C., and Sheltzer, J.M. (2018). Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. ELife *7*, e39217.

Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., Geyer, C.E., Dees, E.C., Goetz, M.P., Olson, J.A., et al. (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N. Engl. J. Med. *379*, 111–121.

Srigley, J.R., Delahunt, B., Samaratunga, H., Billis, A., Cheng, L., Clouston, D., Evans, A., Furusato, B., Kench, J., Leite, K., et al. (2019). Controversial issues in Gleason and International Society of Urological Pathology (ISUP) prostate cancer grading: proposed recommendations for international implementation. Pathology (Phila.) *51*, 463–473.

Stark, J.R., Perner, S., Stampfer, M.J., Sinnott, J.A., Finn, S., Eisenstein, A.S., Ma, J., Fiorentino, M., Kurth, T., Loda, M., et al. (2009). Gleason score and lethal prostate cancer: does 3 + 4 = 4 + 3? J. Clin. Oncol. *27*, 3459–3464.

Stopsack, K.H., Whittaker, C.A., Gerke, T.A., Loda, M., Kantoff, P.W., Mucci, L.A., and Amon, A. (2019). Aneuploidy drives lethal progression in prostate cancer. Proc. Natl. Acad. Sci. U S A *116*, 11390–11395.

Stouffer, S.A. (1949). The American Soldier (Princeton University Press).

Sun, L., Zhang, L., Yu, J., Zhang, Y., Pang, X., Ma, C., Shen, M., Ruan, S., Wasan, H.S., and Qiu, S. (2020). Clinical efficacy and safety of anti-PD-1/PD-L1 inhibitors for the treatment of advanced or metastatic cancer: a systematic review and meta-analysis. Sci. Rep. *10*, 2083.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. *45*, W98–W102.

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. *47*, W556–W560.

The Cancer Genome Atlas Research Network; Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Therneau, T.M. (2021). Survival Analysis [R Package Survival Version 3.2-11] (Comprehensive R Archive Network (CRAN)). https://cran.r-project.org/web/packages/survival/index.html.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*, eaan2507.

Unni, A.M., Lockwood, W.W., Zejnullahu, K., Lee-Lin, S.-Q., and Varmus, H. (2015). Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. ELife *4*, e06907.

Vasudevan, A., Baruah, P.S., Smith, J.C., Wang, Z., Sayles, N.M., Andrews, P., Kendall, J., Leu, J., Chunduri, N.K., Levy, D., et al. (2020). Single-chromosomal gains can function as metastasis suppressors and promoters in colon cancer. Dev. Cell *52*, 413–428.e6.

Vasudevan, A., Schukken, K.M., Sausville, E.L., Girish, V., Adebambo, O.A., and Sheltzer, J.M. (2021). Aneuploidy as a promoter and suppressor of malignant growth. Nat. Rev. Cancer *21*, 89–103.

Venet, D., Dumont, J.E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput. Biol *7*, e1002240.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Whitfield, M.L., George, L.K., Grant, G.D., and Perou, C.M. (2006). Common markers of proliferation. Nat. Rev. Cancer *6*, 99–106.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

Young, R.C. (2003). Early-stage ovarian cancer: to treat or not to treat. JNCI J. Natl. Cancer Inst. *95*, 94–95.

Zaniboni, A., and Labianca, R. (2004). Adjuvant therapy for stage II colon cancer: an elephant in the living room? Ann. Oncol. *15*, 1310–1318.

Zeeberg, B.R., Riss, J., Kane, D.W., Bussey, K.J., Uchio, E., Linehan, W.M., Barrett, J.C., and Weinstein, J.N. (2004). Mistaken Identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinf. *5*, 80.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713.

