Received: 2 April 2021 | Revised: 25 July 2021 | Accepted: 30 September 2021 | Accepted article online: 12 October 2021

DOI: 10.1002/pmic.202000311

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Universal and cross-cancer prognostic biomarkers for predicting survival risk of cancer patients from expression profile of apoptotic pathway genes

Chakit Arora | Dilraj Kaur | Gajendra P. S. Raghava

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Correspondence
Gajendra P. S. Raghava, Head of Department, Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi-110020, India.
Email: raghava@iiitd.ac.in

Abstract

Numerous cancer-specific prognostic models have been developed in the past, wherein one model is applicable for only one type of cancer. In this study, an attempt has been made to identify universal or multi-cancer prognostic biomarkers and develop models for predicting survival risk across different types of cancer patients. In order to accomplish this, we gauged the prognostic role of mRNA expression of 165 apoptosis-related genes across 33 cancers in the context of patient survival. Firstly, we identified specific prognostic biomarker genes for 30 cancers. The cancer-specific prognostic models achieved a minimum Hazard Ratio, $HR_{SKCM} = 1.99$ and maximum $HR_{THCA} = 41.59$. Secondly, a comprehensive analysis was performed to identify universal biomarkers across many cancers. Our best prognostic model consisted of 11 genes (TOP2A, ISG20, CD44, LEF1, CASP2, PSEN1, PTK2, SATB1, SLC20A1, EREG, and CD2) and stratified risk groups across 27 cancers ($HR_{OV} = 1.53$-$HR_{UVM} = 11.74$). The model was validated on eight independent cancer cohorts and exhibited a comparable performance. Further, we clustered cancer-types on the basis of shared survival related apoptosis genes. This approach proved helpful in development of cross-cancer prognostic models. To show its efficacy, a prognostic model consisting of 15 genes was thereby developed for LGG-KIRC pair ($HR_{KIRC} = 3.27$, $HR_{LGG} = 4.23$). Additionally, we predicted potential therapeutic candidates for LGG-KIRC high risk patients.

KEYWORDS
apoptosis, biomarker, cancer, gene expression, prognosis

List of Abbreviations: ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; LGG, brain lower grade glioma; BRCA, beast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; LCML, chronic myelogenous leukemia; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and Paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumours; THYM, thymoma; THCA, thyroid carcinoma; UCS, uterine carcinosarcoma; UCEC, uterine corpus endometrial carcinoma; UVM, uveal melanoma; TCGA, the cancer genome atlas; GO, gene ontology; HR, hazard ratio; KM, Kaplan Meier; R, Pearson correlation coefficient; GPM, good prognostic marker; BPM, bad prognostic marker; C, concordance index; CI, confidence interval; Cox-PH, cox proportional hazard; OS, overall survival; PI, prognostic index; RSEM, RNA-seq by expectation-maximization; UPGMA, unweighted pair group method with arithmetic mean; DGIdB, drug gene interaction database; J, Jaccard similarity index

