



Article

Characterization of the Survival Influential Genes in Carcinogenesis

Divya Sahu [†], Yu-Lin Chang [†], Yin-Chen Lin and Chen-Ching Lin ^{*}

Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei 11221, Taiwan; sahu.divya786@gmail.com (D.S.); skyblue8863@gmail.com (Y.-L.C.); ying87000@gmail.com (Y.-C.L.)

^{*} Correspondence: ChaosLin@ym.edu.tw or chenching.lin@nycu.edu.tw

[†] Authors contributed equally.

Abstract: The genes influencing cancer patient mortality have been studied by survival analysis for many years. However, most studies utilized them only to support their findings associated with patient prognosis: their roles in carcinogenesis have not yet been revealed. Herein, we applied an in silico approach, integrating the Cox regression model with effect size estimated by the Monte Carlo algorithm, to screen survival-influential genes in more than 6000 tumor samples across 16 cancer types. We observed that the survival-influential genes had cancer-dependent properties. Moreover, the functional modules formed by the harmful genes were consistently associated with cell cycle in 12 out of the 16 cancer types and pan-cancer, showing that dysregulation of the cell cycle could harm patient prognosis in cancer. The functional modules formed by the protective genes are more diverse in cancers; the most prevalent functions are relevant for immune response, implying that patients with different cancer types might develop different mechanisms against carcinogenesis. We also identified a harmful set of 10 genes, with potential as prognostic biomarkers in pan-cancer. Briefly, our results demonstrated that the survival-influential genes could reveal underlying mechanisms in carcinogenesis and might provide clues for developing therapeutic targets for cancers.



Citation: Sahu, D.; Chang, Y.-L.; Lin, Y.-C.; Lin, C.-C. Characterization of the Survival Influential Genes in Carcinogenesis. *Int. J. Mol. Sci.* **2021**, *22*, 4384. <https://doi.org/10.3390/ijms22094384>

Academic Editor: Takaaki Masuda

Received: 30 March 2021

Accepted: 20 April 2021

Published: 22 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: survival influential genes; pan-cancer; prognostic biomarkers

1. Introduction

Genes with an impact on the survival of tumor cells [1,2] likely influence the survival of cancer patients. So far, cancer-essential genes have been discovered by the clustered regularly interspaced short palindromic repeats (CRISPR) method [3] and the cancer-dependent genes with the Cancer Dependency Map (DepMap) [4]. The discovered cancer-essential genes could facilitate the development of promising cancer therapies [5] and carcinogenesis mechanisms. Although experimental studies hold promise for the accurate detection of essential genes, they are capital and labor intensive, and time consuming. Therefore, several computational approaches have been implemented to predict essential genes in *Saccharomyces cerevisiae*, *Escherichia coli* [6], and humans [7,8]. However, the identification of survival-influential genes (SIGs) in cancer patients and the investigation of their role in carcinogenesis are largely unexplored.

Survival analysis is a branch of statistics that analyzes data, where the outcome variable is the time until one or more events occur, e.g., death [9]. That is, the methods of survival analysis measure the proportion of a population that will survive to passing a certain time point, i.e., the survival rate of the population. Similarly, the experimental approaches identifying essential genes measure the growth curves of the studied organism population, such as yeast or *E. coli* [10–14]. However, it is impossible to recruit humans as experimental targets with the purpose of screening essential genes or synthetic lethal gene pairs in this way. Recently, benefiting from the large projects collecting information on cancer patients' genomics and transcriptomics data, such as TCGA (The Cancer Genome Atlas), many patients' clinical information, including survival time, was also gathered.

These sets of clinical data make survival analysis a promising strategy for recovering SIGs in cancer patients.

To identify and analyze the SIGs in cancer patients, we designed an *in silico* approach, incorporating effect size derived from the Monte Carlo algorithm into the Cox regression model, to assess the influence of gene expression on patient survival in cancer. We demonstrated that the pan-cancer SIGs identified by our approach are significantly associated with known oncogenic genes, and the SIGs identified in the individual cancer types displayed cancer-dependent properties. Furthermore, we found that the harmful SIGs might be involved in cell proliferation to promote carcinogenesis. However, the diverse functions of protective SIGs across cancer types demonstrate that the strategies against carcinogenesis developed by patients with different cancer types might be variable. Here, we present list of SIGs which could reveal the underlying mechanisms in carcinogenesis and propose a set of prognostic biomarkers that might be potential therapeutic targets for cancer.

2. Results

2.1. Overview of the Identified Survival Influential Genes in Cancers

In this study, we proposed a cancer-dependent approach that incorporated the Monte Carlo algorithm into Cox regression to identify the SIGs in cancers. Briefly, this approach created the simulated gene expression profiles of cancers through the Monte Carlo processes. Then, by comparing with the simulated expression profiles, this approach can estimate the effect size of the survival influence for genes in a certain cancer type. We then calculated and investigated the effect sizes for all genes in every cancer separately, to systematically define the proper threshold to identify SIGs for each cancer. We observed that as the p -value increased, the effect size increased at first, and then decreased afterward (Figure 1A,B, left panel). This result indicated the existence of a global maximal effect size. Furthermore, we observed that the p -values reaching the maximal effect size were distinct among cancer types (Figure 1A,B, right panel). This observation confirmed that a distinct threshold of significance level is necessary for different cancer types when performing a survival analysis [15,16]. Accordingly, in different cancer types, we used a p -value reaching the maximum effect size as the threshold to identify the SIGs (Figure 1B, right panel). The identified SIGs are listed in Supplementary Dataset S1. We observed that, among the 16 cancer types, eight cancer types included more harmful genes and the others included more protective genes (Figure 1C). Additionally, the proportion of SIGs ranged from 5% to 11%, except for low grade glioma (LGG) and kidney renal clear cell carcinoma (KIRC) with 31% and 20%, respectively (Figure 1C). The high proportion of SIGs in LGG and KIRC implied that the mortality of the patients with LGG or KIRC might be more sensitive to aberrant gene expression. Similarly, a previous study reported a significant association of ribonucleoproteins with poor patient survival in KIRC, LGG, and kidney renal papillary cell carcinoma (KIRP) [17]. Additionally, we performed the proposed approach to detect SIGs in pan-cancer. We finally identified 2450 pan-cancer SIGs: 1196 harmful and 1254 protective genes.

Next, to assess the association between the identified SIGs and cancer, we compiled a cancer-associated gene list from three datasets. Since the data on tumor suppressor genes are limited, this cancer-associated gene list mainly consisted of oncogenes, i.e., cancer essential or cancer dependent genes. In other words, they are potentially harmful to patient survival. On the other hand, the three datasets collected these cancer-associated genes from various cancer types. Therefore, we used them to evaluate the cancer association of the pan-cancer harmful SIGs. Additionally, we used these cancer-associated genes to assess the performance of our approach in the identification of SIGs. We also compared our approach with the conventional method with two fixed thresholds, which were p -value < 0.05 and < 0.01 , and the Cox regression model with the least absolute shrinkage and selection operator (LASSO). We observed that our approach identified with a higher precision the proportion of cancer-associated genes than these three approaches (Figure 1D). Furthermore, compared with the small but highly confident harmful gene set collected from the

literature, the precision of our approach was lower but comparable (Figure 1D). It is worth noting that the compiled cancer-associated genes were significantly enriched in the identified pan-cancer harmful genes, but not significantly enriched in the literature-curated gene set (Figure 1D). This observation might suggest that our pan-cancer harmful genes are more associated with the compiled cancer-associated genes than the literature-curated ones. Briefly, these comparisons further demonstrated the better performance of our approach, and might show the confidence of the identified SIGs.