Zhao, R., Choi, B.Y., Lee, M.-H., Bode, A.M., and Dong, Z. (2016). Implications of genetic and epigenetic alterations of CDKN2A (p16(INK4a)) in cancer. EBioMedicine *8*, 30–39.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA: Copy Number	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg
TCGA: Methylation	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	usc.edu_PANCAN_merged_HumanMethylation27_HumanMethylation450.betaValue_whitelisted.tsv
TCGA: Mutation	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	mc3.v0.2.8.PUBLIC.maf.gz
TCGA: Gene expression	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv
TCGA: miRNA expression	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv
TCGA: Protein expression	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	TCGA-RPPA-pancan-clean.txt
TCGA: Clinical data	PanCanAtlas: https://gdc.cancer.gov/about-data/publications/pancanatlas	TCGA-CDR-SupplementalTableS1.xlsx
Software and algorithms		
GraphPad Prism	https://www.graphpad.com/	RRID:SCR_002798
NetworkX	https://networkx.org/	RRID:SCR_016864
STRING DB	https://string-db.org/	RRID:SCR_005223
SciPy	https://scipy.org/	RRID:SCR_008058
Pandas	https://pandas.pydata.org/	V1.2.4
Other		
Recurrently-mutated driver genes	Bailey et al., Cell 2018	Table S1
FDA-approved cancer therapies	National Cancer Institute	https://www.cancer.gov/about-cancer/treatment/drugs
NCI60 cell line doubling times	CellMiner	https://discover.nci.nih.gov/cellminer/celllineMetadata.do
Suz12 binding sites	Lee et al., Cell 2006	Table S9
Cancer gene dependency scores	http://www.depmap.org	DepMap Public 21Q2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jason Sheltzer (jason.sheltzer@yale.edu).

Materials availability

This study did not generate new unique reagents. The existing datasets that were analyzed and the code used to analyze them are described in "Data and code availability" below.

Data and code availability

Section 1: Data: No new unique datasets were generated for this study. TCGA data was acquired from the TCGA PanCanAtlas, available at https://gdc.cancer.gov/about-data/publications/pancanatlas. Final datasets used for this analysis include:

DNA copy number: *broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg*

DNA methylation: *usc.edu_PANCAN_merged_HumanMethylation27_HumanMethylation450. betaValue_whitelisted.tsv*
Gene expression: *EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv*
miRNA expression: *pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv*
Mutations: *mc3.v0.2.8.PUBLIC.maf.gz*
Protein expression: *TCGA-RPPA-pancan-clean.txt*

TCGA patient survival data and final clinical annotations were acquired from (Liu et al., 2018). Selection of the clinical endpoint for each cancer type was based on the recommendations provided by (Liu et al., 2018) based on data quality, cohort size, and the number of events that were observed.

Pan-cancer oncogenes, oncogene-cancer type pairs, and recurrently-observed point mutations were acquired from (Bailey et al., 2018). Tumor mitotic activity and prostate cancer Gleason scores were acquired from the provisional TCGA annotations available at http://www.cbioportal.org (Gao et al., 2013). NCI-SEER survival statistics were acquired from https://seer.cancer.gov/statistics/.

The list of FDA-approved cancer therapies was acquired from the NCI Drug Dictionary, available at https://www.cancer.gov/about-cancer/treatment/drugs. Drug targets were identified from the NCI Drug Dictionary, from (Corsello et al., 2020), and from (Wishart et al., 2018). Multi-targeted kinase inhibitors (sorafenib, sunitinib, etc.) and other drugs where the mechanism-of-action is unclear were excluded from this analysis (Gentile et al., 2017; Lin and Sheltzer, 2020; Lin et al., 2019). FDA approval dates were acquired from https://www.fda.gov/drugs/resources-information-approved-drugs/hematologyoncology-cancer-approvals-safety-notifications and from https://www.drugs.com/history/. Cancer dependency scores were acquired from http://www.depmap.org (Meyers et al., 2017).

Doubling times for the NCI-60 cell line panel were acquired from (Reinhold et al., 2012). Suz12 binding sites in embryonic stem cells were acquired from (Lee et al., 2006).

Section 2: Code: The code used to perform the analysis in this paper is available at github.com/joan-smith/comprehensive-tcga-survival. An automatic download script has been provided that will set up the correct directory structure for running the complete suite of analyses. This download script downloads all relevant data from the PanCan Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas), and supporting files from a number of other sources. To fully regenerate the analysis performed in this paper, run the download script (download_data.py) and the main analysis script (main.py). Detailed instructions for reproducing the results are provided in the github repository.

Section 3:Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Survival analysis in TCGA cohorts