https://doi.org/10.1002/pmic.202000311

1 | INTRODUCTION

Cancer is the leading cause of death worldwide [1] and its development has been attributed to various regulatory factors [2]. The exploration of these regulatory mechanisms that lead to cancer has been a hot topic in recent years. It provides important insights into fundamental biological processes linked with carcinogenesis. Besides this, the knowledge about underlying mechanistic processes has led to development of novel drugs and therapeutic strategies. It has also resulted in the identification of several prognostic biomarkers for risk assessment in cancer patients. However, a major challenge in cancer prognosis is the unavailability of a universal biomarker applicable across all the cancers. The availability of even at-least a handful of cross-cancer prognostic biomarker(s) that can be used for risk assessment across maximum cancers remains absent. An ideal universal biomarker would bypass the need of several different biomarkers currently in use. With an added advantage of being time effective, these biomarkers would also be of great financial benefit to already burdened cancer patients. Such patients would also not have to undergo a variety of agonizing prognostic examinations, thereby reducing any further effect on their previously compromised health conditions. However, instead of such biomarkers, there exists a plethora of biomarkers and risk prediction methods. Majority of these biomarkers/methods are specific only to a particular cancer and fail when employed for other cancers. However, with the increase in omics data, a few pan-cancer prognostic biomarkers have also been developed. Recent examples include a comprehensive analysis presented by Ning Zhao et al. [3] wherein, the multi-omics data for 13 cancers was used to identify prognostic biomarker genes specific to the cancers as well as seven genes associated with survival in 13 cancers. Sialic acid-binding immunoglobulin-like lectin 15 (Siglec-15) expression levels were found to be associated with eight cancers in the context of overall-survival. Siglec-15 is a member of family of lectins that recognize sialylated glycans and control immune function. A maximum risk stratification with HR = 3.03, $p = 0.044$ was obtained in THCA (Thyroid carcinoma) patients by employing mRNA expression of Siglec-15 [4]. Yet another study showed that the mRNA expression levels of the gene, Long intergenic non-coding RNA 1614 (LINC01614) can be used to segregate risk groups in 11 cancers based on overall survival, with the maximum separation achieved in THCA patients with HR = 4.047 and $p = 0.010$ [5]. A few other studies have also elucidated the prognostic potential of genes such as FUNDC1 whose expression was linked to prognosis in eight cancers with a maximal risk separation in LIHC (HR = 1.73, logrank-$p = 0.0022$) [6] and HSP90AA1 whose differential expression was observed in eight cancers and was found to be a prognostic biomarker in hepatocellular carcinoma [7]. Apart from these, tumour mutational burden and indel burden have also been recently shown to be linked with prognosis in 14 cancers [8] with the best performance in CHOL (HR = 6.10, $p = 0.002$). These studies are promising, however one of the major pitfalls associated with these studies is the applicability range. Most of the prognostic biomarkers suggested in these studies are applicable to just a handful of cancers. Another crucial disadvantage is the low-risk stratification achieved in terms of Hazard ratios and p-values. Therefore, the challenge for finding more accurate biomarkers which offer prognostic value across a large number of cancers remains open. As discussed earlier, since a multitude of factors cause heterogeneity of cancer, more efforts are required towards thorough investigation of cardinal molecular processes that have been associated with cancer progression and development in the past.

One of these fundamental processes is the programmed cell death process, which, when defunct, leads to accumulation of malignant cells in the body [9]. This cell death mechanism, also known as Apoptosis, is a multistep cellular process involving a large number of regulatory molecules. It primarily consists of genes and their encoded proteins, which orchestrate the death of a cell as a result of various stresses. The downregulation of tumour suppressor gene, p53, leading to tumour development and progression is perhaps the most popular example [10]. Other examples include the downregulation of levels of pro-apoptotic BCL2 family proteins such as BCL2, BCL-XL, MCL1, and upregulation of anti-apoptotic BCL2 family proteins such as BAX, BAK in many cancers [11]. Apoptosis is also one of the widely studied processes in the context of development of prognostic biomarkers and therapeutics which target its key components [12]. Few prominent examples include biomarkers based on correlation of protein/gene expression of apoptotic molecules such as Caspase 3/6, XIAP, Apaf-1, ML-IAP, and XAF1 with patient survival and/or other clinico-pathological features in Melanoma as reviewed by Charles EM et al. [13]; identification of apoptosis related biomarkers such as BCL2 (HR = 5.21, $p = 0.0063$), Fas/FasL (HR = 3.49, $p = 0.005$), p53 (HR = 2.48, $p = 0.046$), TRAIL (HR = 1.21, $p = 0.026$) in the context of overall survival of colorectal cancer patients along-with various statistical and

STATEMENT OF SIGNIFICANCE

- In the past, several cancer-specific prognostic biomarkers/models have been identified/developed, which are applicable to only a particular type of cancer population. In this study, we have attempted to identify universal or multiple cancer prognostic biomarkers and develop prognostic models which are applicable across a number of cancer types. By means of a comprehensive evaluation of prognostic potential of expression of apoptotic pathway genes, we were able to catalogue survival associated genes in 30 cancers. Based on this, risk prediction models were first developed for each cancer-type. Henceforth, we were able to identify a gene signature that is significantly associated with survival risk across 27 cancers. Briefly, this study finds its significance by proposing a versatile prognostic method for a large number of cancer types complemented by a computational strategy for devising similar methods based on a novel clustering-based approach. A practical realisation of this study can be monumental in cancer care and therapeutic management.