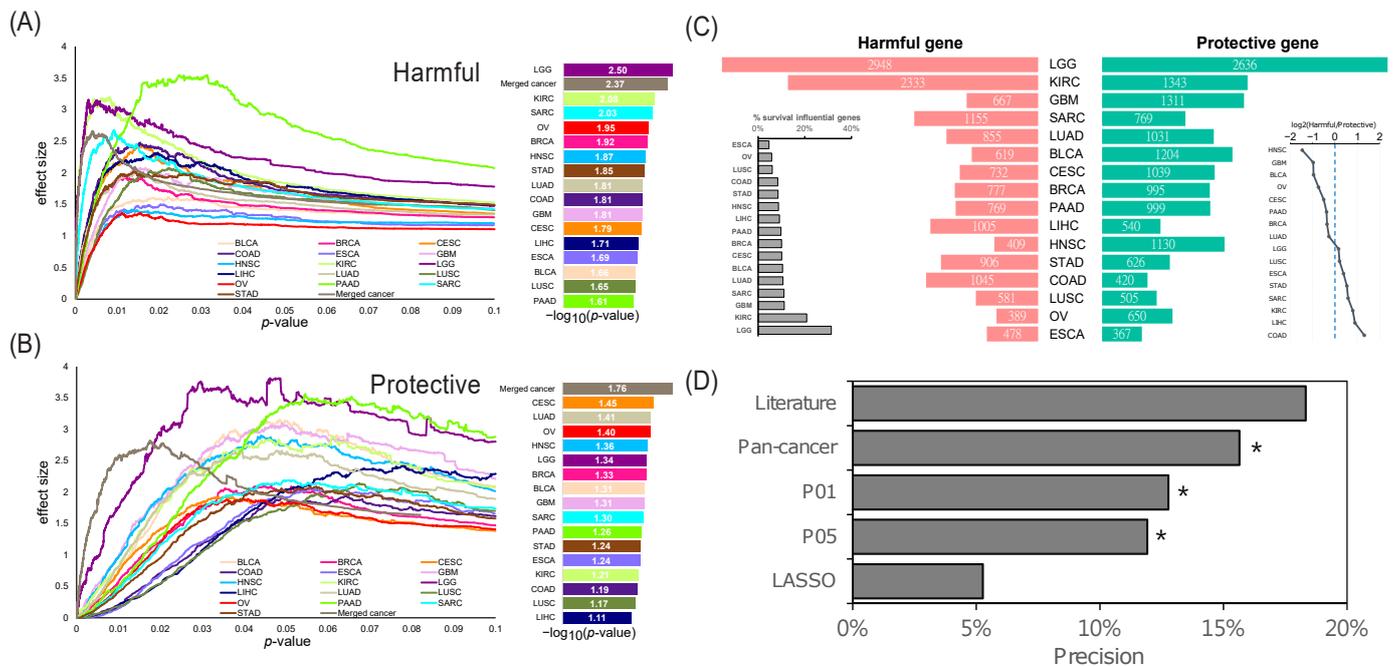


Figure 1. Summary of the survival-influential genes in cancers. (A,B) The effect sizes of the different threshold p -values for each cancer type. The effect sizes of p -values were estimated via 1000-time permutation test. The p -value producing the largest effect size (right panel) was denoted as the cut-off of significance level for identifying survival influential genes. (C) The number of identified survival influential genes in each cancer type. Red and green bars represent harmful and protective genes respectively. The proportion of survival influential genes and the ratio of harmful to protective genes in each cancer genome are also displayed. (D) Cancer association of the SIGs. Precision is the proportion of cancer-associated genes in the SIGs identified from LASSO (Cox regression model with LASSO regularization), P05 and P01 (Cox regression model with the threshold of p -value < 0.05 and 0.01), pan-cancer, and the literature. An asterisk signifies a significantly large proportion of cancer-associated genes (p -value < 0.05, Fisher’s exact test) in the corresponding category.

2.2. Exclusivity of the SIGs and Identification of the Pan-Cancer SIGs

To understand the characteristics of SIGs, we first investigated their prevalence across cancers. We observed that 5684 genes were identified as survival influential in only one cancer type (Figure 2A, unique region), and 9327 genes in at least two cancer types (Figure 2A, shared region). Moreover, around 90% of harmful and protective genes are in less than two cancer types, and no genes are survival influential in more than eight cancer types (Figure S1). This observation indicates that the SIGs might be exclusive between cancers. However, a too strict cutoff for identifying the SIGs might cause exclusivity. To examine this scenario, we removed the condition of significance level and only used the hazard ratio to determine the survival risk of the tested genes. That is, we denoted one gene with a hazard ratio >1 and ≤ 1 as harmful and protective, respectively. Accordingly, each gene was classified as either harmful or protective. Interestingly, we observed that only six genes were identified as harmful in all sixteen cancer types and no genes were protective in all 16 cancer types. Moreover, only 4.66% and 4.55% of genes were identified as harmful and protective, respectively, in more than 13 cancer types (Figure 2B); and 45.03% of genes had a mixed influence (harmful vs. protective: 8:8, 7:9, or 9:7) to patient survival across

the 16 cancer types. In other words, even though we removed the significance cutoff of SIGs, the survival risk for a gene was still barely consistent across cancers. Accordingly, we concluded that the exclusivity of SIGs across cancers was not biased by the cutoff significance. Furthermore, these above results also demonstrated that our approach might be able to identify the particular SIGs in each cancer type.

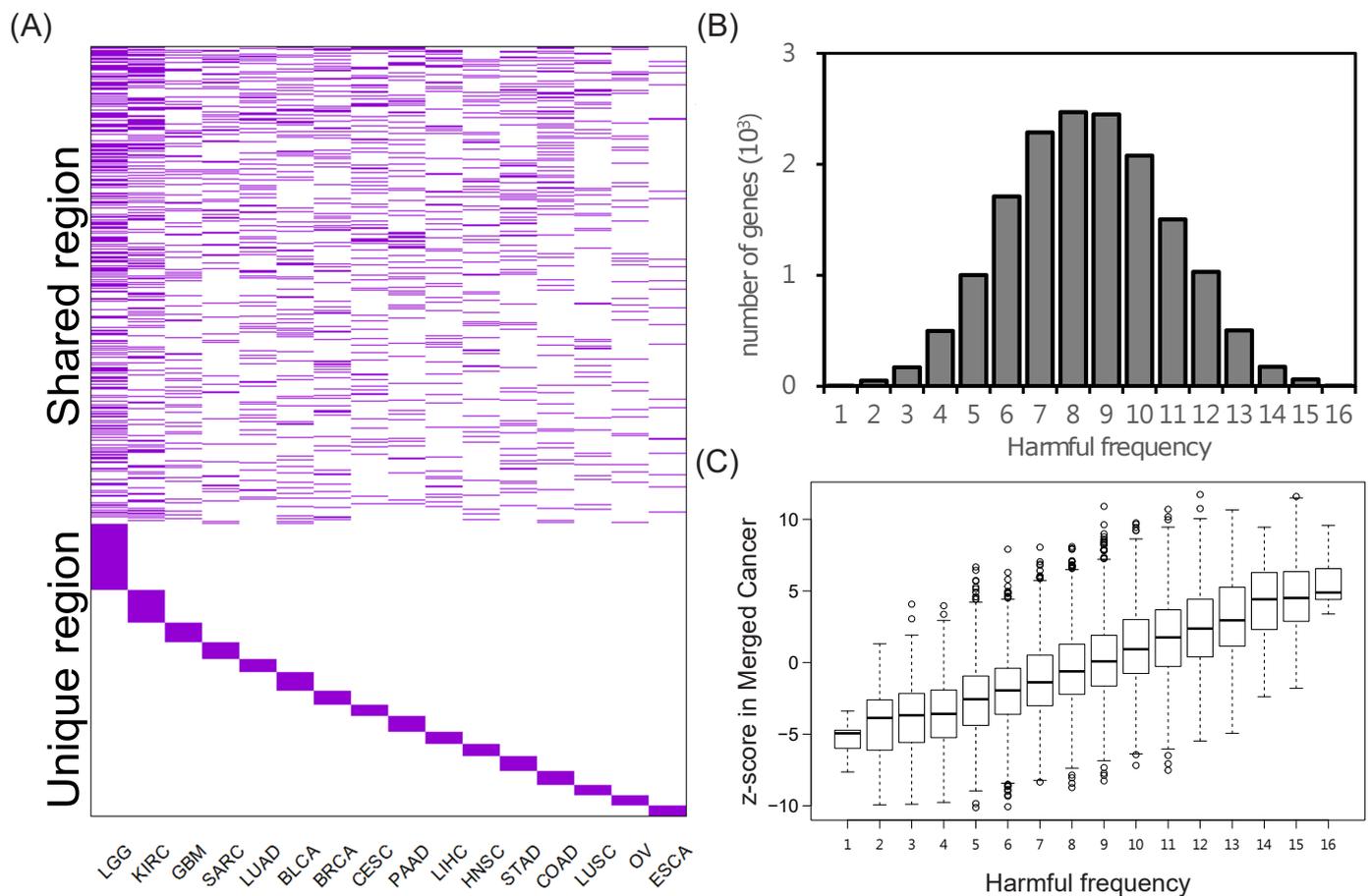


Figure 2. Exclusiveness and generality of SIGs. (A) Presence matrix of the survival influential genes across cancers. Each column indicates cancer type and each row the union of both harmful genes and protective genes of 16 cancer types. (B) The distribution frequency in which the genes were identified as harmful across cancers. (C) The association between the z-score of genes in the merged cancer data set and the frequency with which the genes were identified as harmful across cancers.

On the other hand, the above observations unveiled that the genes identified as harmful or protective in a large number of cancer types might possess the potential to be pan-cancer SIGs. Indeed, genes with a more consistent pattern of survival influence across multiple cancers tended to possess a higher survival influence (z-score) in the merged cancer data set (Figure 2C). In other words, the identified pan-cancer SIGs tended to keep a consistent pattern of survival influence across cancers. However, the pan-cancer harmful and protective genes were not observed to affect patient survival significantly in the individual cancer types, though they possessed stronger effects than other non-SIGs (also excluding pan-cancer SIGs) in most cancer types (Figures S2 and S3). This observation showed that the survival influence of pan-cancer harmful and protective SIGs was amplified and recognized in the merged cancer data set, but was easy to ignore when they were considered for each cancer type. More importantly, without using the merged cancer data set, these pan-cancer SIGs could be omitted.

2.3. Analysis of SIG Roles in the Human Co-Expressed Protein Interaction Network

Previous studies have demonstrated that the investigation of genes/proteins in the biological network is powerful for elucidating their roles in the molecular mechanisms during cancer development [18,19]. We investigated the properties of SIGs in the biological system through studying their interactions in the human protein interaction network (PIN), which was obtained from InBio Map [20]. Notably, the PIN used here was static and therefore might not provide conditional information, e.g., cancer-activated protein-protein interactions (PPIs). We further utilized the co-expressed PPIs (CePPIs) to extract the context-dependent PPIs in each cancer type. In this study, we denoted a PPI formed by two genes significantly co-expressed with each other (z-score of Spearman correlation coefficient ≥ 2) as a CePPI. We observed that the PPIs within harmful or protective SIGs showed a significantly higher coherent expression pattern than other PPIs in the PIN (Figure 3A and Figure S4), demonstrating the stronger functional association among SIGs with similar survival influence. On the contrary, PPIs between harmful and protective genes tended to exhibit a negative correlation of expression pattern (Figure 3A and Figure S4). This observation might suggest an antagonistic effect between harmful and protective genes in cancer patients.