The TCGA project was initiated to facilitate the molecular characterization of the major cancer types found in the US. While clinical and pathological data were collected for each patient, genomic analysis was prioritized over clinical follow-up. As justified by the analyses described in this manuscript and in other publications, we posit that performing survival analysis on the TCGA cohorts remains appropriate for several reasons. First, as described in Liu et al., the TCGA clinical data has been rigorously reviewed, harmonized, and validated through independent analyses (Liu et al., 2018). Liu et al. also reported that, as expected, stage III/IV tumors in TCGA had uniformly worse outcomes compared to stage I/II tumors, and the median survival times for certain cancers fall within established ranges based on published case series. Secondly, in this work, we demonstrate that high-grade tumors have worse outcomes than low-grade tumors, that older patients have worse outcomes than younger patients, and that the survival times within TCGA cancer types are highly-correlated with the survival times reported in the nationally-representative SEER database (Figure S1). Thirdly, our analysis has verified the prognostic power of several established molecular biomarkers, including the adverse effects of p53 mutations and the strong association between cell cycle gene expression and aggressive disease (Baak et al., 2009; Cuzick et al., 2011; Mosley and Keri, 2008; Robles and Harris, 2010; Venet et al., 2011). Finally, we and others have verified that the frequencies of mutations in specific oncogenes and tumor suppressors in TCGA are very close to the frequencies observed in other clinical series, further demonstrating that the TCGA cohorts are broadly representative of cancer patients as a whole (Amar et al., 2017; Jordan et al., 2017; Smith and Sheltzer, 2018; Zehir et al., 2017). Thus, while facilitating prognostic biomarker discovery was not the major goal of the TCGA project, we believe that the TCGA populations are representative cohorts and that the survival analysis we have conducted is appropriate.

For all cancer types except LAML and SKCM, only primary tumor specimens (TCGA code: 01) were analyzed. For the analysis of SKCM data, both primary and metastatic samples (TCGA code: 06) were analyzed. In the event that both a primary specimen and a metastatic specimen were available for SKCM, only the primary specimen was analyzed. For the analysis of LAML data, blood cancer specimens were given the TCGA code "03" and cancers with this code were included. For the analysis of XIST expression, RPS4Y1 expression and patient sex, the TCGT cohort was excluded due to the reactivation of XIST that has been previously reported in testicular cancers (Looijenga et al., 1997).

Selection of analysis methodology

Several statistical techniques have been developed to perform survival analysis (Kleinbaum and Klein, 2012). In this paper, we chose to apply Cox proportional hazards regression to study the TCGA cohorts. The Cox model is given by the following function:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{n} \beta_i X_i}$$

Where t is the survival time, h (t, X) is the hazard function, $h_0(t)$ is the baseline hazard, X_i is a potential prognostic variable, and β_i indicates the strength of the association between a prognostic variable and survival. In this model, patients have a baseline, time-dependent risk of death $h_0(t)$, modified by time-independent prognostic features that either increase ($\beta_i > 0$) or decrease ($\beta_i < 0$) risk of death. In this work, we report Z scores from these Cox models, which are calculated by dividing the regression coefficient (β_i) by its standard error.

As we have previously described (Smith and Sheltzer, 2018), we utilize Cox proportional hazards modeling for survival analysis for several reasons. First, unlike Kaplan-Meier analysis, Cox models do not require the selection of threshold values, so continuous data like gene expression measurements do not need to be dichotomized. Secondly, Cox models can accept both continuous and discrete input data, allowing this approach to be used to analyze both binary (e.g., mutant vs. non-mutant) and continuous (e.g., gene expression) genomic features. Thirdly, Cox models allow the use of right-censored survival data, in which some patients are lost to follow-up or do not experience a relevant clinical event within the time frame of the study. Right-censored clinical data is appropriate for real-world analyses in which patients with indolent cancers may live for decades without disease recurrence and following all patients until their deaths is not feasible. Fourthly, Cox models can be used to perform both univariate (i = 1) and multivariate (i > 1) analyses. Fifthly, Cox regression allows us to calculate a *Z* score and a p value for each association, as Z scores represent the number of standard deviations from the mean of a normal distribution. Previous qq-analysis has demonstrated the underlying normality of the survival data (Smith and Sheltzer, 2018). Sixthly, Z scores encode the directionality of an association: poor prognostic factors will exhibit βi values greater than 0, while favorable prognostic factors will exhibit βi values less than 0. This allows "favorable" and "adverse" survival features to be directly compared. Seventhly, Z scores are useful for meta-analyses, as they can be combined using Stouffer's Method (Stouffer, 1949):

$$Z = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{k}}$$

Eighthly, Cox proportional hazards modeling is commonly used in both previous genome-wide survival analyses and in numerous clinical biomarkers studies, facilitating comparison with other biomarker discovery efforts (Anaya et al., 2016; Fukuoka et al., 2011; Gentles et al., 2015; Hieronymus et al., 2018; Parker et al., 2009; Smith and Sheltzer, 2018).