mathematical models for risk assessment [14–16]. Apoptosis related molecules have also been identified to play prognostic roles in other cancers such as identification of survival related proteins AGR2, ENO1, GDI2, GRP78, GRP94, PPIA, PRDX1, PTEN, and gene KIF15 by means of immunohistochemical staining in gastric cancer [17,18]; high protein and mRNA levels of BIK as markers of tumour recurrence in breast cancer [19]; gene expression of MCL1 as prognostic biomarker in Non-small cell lung cancer [20]; high expression of TOP2A was shown to be linked with poor survival in bladder urothelial carcinoma [21]; a seven gene signature was established to distinguish patients with isocitrate-wildtype glioblastoma [22] and increased FER1L4 expression was found to promote apoptosis and suppress epithelial-mesenchymal transition in osteosarcoma [23]. In thyroid cancer, alterations in apoptotic molecules such as p53, BCL2, BCL-XL, BAX, p73, Fas/FasL, PPARG, TGFb, and NFKb have also been associated with carcinogenesis [24]. However, the scope of these studies was limited to specific cancers with a limited set of genes/proteins. Since apoptosis consists of a large number of regulatory genes/proteins, gauging the prognostic significance of maximum number of genes/proteins involved in apoptosis across several cancers can offer a better understanding. It can also reveal several novel targets and help in development of finer biomarkers for cancer prognosis.

In this study, we utilized the gene-expression data of 33 cancer cohorts from The Cancer Genome Atlas (TCGA) and evaluated the prognostic performance of 165 genes involved in the apoptosis pathway. Apart from identifying cancer-specific prognostic genes, we also constructed multiple gene-based risk prediction models based on the overall survival of patients. Firstly, we developed universal biomarkers which can be used for prognosis across 33 cancers. Secondly, for individual cancer cohorts, we were able to develop risk assessment methods, some of which are superior to previously suggested biomarkers/methods. Thirdly, we developed an 11-gene multi-cancer prognostic biomarker that can be used across 27 cancers. Finally, we propose a strategy for constructing cross-cancer prognostic biomarkers by means of hierarchical clustering of cancers on the basis of common prognostic genes. We show the efficacy of this strategy by developing a 15 gene biomarker applicable for both brain low grade glioblastoma (LGG) and kidney renal cell carcinoma (KIRC) patient prognosis.

2 | METHODS

2.1 | Data collection and preprocessing

Normalized gene expression datasets '.rsem.genes.normalized_results' and raw counts '.rsem.genes.results' for 33 cancer cohorts were obtained from 'The Cancer Genome Atlas' (TCGA) using TCGA Assembler-2 [25]. A '*pan-cancer*' dataset was formed by combining all the samples with raw expression values of genes across 33 cancers and normalizing them using 75th percentile normalization. A list of 165 apoptosis genes was obtained from [26]. The gene expression data for these 165 genes were extracted from the downloaded TCGA cancer datasets and *pan-cancer* dataset. Only those patient samples were retained in all the datasets for whom overall-survival and censoring information were available. The number of samples in *pan-cancer* dataset was 9569, while the number of samples in each cancer cohort, N, is mentioned in Table 2. TCGA abbreviations for cancers are provided in S2 Table 1.

2.2 | Survival analysis

Univariate unadjusted Cox proportional hazards (Cox-PH) regression models were used to screen survival-associated genes from their expression data using the formula:

$$h\,(G, t) \; = h_0\,(t)\,exp\,(\beta G) \qquad (1)$$

where the variable G is the expression data of a gene, h is the hazard function and the variable t is the overall-survival time. The Cox regression coefficient, $\beta > 0$ signifies that the elevated gene expression is unfavourable while inverse applies for $\beta < 0$. R packages 'survival' and 'survminer' were used to implement the Cox-PH models. Using these, Hazard ratios (HR) were computed along-with confidence intervals (%95 CI) and *p*-values. HR is the ratio of hazard rates representing the death risk associated with one group as compared with another by using an appropriate cut-off of gene-expression. Based on its definition, HR > 1 implies the increase in death risk when the expression of the gene is increased, and vice-versa for HR < 1. However, the expression of gene has no effect on survival if HR = 1. For comparison of survival curves between two risk groups, we used Kaplan-Meier (KM) plots and log-rank tests. Survival associated genes were identified with HR greater than or less than 1 and $p < 0.05$. The metric Concordance index (C) was used to evaluate the model's predictive performance [27–29].