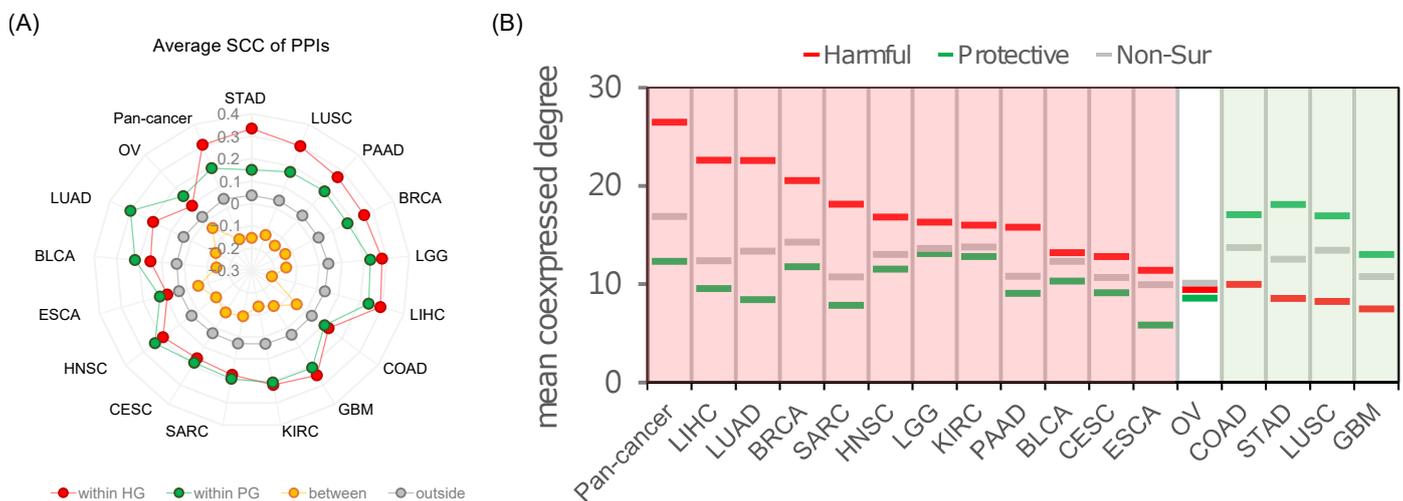


Figure 3. Co-expressed protein-protein interaction among SIGs. **(A)** The average Spearman correlation coefficient (SCC) between SIGs in cancers. The average SCCs were calculated between two genes forming PPI only. The red and green circles represent the mean SCCs within harmful and protective SIGs, respectively; yellow circles are the mean SCCs between harmful and protective genes, between SIGs with different types; grey circles are SCCs between non-SIGs. **(B)** Comparison of the co-expressed degree between harmful and protective SIGs and non-SIGs. The co-expressed degree is the number of significantly co-expressed (z-score of SCC ≥ 2) PPI partners of one gene.

The proteins encoded by cancer genes have been observed to occupy more pivotal positions, e.g., hubs, bottlenecks, or the center, than other proteins in the human interactome [18]. Indeed, we found that in ten cancer types the SIGs, either harmful or protective, had a significantly higher degree of co-expression than the non-SIGs (Figure 3B and Table S1), that is, they tended to be hubs in the cancer CePINs. Additionally, we found that, in the corresponding cancer type, except for OV, when one type of SIG showed a higher co-expression degree than non-SIGs, the other type showed a lower (Figure 3B). This competition of hub positions further supported the antagonistic effect between harmful and protective genes in cancer patients. Moreover, in the static human PIN, which does not filter out non-co-expressed PPIs, we did not observe a significant difference of degree between the SIGs and non-SIGs in most cancer types (Tables S2 and S3). Therefore, the SIGs identified in the individual cancer type might be cancer-specific instead of general

hubs. These results also confirmed that our approach could identify the context-sensitive SIGs, and cancer-specific, for cancers.

2.4. The Survival Influential Functional Modules

Next, we performed a network-based functional enrichment analysis to identify the survival-influential functional modules, which are the protein interacting functional modules formed by the coherently expressed SIGs in the corresponding cancer type. However, we did not identify harmful modules in ESCA, HNSC, and OV or protective modules in ESCA and PAAD under the predefined criterion. Through this analysis, we could pinpoint the molecular mechanisms by which the SIGs are involved in carcinogenesis affect patient prognosis. In the harmful SIGs, we observed that the functional modules associated with cell cycle were consistent across six cancer types: BRCA, KIRC, LGG, LIHC, LUAD, and PAAD (Figure 4). Interestingly, except for CESC, other cell cycle-related modules also appeared in the remaining cancer types: positive regulation of cell proliferation in COAD and GBM; positive regulation of cell differentiation in BLCA, LUSC, and STAD; and G1/S transition of mitotic cell cycle in SARC. Briefly, twelve cancer types presented the harmful modules involved in cell cycle, implying that the dysregulation of cell cycle could be harmful to patients across cancers [21,22]. Indeed, we found that, in the merged cancer dataset, the harmful SIGs were also significantly overrepresented in the functional modules associated with cell cycle, such as “mitotic nuclear division”, “cell cycle phase transition”, and “mitotic cell cycle process” (Figure 4). This observation also demonstrates that the harmful SIGs between different cancers might be exclusive, but the biological processes of harmful SIGs (regulation of cell cycle) might demonstrate the generality of harmful SIGs’ survival characteristics across cancers. However, the functional modules from the protective SIGs were more diverse than the harmful modules. The most prevalent modules, which were discovered in CESC, HNSC, LUAD, and SARC, are relevant for immune response (Figure 5). The activation of immune response-related functions is known to improve patient prognosis [23,24]. The functional modules of protective SIGs identified in the merged cancer dataset are involved in the regulation of gene expression and protein trafficking (Figure 5). This is inconsistent with the modules identified from the individual cancer types. This observation may confirm the exclusivity of protective SIGs between cancers, and imply that patients with different cancer types might develop different mechanisms against carcinogenesis.

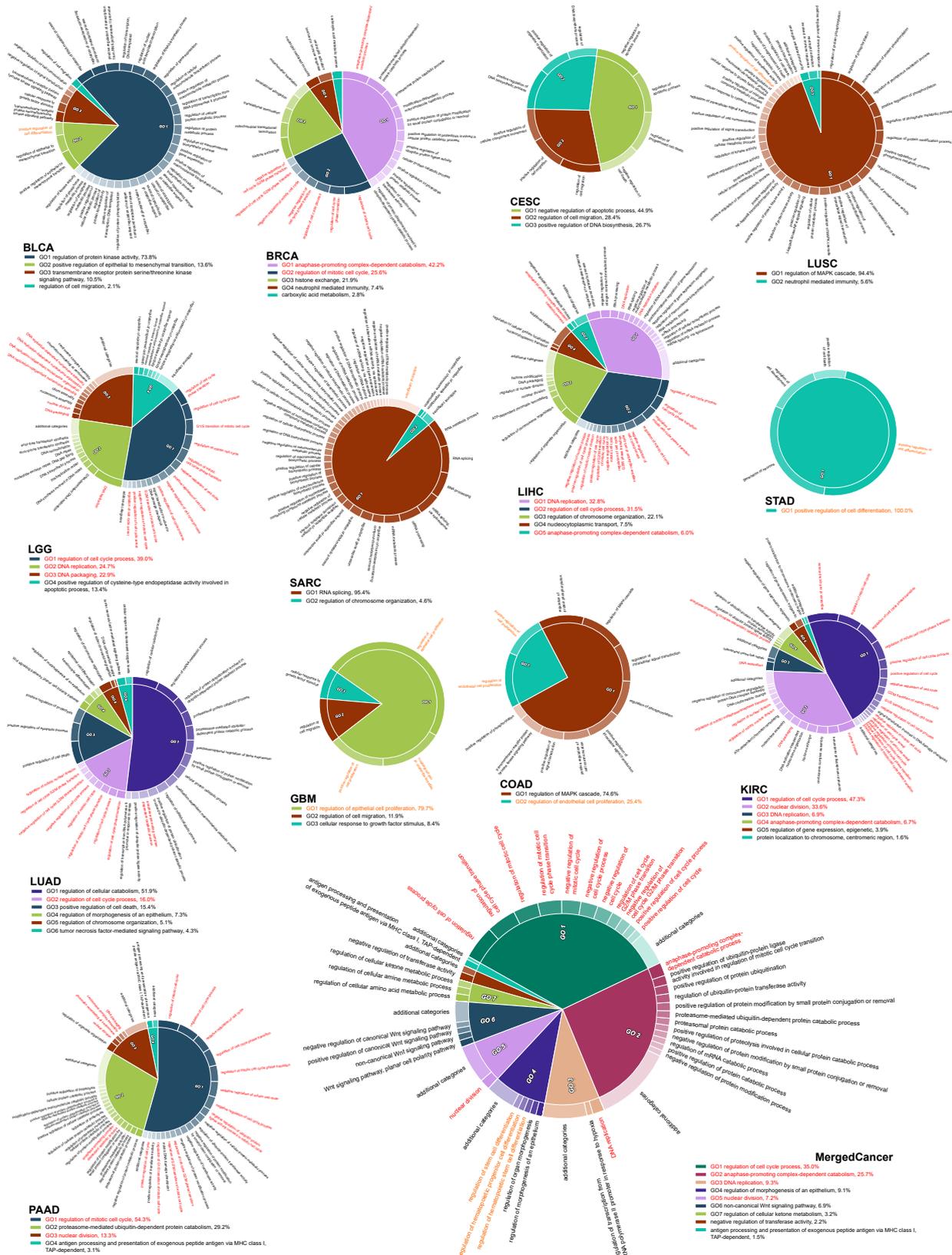


Figure 4. The survival influential modules formed by harmful SIGs. Each pie chart shows the identified functional modules in which the harmful SIGs are involved in the corresponding cancer type. The percentage is the relative significance of one functional module to the others. The functional modules associated with cell cycle are marked by red; modules related to cell proliferation or differentiation are marked by orange.