Overall analysis strategy

In this paper, we describe the comprehensive and unbiased generation of Cox proportional hazard model Z scores for every genomic feature available in TCGA (CNAs, methylation, mutation, gene expression, miRNA expression, and protein expression), and for every cancer type. All analysis was performed in Python, using pandas (Reback et al., 2020), matplotlib (Hunter, 2007), numpy (Harris et al., 2020) and scipy (Virtanen et al., 2020). Cox proportional hazards were computed using the R survival package (Therneau, 2021), and rpy2 was used to integrate the R computations with Python scripts.

The software for these analyses is structured to be repeatable, modular, and debuggable. The same code was used for computing all Z scores, with swappable functions for preparing and cleaning each data type. Similarly, all clinical data was prepared identically across all analyses, using a custom-built parser for the clinical data provided by (Liu et al., 2018).

TCGA copy number data was generated as relative copy number values for particular chromosomal intervals. This data was translated to produce a single copy number value on a per-gene basis, based on the observed copy number at each gene's transcription start site. This annotation was performed using mapping data from GENCODE v32 (Frankish et al., 2019). Interval trees, from the Python package intervaltree, were used to facilitate efficient mapping of segmental data to the appropriate genes (https://github.com/chaimleib/intervaltree). The copy number value at a gene's transcription start site was used as the input for the Cox models. Note that Cox proportional hazards modeling is threshold-independent, and so no minimum or maximum copy number value was required to specify a deletion or an amplification.

For protein expression data, the normalized and batch-corrected RPPA expression values were used as inputs for the Cox models. For the gene expression and microRNA expression data, values were log2-transformed and clipped at 0, then used as inputs for the Cox models.

For the methylation data, each probe was mapped to the relevant gene(s) that it recognized using the probeset annotations provided by Illumina. Beta values that mapped to the same gene were collapsed by averaging. This single average Beta value was used as input for the Cox models.

For CNAs, methylation, gene expression, miRNA expression, and protein expression, Cox models were only generated if at least 10 patients in a cohort had data for a particular feature.

For mutations, Cox models were generated if 2% or more of all sequenced patients for a particular cancer type had a non-synonymous mutation in the relevant gene. This 2% cut-off is based on qq tests for normality that we have previously conducted (Smith and Sheltzer, 2018). Non-synonymous mutations included: missense, nonsense, frameshift deletion, splice site, frameshift insertion, in-frame deletion, translation start site, nonstop mutation, and in-frame insertion. For each gene in each patient, a gene was considered to be mutated if there was a single of these non-synonymous mutations at any codon within the gene. In the driver gene analysis, a patient was marked as mutated if there was a single non-synonymous mutation at any of the relevant codons.

Multivariate analysis was performed using age, sex, stage, and grade data from (Liu et al., 2018) The variables used for each cohort are listed in Table S2G. The divisions for ''stage'' as a variable are listed in Table S2H. The divisions for ''grade'' as a variable are listed in Table S2I.

Single apostrophes were prepended to gene names in the output files from these analyses in order to allow the data tables to be read in Microsoft Excel without auto-formatting (Zeeberg et al., 2004).

Kaplan-Meier analysis

Kaplan-Meier plots were generated using GraphPad Prism. Gene expression, miRNA expression, protein expression, and methylation values were dichotomized based on their mean values within the indicated cohorts. For copy number analysis, CNA values > 0.3 were classified as amplified and CNA values < -0.3 were classified as deleted (Smith and Sheltzer, 2018).

Gene ontology analysis

Gene ontology and transcription factor enrichment analysis were performed using g:Profiler (Raudvere et al., 2019). Genes used to calculate a cell cycle gene score and transcription factor methylation score were also identified using the appropriate GO term via g:Profiler.

Additional tools and resources

Gene set permutations were performed in Python by sampling 1000 random permutations of column data, in which observed Z scores were randomly assigned to gene or feature labels within each cancer type.

Peak finding was performed using the standard scipy signals library (Virtanen et al., 2020). Gene network analysis for mitotic genes and developmental transcription factors was performed using NetworkX and STRING (Szklarczyk et al., 2019).

QUANTIFICATION AND STATISTICAL ANALYSIS

Sample sizes for this work were based on the number of patients with each cancer type that were analyzed on each genomic platform (Table 1). No patients were excluded from analysis.

For Cox proportional hazards modeling, a Z score greater than 1.96 or less than -1.96 were considered significant. For Kaplan-Meier plots, p values were determined using a log rank test. Additional statistical tests are described in the figure legends, including Student's t-tests (Figures 2D–2J) and hypergeometric tests (Figures S5E and S6E).

ADDITIONAL RESOURCES

A website facilitating access to the results of this analysis is available at http://www.tcga-survival.com.