2.3 | Prognostic index (PI)

As implemented in [30–33], for *n* genes, $g_1, g_2, \ldots g_n$ with cox coefficients $\beta_1, \beta_2 \ldots \beta_n$ obtained from the univariate Cox-PH analyses using median cut-offs, Prognostic Index (PI) was defined as:

$$PI \; = \; B.g \qquad (2)$$

where $g = [g_1\ g_2\ g_3\ \ldots.\ g_n]$ and $B = [\beta_1\beta_2\beta_3\ \ldots\ \beta_n]$. Thereafter, risk groups were segregated by using univariate Cox-PH regression model, as defined earlier with G replaced by PI. The cut-off value for PI was evaluated using cutp from 'survMisc' package in R. Model's performance is estimated using HR, *p*, %95 CI and C values.

2.4 | Voting model

As implemented in [34], for an *n*-gene voting model, a *n*-bit vector is assigned to each patient sample. Thereafter, each bit is labelled as high or low risk on the basis of corresponding classification by individual genes, using Cox-PH univariate models. Finally, the sample is allotted an overall risk label decided by majority of the labelled bits (i.e., greater than *n*/2 labels).

2.5 | Gene ontology (GO) functional enrichment analysis

A GO enrichment analysis was performed for finding the molecular function (MF) associated with survival associated genes in specific cancers. The STRING application programming interface (API) for GO functional enrichment was implemented in python and used for performing this analysis (https://string-db.org/help/api/). Thereafter, the GO MF enriched terms were analysed across different cancers to find out overlapping terms.

2.6 | Screening of candidate therapeutic molecules

The univariate Cox-PH survival analysis provides information about the risk associated with an increase in the expression of specific genes. These genes can be categorized into two groups: genes whose elevated expression is a risk factor and genes whose elevated expression is favourable for patient survival. The modulation of the expression profile of these genes can be a potential therapy for cancer patients. For this, we employ the Connectivity Map2 database [35,36], which provides a list of enriched small molecules based on the list of upregulated or downregulated genes. The predicted molecules are ranked based on enrichment values and can act as potential therapeutic candidates for altering the gene expression.

3 | RESULTS

3.1 | Survival-risk prediction for all types of cancer patients

The dataset comprising of patient samples belonging to all cancer types (pan-cancer) was utilized here to develop risk prediction models which can be applicable across all the cancers. Top survival associated genes were first screened and after that used to develop models which stratify the patients into High or Low risk groups. Apart from risk prediction, survival probabilities were also estimated using the KM plots. In order to screen top genes, two different approaches were implemented: (i) correlation analysis-based approach, and (ii) Cox-PH survival analysis-based approach. Details are provided below.

3.1.1 | Correlation based universal prognostic biomarkers

Pearson correlation coefficient (R) was used to determine the relationship of the expression of 165 apoptotic genes with OS in the pan-cancer dataset. It was observed that gene expression was weakly related to OS when the pan-cancer dataset was used. Figure 1A shows the genes having $|R|> 0.04$, LEF1 being the top correlated gene with $R = 0.077$ (see S1 Table 1 for complete results). Figure 1B shows the histogram plot for the correlation analysis. Most of the genes were found to be negatively related to OS implying the inverse role of their expression with OS. For example, CASP7 and BMP2 upregulation was observed to weakly reduce OS of the patients, while LEF1 expression showed an opposite nature. Further, risk stratification models based on PI and voting methods were developed for the top five correlated genes (LEF1, CASP7, BMP2, LPPR4, and ANXA1). The PI model defined as, $PI_{corr} = 0.098 \times LEF1 + 0.182 \times CASP7 + 0.333 \times BMP2 - 0.257 \times LPPR4 + 0.465 \times ANXA1$, was able to stratify high and low risk patients with HR = 1.67, *p*-value = 9.75×10^{-33}, C = 0.58, %95CI 1.54–1.82 and logrank-*p* = 2.85×10^{-32}. Whereas, voting model performed the second best with HR = 1.60, *p*-value = 2.67×10^{-26}, C = 0.56, %95CI 1.46–1.74 and logrank-*p* = 2.29×10^{-25}. KM plots for these models are shown in Figure 1C and Figure 1D.