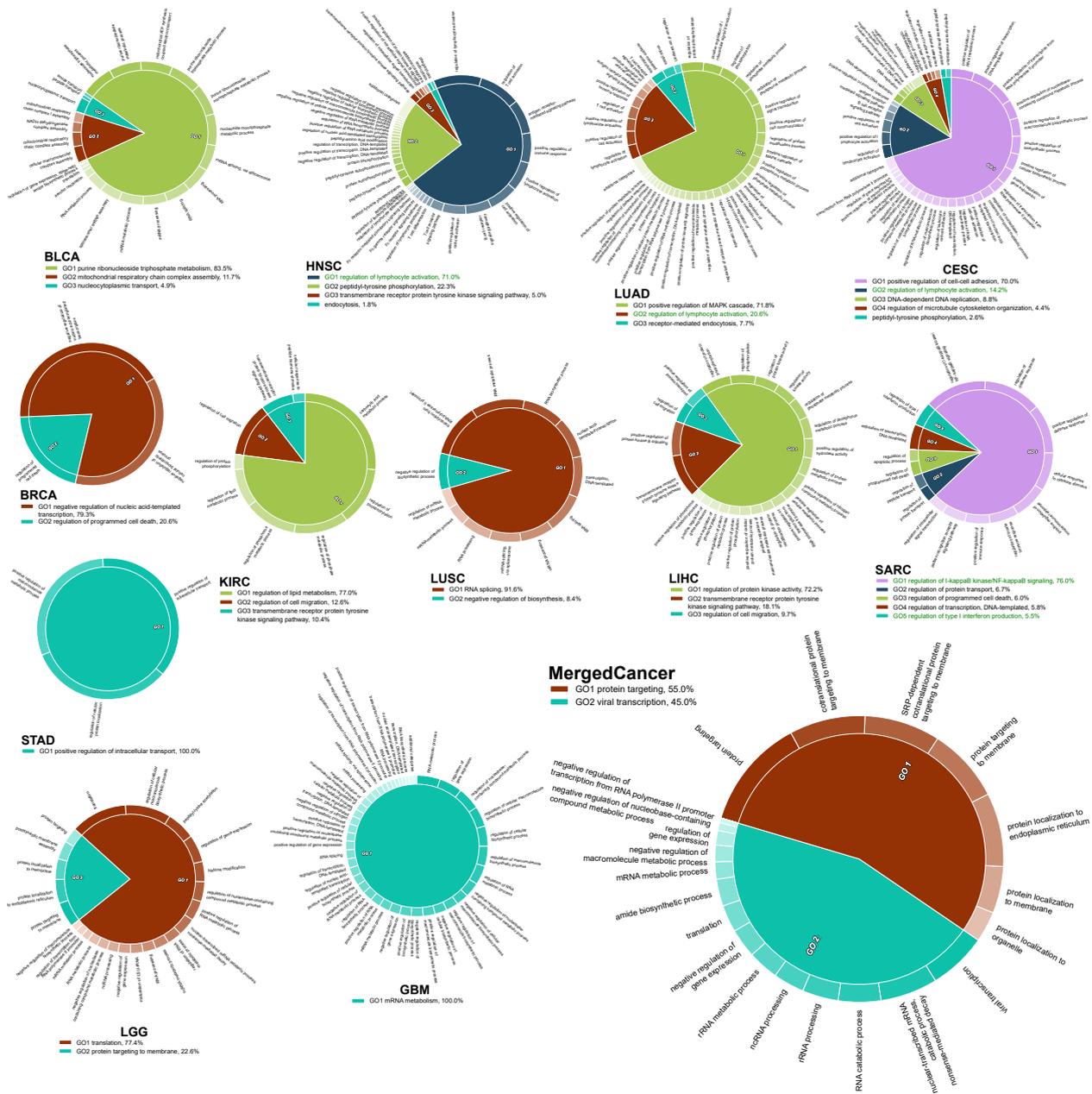


Figure 5. The survival influential modules formed by protective SIGs. Each pie chart shows the identified functional modules in which the protective SIGs were involved in the corresponding cancer type. The percentage is the relative significance of one functional module to the others. The functional modules associated with immune response are marked by green.

2.5. The Cancer Hallmarks of SIGs in Pan-Cancer

To investigate how these SIGs influence carcinogenesis in pan-cancer, we obtained hallmark gene sets from the Molecular Signature Database (MSigDB) [25–27] and performed enrichment analysis using Fisher’s exact test. These hallmark gene sets are well-defined and representative biological processes and pathways in cells. We observed that pan-cancer harmful genes were significantly over-represented for all the hallmark categories (Figure 6A). This observation implies that the way harmful SIGs are involved in carcinogenesis might be variable and prevalent in respect to cellular function. More specifically, we found that the pan-cancer harmful SIGs were significantly enriched with several cancer-relevant hallmarks, such as epithelial mesenchymal transition, DNA repair, glycolysis, hypoxia, apoptosis, MYC target, p53 pathway, and E2F targets (see full significant list

in Table S4) [28,29]. This result suggests that the pan-cancer harmful SIGs might promote carcinogenesis through participating in these cancer hallmark processes. In contrast, we observed that pan-cancer protective genes were significantly overrepresented only in the metabolic category, but underrepresented in the immune response and pathway categories (Figure 6B). Interestingly, two cancer-relevant hallmarks, bile acid and fatty acid metabolism, in which pan-cancer protective SIGs are enriched have been reported to contain cancer-therapeutic potential [30,31].

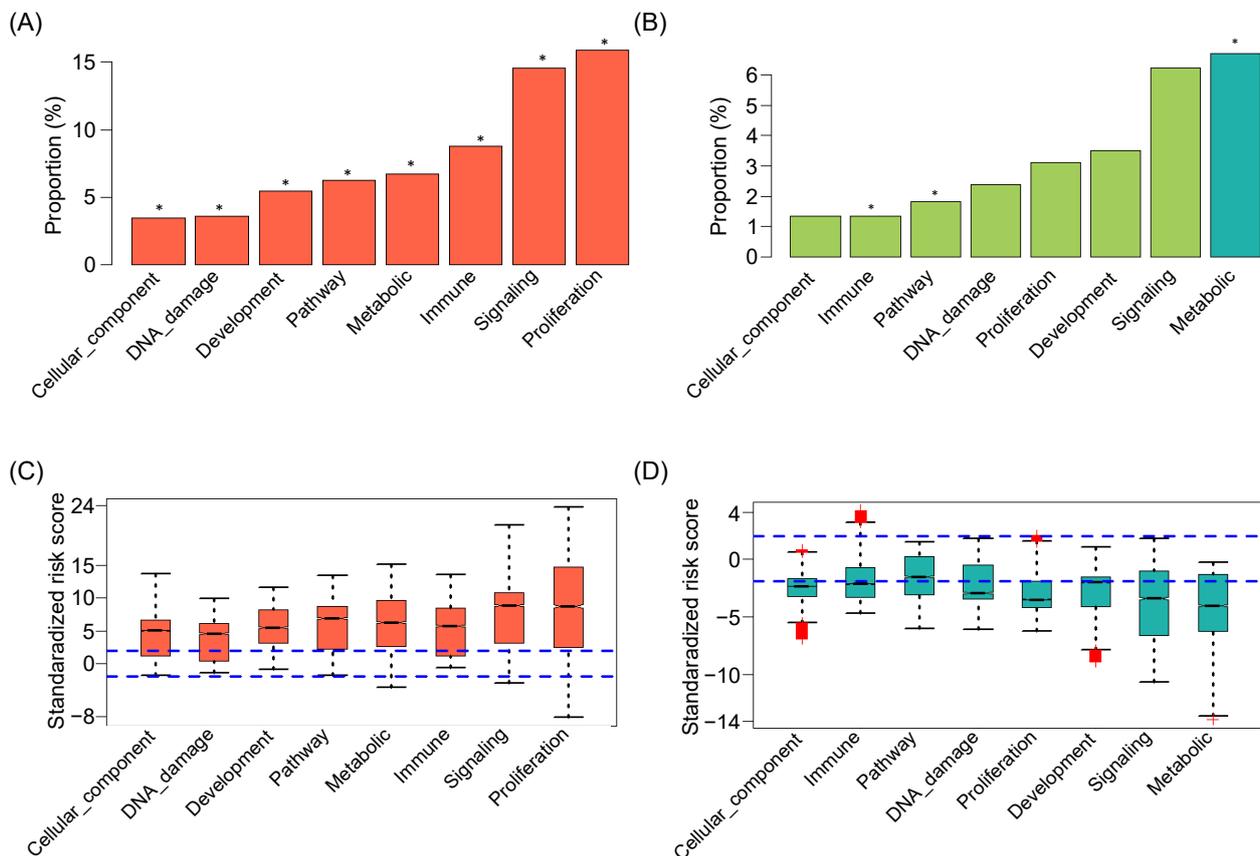


Figure 6. Enrichment analysis of pan-cancer SIGs in cancer-relevant hallmarks. (A,B) Overrepresentation or underrepresentation of pan-cancer harmful (A) and protective (B) genes in MSigDB hallmark gene sets. The x-axis shows the MSigDB hallmark category. The y-axis shows the proportion of the pan-cancer harmful (B), and protective (C) genes of the total genes in the MSigDB category. Dark and light bars represent overrepresentation and underrepresentation, respectively. An asterisk signifies a significant proportion of pan-cancer SIGs (p -value < 0.05 , Fisher's exact test) in the corresponding hallmark category. (C,D) Risk score of pan-cancer harmful (C), and protective (D), SIGs in the MSigDB category compared to overall patient survival. Positive scores represent a high expression level of genes significantly associated with poor survival, and negative scores indicate better survival. The dashed lines represent an absolute standardized risk score value of ± 1.96 (corresponding to a p -value < 0.05).

To investigate the pan-cancer harmful and protective SIGs in hallmark categories associated with patient survival, we built a risk score using the univariate z-score for overall survival for each patient. We observed that pan-cancer harmful and protective SIGs in the MSigDB category tended to have a positive and negative standardized risk score, respectively (Figure 6C,D). This observation suggested that the pan-cancer harmful genes in all the MSigDB categories were significantly associated with poor patient survival, emphasizing the role of the pan-cancer harmful SIGs in promoting carcinogenesis through involving the cancer-relevant hallmarks. In contrast, the pan-cancer protective genes in the signaling and metabolic category were highly significantly associated with good patient

survival. Collectively, these observations further highlighted the potential roles played by the identified pan-cancer SIGs in carcinogenesis.

2.6. Identification of Clinically Relevant Pan-Cancer Harmful SIGs in the Proliferation Hallmark

According to the above results, we could conclude that the harmful genes involved in proliferation are critical to the poor prognosis of cancer patients. Therefore, we selected pan-cancer harmful genes in the proliferation category and performed a series of analyses. First, we performed a pairwise co-expression analysis by applying spearman correlation, and then transformed the obtained SCCs into z-scores using Fisher's z-transformation in the merged cancer dataset, and also in each cancer. With a z-score cut-off of three, we identified 5950 co-expressed gene pairs in common, suggesting a strong correlation between harmful genes in all 16 cancer types. To have a more detailed view of these co-expressed genes, we identified 174 of them possessing a unique protein-protein interaction among the 96 pan-cancer harmful genes (harmful proteins) in the human interactome. Highly connected nodes are usually defined as "hubs". We filtered the 10 genes (*CDK1*, *CDC20*, *PLK1*, *AURKA*, *AURKB*, *BRCA1*, *BUB1B*, *MCM2*, *BUB1*, and *MCM7*) which were the top 10 highly connected and having a maximum degree of 20 and a minimum degree of 7 (Figure 7A). In addition, two distinct sub-networks were observed. The large sub-network consisted of key cancer proliferation genes mainly categorized as kinases, while the other small sub-network were mainly categorized as proteasomes (Figure 7A). We also identified that these genes were differentially expressed in primary tumor and matched normal samples in pan-cancer (Figure 7B) and also upregulation in 10 cancer types (Figure S5), further suggesting the oncogenicity of these pan-cancer harmful SIGs. We further categorized the expression profile of these 10 highly co-expressed genes into low expression groups and high expression groups, based on their median expression. We found the high expression levels of these genes were significantly associated with poor survival of patients through pan-cancer analyses (Figure 7C–L). We also found that some of the 10 genes showed clinical relevance by associating with patient survival in at least 11 cancer types (Figure S6). Interestingly, the Kaplan-Meier plot of *AURKA*, *AURKB*, *BUB1*, *BUB1B*, *CDC20*, *CDK1*, and *PLK1* showed a wide clear separation between low expression and high expression curves in KIRC and LGG (Figure S6), suggesting them as prognostic biomarkers. For example, *CDC20*, a potential novel target for cancer therapy has been reported to be dysregulated in the majority of human cancers, including oral squamous cell carcinoma [32], human bladder carcinoma [33], pancreatic cancer [34], colorectal cancer [35], breast cancer [36], glioblastoma [37], and non-small cell lung cancer [38], and showed clinical relevance in six cancer types, i.e., BLCA, KIRC, LIHC, LUAD, LGG, and PAAD (Figure S6), while *PLK1*, another potential target for cancer therapy [39] showed clinical relevance in eight cancer types, including BLCA, COAD, HNSC, KIRC, LIHC, LUAD, LGG, and PAAD (Figure S6). Taken together, a strong correlation between these ten SIGs and their significant association with major cancer types further suggested their potential role in promoting carcinogenesis.

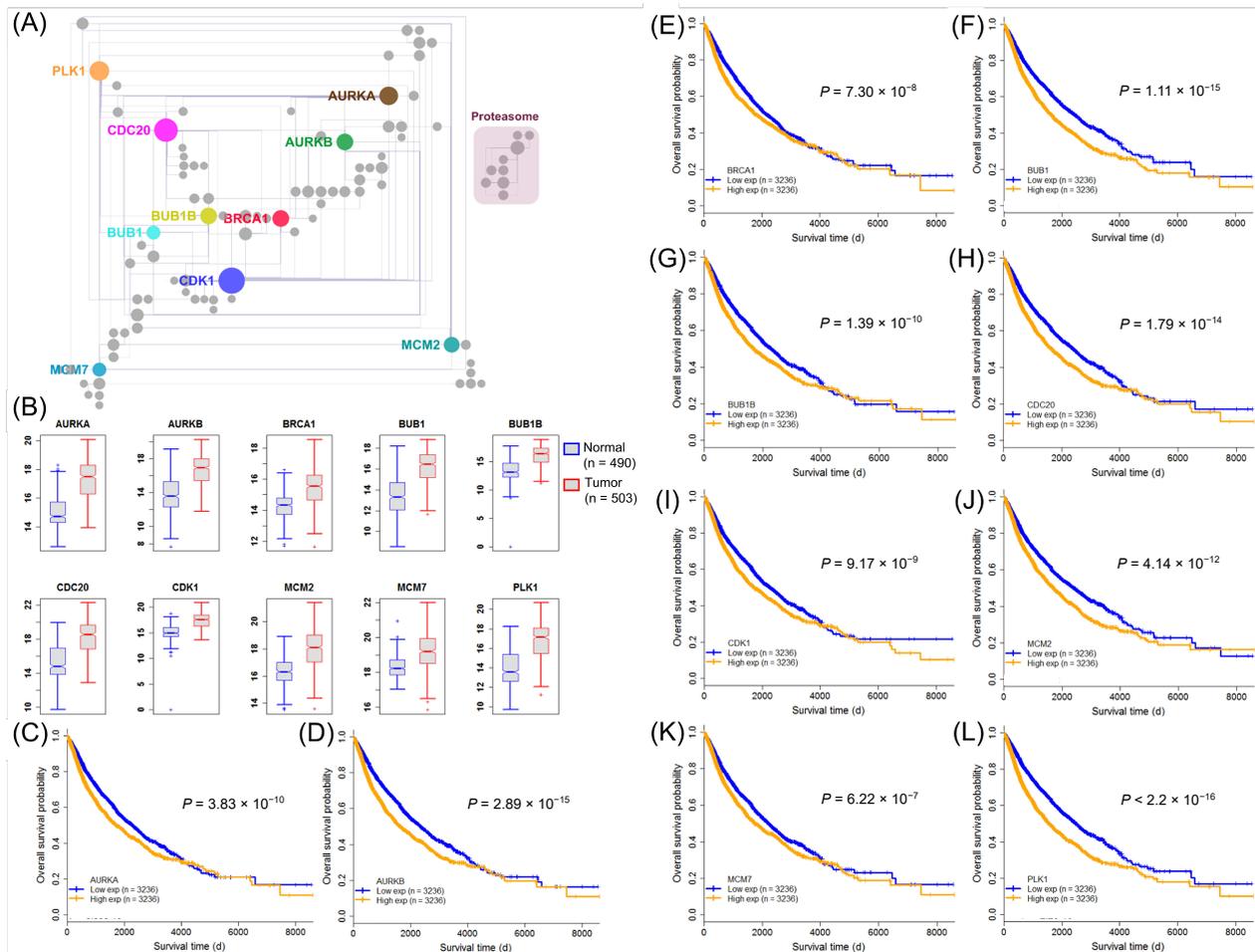


Figure 7. Clinically relevant genes in pan-cancer. **(A)** The human interactome of significantly co-expressed pan-cancer harmful genes (harmful proteins) in the MSigDB category of proliferation. Nodes represent genes and edges protein-protein interactions. Node size is proportional to the degree of the node. Nodes shown in color represent nodes with the 10 highest degrees. **(B)** Expression profile of clinically relevant genes in primary tumor and matched normal samples of pan-cancer. The RSEM normalized expression value is displayed in the $\log_2(x + 1)$ scale. **(C–L)** Survival estimates of overall survival in pan-cancer patients ($n = 6584$). Kaplan-Meier plots of low expression and high expression groups, based on the median expression. The p -values were obtained using a Mantel log-rank test.