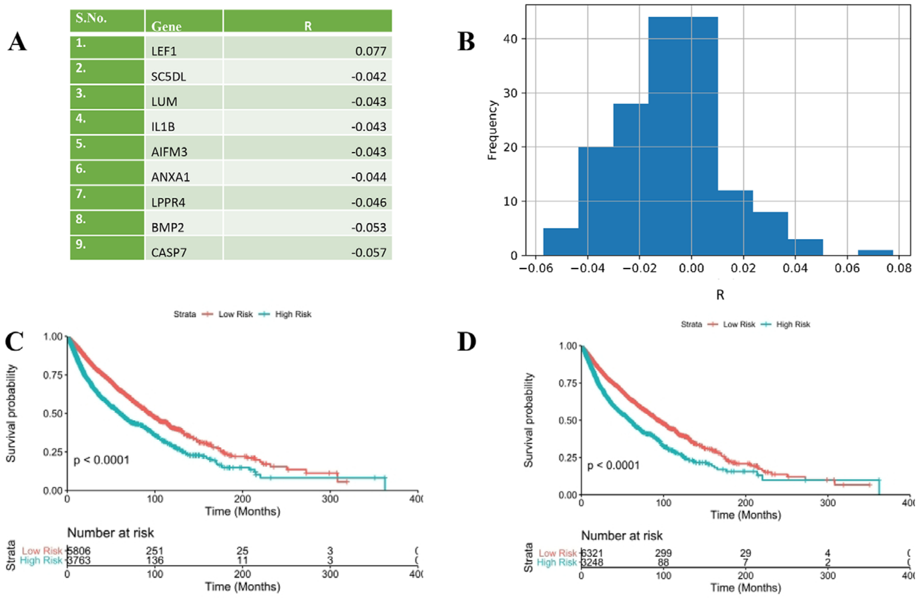
TABLE 1 Top survival associated genes in the *pan-cancer* cohort

S.no.	Gene	β	HR	*p*-value	1/HR	C
1.	EREG	0.68	1.98	1.69×10^{-54}	0.50	0.60
2.	IL1A	0.65	1.91	3.41×10^{-49}	0.52	0.59
3.	IL18	0.51	1.66	5.02×10^{-31}	0.60	0.58
4.	BAK1	0.51	1.66	1.03×10^{-30}	0.60	0.58
5.	BID	0.50	1.65	3.44×10^{-30}	0.61	0.58
6.	CDC25B	0.50	1.64	1.10×10^{-29}	0.61	0.58
7.	IL1B	0.49	1.63	4.70×10^{-29}	0.61	0.58
8.	ANXA1	0.46	1.59	3.16×10^{-26}	0.63	0.57
9.	TOP2A	0.46	1.59	4.32×10^{-26}	0.63	0.58
10.	BRCA1	0.46	1.58	1.94×10^{-25}	0.63	0.58

HR: Hazard Ratio, C: Concordance Index, β: Cox regression coefficient. Cox-PH univariate regression analysis with >median expression cut-off was used to estimate survival risk associated with the two risk groups, for each gene.

3.1.2 | Cox-PH survival analysis based universal prognostic biomarkers

A univariate unadjusted Cox-PH analysis was performed on the pan-cancer dataset. For each gene out of 165 genes, median expression cut-off was used to segregate high and low risk patients. 119 genes were found to be significantly related with OS ($p < 0.05$), with HR ranging from 0.63–1.98 (C values 0.52–0.56). Expression of 24 genes was related to overall good prognosis, while expression of remaining 95 genes was related with overall poor prognosis of cancer patients (S1 Table 1). Table 1 shows the results for the top ten genes based on *p*-values. The top genes were observed to be bad prognostic biomarkers that is, their upregulated expression was linked with an unfavourable survival in cancer patients. The next step was to develop universal



S.No.	Gene	R
1.	LEF1	0.077
2.	SC5DL	-0.042
3.	LUM	-0.043
4.	IL1B	-0.043
5.	AIFM3	-0.043
6.	ANXA1	-0.044
7.	LPPR4	-0.046
8.	BMP2	-0.053
9.	CASP7	-0.057

FIGURE 1 Correlation based universal prognostic biomarkers. (A) Top genes whose expression is correlated with OS having absolute Pearson correlation coefficient, |R| > 0.04. (B) Histogram plot showing number of genes with respect to the correlation coefficient (R). (C) KM plot representing risk stratification using $\mathrm{PI}_{\mathrm{corr}}$ (HR = 1.67, *p*-value = 9.75×10^{-33}). (D) KM plot representing risk stratification by voting model (HR = 1.60, *p*-value = 2.67×10^{-26})

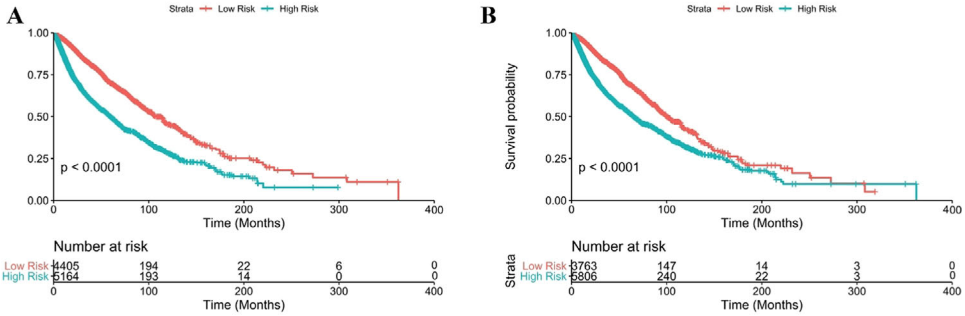


FIGURE 2 Universal prognostic biomarkers based on survival associated genes. (A) KM plot representing risk stratification by voting model (HR = 2.14, *p*-value = 7.81×10^{-62}). (B) KM plot representing risk stratification using $\mathrm{PI}_{\mathrm{surv}}$ (HR = 1.87, *p*-value = 1.29×10^{-39})

prognostic biomarkers using the expression profile of these genes, which could be used for all the cancers. For this, we constructed PI and voting models. Here, PI denoted as $\mathrm{PI}_{\mathrm{surv}}$ was able to stratify high and low risk patients with HR = 1.87, *p*-value = 1.29×10^{-39}, C = 0.59, %95CI 1.71–2.06 and logrank-*p* = 1.18×10^{-42}. The voting model, on the other hand, was able to stratify high and low risk patients with HR = 2.14, *p*-value = 7.81×10^{-62}, C = 0.62, %95CI 1.96–2.34 and logrank-*p* = 6.41×10^{-66}. Figure 2 shows the KM plots corresponding to these.

Amongst these genes, expression of EREG has been previously linked with tumour progression, metastasis and gastric cancer prognosis [37], IL18 was found to be a prognostic biomarker in melanoma patients with strong correlation with CD8^{+}T and NK cell infiltration [38], BAK1 expression was recently found to be a prognosis predictor in chemotherapy treated gastric cancer patients [39], role of IL1B as prognostic biomarker and therapeutic target for multiple sclerosis patients was shown in [40], ANXA1 was found to act as tumour suppressor in HNSC [41], TOP2A expression was related to prognosis of BLCA patients [21] and while, the role of the mutations of DNA repair gene BRCA1 in breast cancer is very well established, it has recently been observed to play a prognostic role in prostate cancer [42].

3.2 | Survival-risk prediction and analysis for different types of cancer patients

Cox-PH survival analysis was utilized to identify prognostic biomarker genes that were significantly associated with overall patient survival in different cancer cohorts. After the screening of biomarker genes, risk prediction models were developed for each cancer type. We also performed a GO enrichment analysis to find out molecular functions associated with top prognostic biomarker genes in each cancer and their overlap across different cancers. Details are provided below.

3.2.1 | Identification of prognostic biomarker genes

Univariate Cox-PH survival analysis was performed for 165 genes using each cancer's dataset. Genes were classified as good prognostic marker (GPM) or bad prognostic marker (BPM) on the basis of whether a gene's expression is positively correlated with OS (GPM) or negatively correlated with OS (BPM). BPM genes are identified with an HR > 1, while GPM genes have an HR < 1. For BPM genes, patients with gene expression more than the median gene expression were classified as high risk and patients with gene expression less than/equal to the median gene expression as low risk. The inverse applies for a GPM gene. S1 Table 2 shows the results of survival analysis for each of 165 genes across 33 cancers (HR values and p-values), while S1 Table 3 shows the cancers vs. genes matrix with HR values of only significant genes ($p < 0.05$). Table 2 shows the number of survival associated genes for each cancer among other details. It is seen that in most of cancers BPM genes are more than GPM genes, showing the detrimental role of the upregulated expression of some apoptotic genes in cancer. Table 2 also mentions the top genes (at most ten) for each cancer on the basis of p-values obtained from univariate survival analysis. None of the 165 genes was significantly associated with survival in three cancers: DLBC, TCGT, and PCPG.