3. Discussion

In this study, we identified genes influencing patient survival in 16 individual cancer types and pan-cancer. Moreover, our approach tried to identify the survival influential genes depending on each cancer type, instead of using a global and fixed cut-off across all the studied cancer types. Intuitively, the survival influential genes should not be prevalent in the genome to control the fatal sensitivity to the perturbations inside or outside the cells. Accordingly, previous studies have reported that the proportion of essential genes in the human cancer genome is only about 10% [1,40,41]. Likely, the proportion of SIGs identified by our approach was around 10%, except for LGG and KIRC, which were still with a relatively higher proportion of around 31% and 20%, respectively (Figure 1C). This observation might reveal a limitation of our approach. Even so, our approach outperformed the conventional method of a fixed threshold and LASSO algorithm, and successfully identified the SIGs possessing cancer-dependent properties from co-expressed network analysis. However, the permutation strategy is computationally expensive, thus a more effective approach for assessing cancer-dependent thresholds to determine the survival-influential genes in different cancers is needed.

Furthermore, we observed that the expression level of survival-influential genes in the corresponding cancer type is distinguishable from the other cancers (Figure S7). Interestingly, in the corresponding cancer type, the harmful SIGs were down-regulated, and the protective SIGs were up-regulated, compared to other cancer types (Figure S7). This observation might suggest one imaginable scenario: cancer patients fight against carcinogenesis through inhibiting the harmful SIGs and/or activating the protective SIGs. It is worth noting that most samples in the data set are biased to alive patients; namely the harmful and protective genes could be repressed and promoted respectively to aid patient survival. Indeed, we observed that, in deceased patients, the expression level of harmful and protective genes was elevated and decreased, respectively (Figure S8). However, this observation could be biased by the objective of the Cox regression model, to assess the association between covariates (gene expression level in this study), and the risk of patient survival. That is to say, genes increasingly or decreasingly expressed in deceased samples compared to censored ones have a higher probability of being identified as harmful or protective genes, respectively. Indeed, the z scores of coefficients from the Cox regression model were significantly and positively correlated with the Cohen's D of gene expression level between deceased and censored samples (deceased-censored) for all 16 cancer types (the averaged Spearman correlation coefficient was 0.6, all p -values were smaller than 10^{-128}). Interestingly, we found that the harmful and protective genes were expressed less and more, respectively, in censored (alive) patients with early stage tumors compared to late stage in 7 out of 13 cancer types (Figure S9). This observation implies that, in the patients with a lower risk of survival, the harmful and protective genes could be repressed and promoted, respectively. However, experiments to demonstrate the effect of survival influential genes on carcinogenesis or patient survival are required to validate this observation.

Additionally, we have identified ten clinically relevant genes whose expression can identify the high-risk group of patients with poor outcome in various cancer types. Our findings were also supported by various published literature reports. High expression of *CDC20* showed an association with tumor recurrence and patient death in bladder cancer and pancreatic cancer [33,34]. Additionally, high expression of *CDC20* was significantly associated with overall survival in advanced clinical stage (stage III and IV) patients with colorectal cancer [35]. A meta-analysis report from five studies on approximately 700 colorectal patients showed that high expression of *PLK1* was associated with worse patient survival [42]. Moreover, multivariate cox regression analysis, after adjusting for clinicopathological factors confirmed that high *PLK1* expression at the protein level was independently associated with poor outcome in patients with lung squamous cell carcinoma [43].

4. Materials and Methods

4.1. Data Collection and Preprocessing

The RNA-Seq V2 transcriptome expression datasets and corresponding clinical information from The Cancer Genome Atlas (TCGA) were downloaded from the UCSC Xena browser [44]. We used normalized read counts inferred via RSEM (RNA-Seq by Expectation Maximization) algorithm from RNA-Seq V2 data as mRNA expressions. To increase the statistical power of the survival analysis, we only included the cancer types with sufficient samples (>150) and deceased events (>50); and kept genes and samples with less than 50% unexpressed (RSEM = 0) samples and genes, respectively. Finally, we obtained 6584 samples in total (Figure S10A) and around 17,690 detectable genes on average (Figure S10B) across 16 cancer types (Table S5). To identify survival influential genes in pan-cancer, we assembled a pan-cancer gene expression profile that covered 6584 samples and 16,056 intersecting genes across 16 used cancer types, and termed it merged cancer. The studies in the pan-cancer atlas from TCGA developed and adopted this strategy to enhance the pan-cancer features cross cancers [45–47].

4.2. Identification of Survival Influential Genes in Cancers and Pan-Cancer

In this study, we applied the Cox regression model as the main framework to discover survival influential genes in cancers. We first used the univariate Cox regression model to examine the association between clinical variables and patient survival time and to identify the potential clinical confounding factors (Supplementary Dataset S2). Then, we utilized the multivariate Cox regression model (the covariates are the expression level of the tested gene and the identified significant clinical confounding factors) to assess the influence of a gene on patient survival. In this study, we denoted the genes with positive and negative coefficients as harmful and protective to patient survival, respectively. The significance of the coefficient β was determined by comparing with the null model, which hypothesizes that the changes of tested gene expression level have no effect on patient survival, and was further tested by a likelihood ratio test and the Wald test [48,49]. The Wald test examines whether the observed regression coefficient statistically differs from zero, a reported a z-score (standard score), and a p -value estimated by z-score for the observed regression coefficient. The Cox regression model was performed using the *Survival* R package [50].

Previous studies have observed that the distributions of p -values derived from the Cox-regression model vary across cancers [15,16]. Accordingly, we applied the Monte Carlo algorithm combined with effect size to determine the p -value threshold of the regression coefficient for each cancer type independently. First, we created 1000 simulated cohorts for each cancer type. Each simulated cohort was provided with the same number of patients in the real data, but with randomly permuted survival information. Notably, both survival time and deceased status were coupled as a pair to be permuted. Accordingly, the simulated cohort contained identical survival profiles but randomly permuted gene expression profiles. Then, we used a Cox-regression model on each simulated cohort to assign the simulated p -values of regression coefficient to genes. We then calculated the odds ratio to estimate the effect size of each significance level. The odds ratio can be described as below:

$$OR_p = \frac{O_I}{O_S} = \frac{N_{SI}}{N_{SS}} / \frac{N_{II}}{N_{IS}} \quad (1)$$

where OR_p is the odds ratio of the significance level p -value $< p$; O_I (O_S) is the odds of significant genes in real data being identified as insignificant (significant) in the simulated data; N_{SI} and N_{SS} (N_{II} and N_{IS}) represent the number of significant (insignificant) genes in real data being identified as insignificant and significant in the simulated data, respectively. In other words, the odds ratio in this study described the disagreement between the significant genes in the real data and the insignificant genes in the simulated data under the given significance level.

For each cancer type, the above steps were repeated for 1000 simulated cohorts, and the mean odds ratio was used as the effect size of each significance level. We then applied a p -value reaching the maximum effect size as the threshold to determine the survival-influential genes. The effect size was calculated for harmful and protective genes separately.

We then performed the same procedure on the merged cancer data set to identify the pan-cancer SIGs. To increase the generality of the pan-cancer SIGs, we created 16 cancer minus gene expression profiles that merged gene expression profiles from 15 cancer types: for example, the BRCA minus merged the gene expression profiles from 15 out of the 16 used samples, except for BRCA. We then performed the same procedure on each cancer minus data set to identify cancer minus SIGs. Subsequently, we defined the genes identified as SIGs in (1) the merged cancer dataset, (2) at least one cancer type, and (3) all cancer minus datasets as the pan-cancer SIGs.

4.3. Compilation of the Cancer-Associated Genes

Next, to assess the association between the identified SIGs and cancer, we compiled a cancer-associated gene list from three datasets: (1) 688 cancer genes from the COSMIC release (v85) [51]; (2) 550 cancer essential genes screened from the CRISPR system in hu-

man cancer cell lines [3]; and (3) 769 cancer genes from the cancer dependency map [4]. Additionally, we collected a small but highly confident harmful gene set from the literature [52–57]. We then performed enrichment analysis using Fisher’s exact test to evaluate the association between SIGs and cancer.