3.2.2 | Cancer specific risk prediction models

The top genes mentioned in Table 2 were used to construct models for risk stratification in 30 cancers, excluding TCGT, PCPG, and DLBC. Both gene voting based models and PI models were used to segregate patients into risk groups. HR, p-values and C index were then calculated. Voting models showed the best results and are shown in Table 3. S1 Table 4 shows the results for PI models. For the case of PRAD and READ (two genes each), a tie case was considered as High Risk.

3.2.3 | Gene Ontology (GO) functional enrichment analysis

We performed a GO functional enrichment for finding out the top molecular function (most significant p-value) in the case of these cancers for top genes. Figure 3A shows the results for this. Figure 3B shows the distribution of cancers enriched to each function. We find that the molecular function 'enzyme binding' was enriched in most of the cancers viz. ACC, CESC, LUSC, SARC, STAD, and UVM. Amongst these, CESC and LUSC also have 'enzyme binding' as their top specific enriched function with $p{\sim}10^{-5}$. There was a total of 26 genes from the apoptotic pathway related to this common function. S2 Table 2 shows these genes and the PMIDs of the studies that relate them with different cancers. The analysis was done to see which are the underlying molecular functions where these prognostic genes are involved. 'enzyme binding' was the most common function amongst cancers.

3.3 | Universal model for risk stratification of different types of cancer patients

Heatmap S2 Figure 1 shows all the survival genes in 30 cancers. We found that there are 11 genes that play a prognostic role in more than or equal to eight cancers (in at least 25% cancers). S2 Table 3 shows the list of genes. It mentions the HR range, top enriched molecular function (GO) and PMIDs related to studies that mention the association of these genes in the context of cancer using Candidate Cancer Gene Database [43]. Figure 4A shows the role of these genes as BPM or GPM in different cancers. Figure 4B shows the 27 cancers associated with these genes. Most of the genes play a BPM role that is, their elevated expression prevents cellular apoptosis and thus promotes tumour progression (High Risk patients). Amongst these, EREG was found to be a BPM in all the associated cancers. The drug gene interaction database (DGIdB) predicts anti-EGFR therapeutic Panitumumab as the most interactable drug molecule for EREG [44]. EREG expression has been known to be linked with advanced stage cancers such as metastatic colorectal cancer. EREG is a known EGFR ligand which promotes tumour cell proliferation and differentiation and thus used as both predictive biomarker for anti-EGFR therapy as well as a potential target [45]. SATB1 on the other hand plays a GPM role in majority of the cancers it is associated with that is, its high expression is linked with better overall survival. Whereas, CASP2 plays both kind of roles. Prognostic index (PI) and voting models were constructed using the multi-cancer genes (11 gene panel) in 27 cancers. Results for voting models are shown in Table 4 and PI models are given in S1 Table 5. This universal model performed best in UVM, THYM, PRAD, KICH, and ACC based on HR and C index, where it can be readily used as a single prognostic test. Though in other cancers, the risk prediction performance of this 11 gene panel was moderate (THCA, UCEC, and PAAD) to poor and thus for them, cancer specific prognostic biomarkers should be relied on for a better risk prognosis.

3.4 | External validation of the universal prognostic model

The evaluation of the performance of the universal model on external cohorts is necessary for its practical translation. Therefore, we assessed the prognostic strength of the obtained eleven gene signature on various datasets. We utilized a specialized tool, 'SurvExpress', developed for the validation of biomarker on multiple cancer types [46]. Table 5 shows the validation of expression profile of EREG and SATB1 in different cancers in the context of their respective negative and positive correlation with overall survival. As depicted, the prognostic role of EREG as BPM and SATB1 as GPM as identified in our analysis is strengthened following these validation results. Furthermore, SurvExpress constructed a prognostic index-based model of the 11 genes that were provided. Table 6 represents the result of the universal model on eight different cancer cohorts. The cohorts for which the expression data was unavailable were rejected for the analysis. As observed from the results the universal model performed

TABLE 2 The table shows the no. of patient samples (N), no. of BPM and GPM genes and top ten survival associated genes for 33 cancers