4.4. Functional Modules of SIGs

To investigate the biological processes of SIGs involved in carcinogenesis, we performed conventional and network-wise functional enrichment analyses to identify the functional modules. The functional annotations of genes were obtained from gene ontology (GO) [58,59]. In the conventional functional enrichment analysis, the significance was established by p -value derived from the hypergeometric test, which is described below:

$$P(X = k) = \frac{\binom{m}{k} \binom{N - m}{n - k}}{\binom{N}{n}} \quad (2)$$

where X denotes the evaluated function; N represents the number of GO annotated genes; m indicates the number of SIGs; n represents the number of genes with the evaluated function; and k indicates the number of SIGs with the evaluated function. To enhance the functional relationship between the identified genes, protein interaction and network-wise functional enrichment analyses were incorporated to discover the functional modules within SIGs [60]. The source of the human protein-protein interaction (PPI) data was InBio Map [20]. The network-wise functional enrichment analysis was modified from the conventional method. The significance of the tested function was based on p -values produced from a modified hypergeometric test. The hypergeometric distribution for the network-wise approach is described below:

$$P_e(X = k_e) = \frac{\binom{m_e}{k_e} \binom{N_e - m_e}{n_e - k_e}}{\binom{N_e}{n_e}} \quad (3)$$

where e is the abbreviation of the PPI. Each symbol has the same meaning as in the conventional hypergeometric distribution, but the counting objects are changed from genes to functional PPIs. The functional PPIs are interactions formed by the two genes involved in the same functions. This approach revealed the significant protein interaction functional modules in which the identified genes were involved. To incorporate the conditional information (related to cancer) into the functional modules, we only studied co-expressed PPI in the network-wise functional enrichment analysis; and genes that were formed by co-expressed PPI in the conventional analysis. Furthermore, we used co-expressed PIN as a background instead of the static PIN to enhance the cancer-dependent constraint. Subsequently, the function modules with adjusted p -value < 0.05 from both analyses were identified as the significantly enriched function modules. These two p -values were adjusted by the Benjamini and Hochberg multiple testing procedures to control the false discovery rate (FDR) [61]. Additionally, only the significant function modules with a depth greater than six in the GO database were used to increase the functional specificity. The depth is the shortest path length from a function to the root, which is the term biological process (GO:0008150), in the GO tree structure. Finally, the significant functional modules with more than five SIGs and five co-expressed PPIs were identified as the functional modules for the following analyses.

The discovered functional modules were further summarized by the REVIGO [62] algorithm, with a similarity ≥ 0.9 , which was calculated from the SimRel algorithm [63] and visualized using the CirGO package [64].

5. Conclusions

In this study, we tried to identify SIGs in cancers and pan-cancer and investigated the characteristics of SIGs in carcinogenesis. In summary, the SIGs are highly exclusive across cancers, and tend to occupy pivotal positions in the co-expressed PIN associated with the cancer. More specifically, the harmful SIGs prefer to participate in the biological processes related to cell cycle and cell proliferation. However, the functions of protective SIGs are variable across cancer types, but might be involved in bile acid or fatty acid metabolism to improve patient survival. Briefly, the identified SIGs may provide a potential biomarker pool for cancer therapy, and facilitate improving our knowledge of the molecular mechanism of carcinogenesis.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/ijms22094384/s1>.

Author Contributions: C.-C.L. conceived, designed, and coordinated this study. D.S., Y.-L.C., Y.-C.L., and C.-C.L. implemented the computational method and carried out the analysis. D.S. and Y.-L.C. drafted the manuscript. C.-C.L. and D.S. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology in Taiwan (MOST 104-2320-B-010-037- and MOST 107-2221-E-010-016-MY2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The list of identified SIGs and used clinical confounding factors are available online as Supplementary Datasets S1 and S2 at www.mdpi.com.

Acknowledgments: We thank Hsuan-Cheng Huang, Yu-Chao Wang, and Cho-Yi Chen in the Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, provided invaluable suggestions in improving this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gerdes, S.; Edwards, R.; Kubal, M.; Fonstein, M.; Stevens, R.; Osterman, A. Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* **2006**, *17*, 448–456. [[CrossRef](#)] [[PubMed](#)]
2. Juhas, M.; Eberl, L.; Glass, J.I. Essence of life: Essential genes of minimal genomes. *Trends Cell Biol.* **2011**, *21*, 562–568. [[CrossRef](#)] [[PubMed](#)]
3. Wang, T.; Birsoy, K.; Hughes, N.W.; Krupczak, K.M.; Post, Y.; Wei, J.J.; Lander, E.S.; Sabatini, D.M. Identification and characterization of essential genes in the human genome. *Science* **2015**, *350*, 1096–1101. [[CrossRef](#)] [[PubMed](#)]
4. Tsherniak, A.; Vazquez, F.; Montgomery, P.G.; Weir, B.A.; Kryukov, G.; Cowley, G.S.; Gill, S.; Harrington, W.F.; Pantel, S.; Krill-Burger, J.M.; et al. Defining a Cancer Dependency Map. *Cell* **2017**, *170*, 564–576.e16. [[CrossRef](#)]
5. Patel, S.J.; Sanjana, N.E.; Kishton, R.J.; Eidizadeh, A.; Vodnala, S.K.; Cam, M.; Gartner, J.J.; Jia, L.; Steinberg, S.M.; Yamamoto, T.N.; et al. Identification of essential genes for cancer immunotherapy. *Nature* **2017**, *548*, 537–542. [[CrossRef](#)]
6. Gustafson, A.M.; Snitkin, E.S.; Parker, S.C.; DeLisi, C.; Kasif, S. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genom.* **2006**, *7*, 265. [[CrossRef](#)]
7. Acencio, M.L.; Lemke, N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinform.* **2009**, *10*, 290. [[CrossRef](#)]
8. Yamada, S.; Lin, C.Y.; Chang, S.L.; Chao, T.F.; Lin, Y.J.; Lo, L.W.; Chung, F.P.; Hu, Y.F.; Tuan, T.C.; Liao, J.N.; et al. Risk of Stroke in Patients With Short-Run Atrial Tachyarrhythmia. *Stroke* **2017**, *48*, 3232–3238. [[CrossRef](#)]
9. Kleinbaum, D.G.; Klein, M. *Survival Analysis: A Self-Learning Text*, 3rd ed.; Springer: New York, NY, USA, 2012.
10. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.A.; Tomita, M.; Wanner, B.L.; Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* **2006**, *2*. [[CrossRef](#)]
11. Butland, G.; Babu, M.; Diaz-Mejia, J.J.; Bohdana, F.; Phanse, S.; Gold, B.; Yang, W.; Li, J.; Gagarinova, A.G.; Pogoutse, O.; et al. eSGA: E. coli synthetic genetic array analysis. *Nat. Methods* **2008**, *5*, 789–795. [[CrossRef](#)]
12. Costanzo, M.; Baryshnikova, A.; Bellay, J.; Kim, Y.; Spear, E.D.; Sevier, C.S.; Ding, H.; Koh, J.L.; Toufighi, K.; Mostafavi, S.; et al. The genetic landscape of a cell. *Science* **2010**, *327*, 425–431. [[CrossRef](#)]
13. Giaever, G.; Chu, A.M.; Ni, L.; Connelly, C.; Riles, L.; Veronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; Andre, B.; et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **2002**, *418*, 387–391. [[CrossRef](#)]

14. Glass, J.I.; Assad-Garcia, N.; Alperovich, N.; Yooseph, S.; Lewis, M.R.; Maruf, M.; Hutchison, C.A.; Smith, H.O.; Venter, J.C. Essential Genes of a Minimal Bacterium. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 425–430. [[CrossRef](#)]
15. Anaya, J.; Reon, B.; Chen, W.M.; Bekiranov, S.; Dutta, A. A pan-cancer analysis of prognostic genes. *PeerJ* **2015**, *3*, e1499. [[CrossRef](#)]
16. Yang, Y.; Han, L.; Yuan, Y.; Li, J.; Hei, N.; Liang, H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **2014**, *5*, 3231. [[CrossRef](#)]
17. Gong, J.; Li, Y.; Liu, C.J.; Xiang, Y.; Li, C.; Ye, Y.; Zhang, Z.; Hawke, D.H.; Park, P.K.; Diao, L.; et al. A Pan-cancer Analysis of the Expression and Clinical Relevance of Small Nucleolar RNAs in Human Cancer. *Cell Rep.* **2017**, *21*, 1968–1981. [[CrossRef](#)]
18. Sun, J.; Zhao, Z. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genom.* **2010**, *11* (Suppl. 3), S5. [[CrossRef](#)]
19. Taylor, I.W.; Linding, R.; Warde-Farley, D.; Liu, Y.; Pesquita, C.; Faria, D.; Bull, S.; Pawson, T.; Morris, Q.; Wrana, J.L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **2009**, *27*, 199–204. [[CrossRef](#)]
20. Li, T.; Wernersson, R.; Hansen, R.B.; Horn, H.; Mercer, J.; Slodkovic, G.; Workman, C.T.; Rigina, O.; Rapacki, K.; Staerfeldt, H.H.; et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **2017**, *14*, 61–64. [[CrossRef](#)]
21. Collins, K.; Jacks, T.; Pavletich, N.P. The cell cycle and cancer. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 2776–2778. [[CrossRef](#)]
22. Otto, T.; Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer* **2017**, *17*, 93–115. [[CrossRef](#)]
23. Adam, J.K.; Odhav, B.; Bhoola, K.D. Immune responses in cancer. *Pharmacol. Ther.* **2003**, *99*, 113–132. [[CrossRef](#)]
24. Gonzalez, H.; Hagerling, C.; Werb, Z. Roles of the immune system in cancer: From tumor initiation to metastatic progression. *Genes Dev.* **2018**, *32*, 1267–1284. [[CrossRef](#)]
25. Liberzon, A.; Birger, C.; Thorvaldsdottir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [[CrossRef](#)]
26. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdottir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [[CrossRef](#)]
27. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [[CrossRef](#)]
28. Phillips, J.W.; Pan, Y.; Tsai, B.L.; Xie, Z.; Demirdjian, L.; Xiao, W.; Yang, H.T.; Zhang, Y.; Lin, C.H.; Cheng, D.; et al. Pathway-guided analysis identifies Myc-dependent alternative pre-mRNA splicing in aggressive prostate cancers. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 5269–5279. [[CrossRef](#)]
29. Malta, T.M.; Sokolov, A.; Gentles, A.J.; Burzykowski, T.; Poisson, L.; Weinstein, J.N.; Kaminska, B.; Huelsken, J.; Omberg, L.; Gevaert, O.; et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **2018**, *173*, 338–354.e15. [[CrossRef](#)] [[PubMed](#)]
30. Di Ciaula, A.; Wang, D.Q.; Molina-Molina, E.; Lunardi Baccetto, R.; Calamita, G.; Palmieri, V.O.; Portincasa, P. Bile Acids and Cancer: Direct and Environmental-Dependent Effects. *Ann. Hepatol.* **2017**, *16* (Suppl. 1), S87–S105. [[CrossRef](#)] [[PubMed](#)]
31. Koundouros, N.; Poulogiannis, G. Reprogramming of fatty acid metabolism in cancer. *Br. J. Cancer* **2020**, *122*, 4–22. [[CrossRef](#)] [[PubMed](#)]
32. Mondal, G.; Sengupta, S.; Panda, C.K.; Gollin, S.M.; Saunders, W.S.; Roychoudhury, S. Overexpression of Cdc20 leads to impairment of the spindle assembly checkpoint and aneuploidization in oral cancer. *Carcinogenesis* **2007**, *28*, 81–92. [[CrossRef](#)]
33. Choi, J.W.; Kim, Y.; Lee, J.H.; Kim, Y.S. High expression of spindle assembly checkpoint proteins CDC20 and MAD2 is associated with poor prognosis in urothelial bladder cancer. *Virchows Arch. Int. J. Pathol.* **2013**, *463*, 681–687. [[CrossRef](#)]
34. Chang, D.Z.; Ma, Y.; Ji, B.; Liu, Y.; Hwu, P.; Abbruzzese, J.L.; Logsdon, C.; Wang, H. Increased CDC20 expression is associated with pancreatic ductal adenocarcinoma differentiation and progression. *J. Hematol. Oncol.* **2012**, *5*, 15. [[CrossRef](#)]
35. Wu, W.J.; Hu, K.S.; Wang, D.S.; Zeng, Z.L.; Zhang, D.S.; Chen, D.L.; Bai, L.; Xu, R.H. CDC20 overexpression predicts a poor prognosis for patients with colorectal cancer. *J. Transl. Med.* **2013**, *11*, 142. [[CrossRef](#)]
36. Karra, H.; Repo, H.; Ahonen, I.; Loytyniemi, E.; Pitkanen, R.; Lintunen, M.; Kuopio, T.; Soderstrom, M.; Kronqvist, P. Cdc20 and securin overexpression predict short-term breast cancer survival. *Br. J. Cancer* **2014**, *110*, 2905–2913. [[CrossRef](#)]
37. Mao, D.D.; Gujar, A.D.; Mahlokozero, T.; Chen, I.; Pan, Y.; Luo, J.; Brost, T.; Thompson, E.A.; Turski, A.; Leuthardt, E.C.; et al. A CDC20-APC/SOX2 Signaling Axis Regulates Human Glioblastoma Stem-like Cells. *Cell Rep.* **2015**, *11*, 1809–1821. [[CrossRef](#)]
38. Kato, T.; Daigo, Y.; Aragaki, M.; Ishikawa, K.; Sato, M.; Kaji, M. Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. *J. Surg. Oncol.* **2012**, *106*, 423–430. [[CrossRef](#)]
39. Liu, Z.; Sun, Q.; Wang, X. PLK1, A Potential Target for Cancer Therapy. *Transl. Oncol.* **2017**, *10*, 22–32. [[CrossRef](#)]
40. Cheng, C.Y.; Krishnakumar, V.; Chan, A.P.; Thibaud-Nissen, F.; Schobel, S.; Town, C.D. Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **2017**, *89*, 789–804. [[CrossRef](#)]
41. Tzafrir, I.; Pena-Muralla, R.; Dickerman, A.; Berg, M.; Rogers, R.; Hutchens, S.; Sweeney, T.C.; McElver, J.; Aux, G.; Patton, D.; et al. Identification of genes required for embryo development in Arabidopsis. *Plant Physiol.* **2004**, *135*, 1206–1220. [[CrossRef](#)] [[PubMed](#)]
42. Ran, Z.; Chen, W.; Shang, J.; Li, X.; Nie, Z.; Yang, J.; Li, N. Clinicopathological and prognostic implications of polo-like kinase 1 expression in colorectal cancer: A systematic review and meta-analysis. *Gene* **2019**, *721*, 144097. [[CrossRef](#)] [[PubMed](#)]

43. Li, H.; Wang, H.; Sun, Z.; Guo, Q.; Shi, H.; Jia, Y. The clinical and prognostic value of polo-like kinase 1 in lung squamous cell carcinoma patients: Immunohistochemical analysis. *Biosci. Rep.* **2017**, *37*, BSR20170852. [CrossRef] [PubMed]
44. Goldman, M.; Craft, B.; Hastie, M.; Repečka, K.; Kamath, A.; McDade, F.; Rogers, D.; Brooks, A.N.; Zhu, J.; Haussler, D. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* **2019**, 326470. [CrossRef]
45. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W.K.; Luna, A.; La, K.C.; Dimitriadou, S.; Liu, D.L.; Kantheti, H.S.; Saghafein, S.; et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **2018**, *173*, 321–337.e10. [CrossRef]
46. Hoadley, K.A.; Yau, C.; Hinoue, T.; Wolf, D.M.; Lazar, A.J.; Drill, E.; Shen, R.; Taylor, A.M.; Cherniack, A.D.; Thorsson, V.; et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **2018**, *173*, 291–304.e6. [CrossRef]
47. Gao, Q.; Liang, W.W.; Foltz, S.M.; Mutharasu, G.; Jayasinghe, R.G.; Cao, S.; Liao, W.W.; Reynolds, S.M.; Wyczalkowski, M.A.; Yao, L.; et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **2018**, *23*, 227–238.e3. [CrossRef]
48. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*, 1st ed.; Springer: New York, NY, USA, 2000.
49. Andersen, P.K.; Gill, R.D. Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Stat.* **1982**, *10*, 1100–1120. [CrossRef]
50. Therneau, T. A Package for Survival Analysis in S. Version 2.38. 2015. Available online: <https://CRAN.R-project.org/package=survival> (accessed on 16 October 2016).
51. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. COSMIC: The Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **2019**, *47*, D941–D947. [CrossRef]
52. Kim, S.K.; Kim, S.Y.; Kim, C.W.; Roh, S.A.; Ha, Y.J.; Lee, J.L.; Heo, H.; Cho, D.H.; Lee, J.S.; Kim, Y.S.; et al. A prognostic index based on an eleven gene signature to predict systemic recurrences in colorectal cancer. *Exp. Mol. Med.* **2019**, *51*, 1–12. [CrossRef]
53. Hou, Z.; Yang, J.; Wang, H.; Liu, D.; Zhang, H. A Potential Prognostic Gene Signature for Predicting Survival for Glioblastoma Patients. *BioMed Res. Int.* **2019**, *2019*, 9506461. [CrossRef]
54. Haider, S.; Wang, J.; Nagano, A.; Desai, A.; Arumugam, P.; Dumartin, L.; Fitzgibbon, J.; Hagemann, T.; Marshall, J.F.; Kocher, H.M.; et al. A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med.* **2014**, *6*, 105. [CrossRef]
55. Peters, C.J.; Rees, J.R.; Hardwick, R.H.; Hardwick, J.S.; Vowler, S.L.; Ong, C.A.; Zhang, C.; Save, V.; O'Donovan, M.; Rassl, D.; et al. A 4-gene signature predicts survival of patients with resected adenocarcinoma of the esophagus, junction, and gastric cardia. *Gastroenterology* **2010**, *139*, 1995–2004.e15. [CrossRef]
56. Van de Vijver, M.J.; He, Y.D.; van't Veer, L.J.; Dai, H.; Hart, A.A.; Voskuil, D.W.; Schreiber, G.J.; Peterse, J.L.; Roberts, C.; Marton, M.J.; et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **2002**, *347*, 1999–2009. [CrossRef]
57. Boutros, P.C.; Lau, S.K.; Pintilie, M.; Liu, N.; Shepherd, F.A.; Der, S.D.; Tsao, M.S.; Penn, L.Z.; Jurisica, I. Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 2824–2828. [CrossRef]
58. Consortium, T.G.O. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338.
59. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
60. Lin, C.C.; Hsiang, J.T.; Wu, C.Y.; Oyang, Y.J.; Juan, H.F.; Huang, H.C. Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst. Biol.* **2010**, *4*, 138. [CrossRef]
61. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
62. Supek, F.; Bosnjak, M.; Skunca, N.; Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **2011**, *6*, e21800. [CrossRef]
63. Schlicker, A.; Domingues, F.S.; Rahnenfuhrer, J.; Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform.* **2006**, *7*, 302. [CrossRef]
64. Kuznetsova, I.; Lugmayr, A.; Siira, S.J.; Rackham, O.; Filipovska, A. CirGO: An alternative circular way of visualising gene ontology terms. *BMC Bioinform.* **2019**, *20*, 84. [CrossRef] [PubMed]