Cancer	N	BPM	GPM	Total	Top Genes
LGG	511	77	17	94	WEE1, BTG3, BMP2, PLAT, SMAD7, ANXA1, PEA15, CDK2, HSPB1, SOD2
KIRC	532	50	32	82	CASP9, F2, TIMP1, IL6, CDC25B, ADD1, CCNA1, BAK1, SLC20A1, TIMP3
MESO	86	33	15	48	HMGB2, TOP2A, BRCA1, PLAT, SLC20A1, WEE1, PPP2R5B, MADD, PDC, 4, LMNA
SKCM	449	10	33	43	TNFSF10, SATB1, DPYD, BIRC3, SOD2, F2R, CYLD, GCH1, CD69, PSEN2
PAAD	178	34	7	41	CASP4, TNFSF10, PSEN1, CD44, CASP2, EMP1, TOP2A, DPYD, CCND1, MGB2
ACC	79	22	14	36	TOP2A, PEA15, BRCA1, H1F0, HMGB2, MADD, CDK2, SPTAN1, CYLD, S, STM1
BRCA	1091	10	25	35	PTK2, NEFH, IGF2R, PLAT, DNM1L, XIAP, ETF1, NEDD9, IRF1, RARA
LAML	173	14	16	30	PDCD4, ISG20, LMNA, NEDD9, CCND2, PSEN1, HGF, SOD1, ADD1, CD44
HNSC	519	19	10	29	CCND1, BMF, CCNA1, BAK1, PSEN1, APP, TIMP1, BCAP31, SLC20A1, TN, RSF12A
UVM	80	17	12	29	ERBB3, ISG20, EREG, TIMP3, LEF1, SATB1, TXNIP, PPP2R5B, ERBB2, PT2
CESC	304	16	10	26	EREG, CASP2, MGMT, CD2, IL1B, IGF2R, APP, NEFH, TIMP2, GCH1
KIRP	287	21	3	24	BCL2L10, TOP2A, PMAIP1, MCL1, LEF1, PPP2R5B, PEA15, DCN, IRF1, H10
SARC	257	7	16	23	CTH, RNASEL, GSN, IRF1, SPTAN1, CASP1, BTG2, CFLAR, TNF, CASP2
BLCA	404	7	15	22	EMP1, GCH1, HMGB2, GSTM1, CASP7, ANXA1, IFNGR1, ETF1, SLC20A1, AIFM3
LIHC	369	12	4	16	MGMT, ETF1, RARA, GPX3, EREG, CD2, DAP3, GPX4, FASLG, CDC25B
STAD	413	13	3	16	CAV1, CD44, PDGFRB, DNAJC3, EREG, TGFB2, CTNNB1, DFFA, BCL2L11, CASP6
LUSC	488	12	3	15	CD14, BTG3, EREG, CCND2, PTK2, PAK1, ADD1, HSPB1, TIMP3, SMAD7
LUAD	497	9	5	14	EREG, VDAC2, BBC3, SLC20A1, BTG2, TOP2A, RELA, CD2, GPX4, ETF1
ESCA	183	6	7	13	ENO2, IL18, TOP2A, DAP, BCL2L1, PMAIP1, ISG20, IL1A, TSPO, SATB1
COAD	297	5	5	10	BCL10, CASP4, FAS, IL6, GSR, TIMP1, BGN, LUM, ERBB2, BTG2
OV	305	4	5	9	DAP, CASP8, EMP1, BIRC3, CASP2, WEE1, PSEN1, NEDD9, SOD1
THCA	505	4	5	9	ANXA1, TGFBR3, CLU, PSEN1, TNFRSF12A, GPX4, TIMP3, LEF1, BNIP3L
KICH	65	6	2	8	IFNB1, MADD, BIK, GSR, TOP2A, PTK2, DAP3, CLU
GBM	160	6	1	7	HSPB1, FDXR, TXNIP, ANKH, EGR3, F2R, IER3
UCEC	541	5	0	5	BCL2L1, MCL1, AVPR1A, SLC20A1, ISG20
UCS	57	2	3	5	MGMT, HGF, BMF, H1F0, PTK2
CHOL	36	3	1	4	PSEN1, BNIP3L, EREG, JUN
THYM	119	2	2	4	IER3, SOD2, CD2, LEF1
PRAD	497	1	1	2	SATB1, IER3
READ	96	1	1	2	BRCA1, DNAJC3
DLBC	47	0	0	0	-
PCPG	179	0	0	0	-
TGCT	133	0	0	0	-

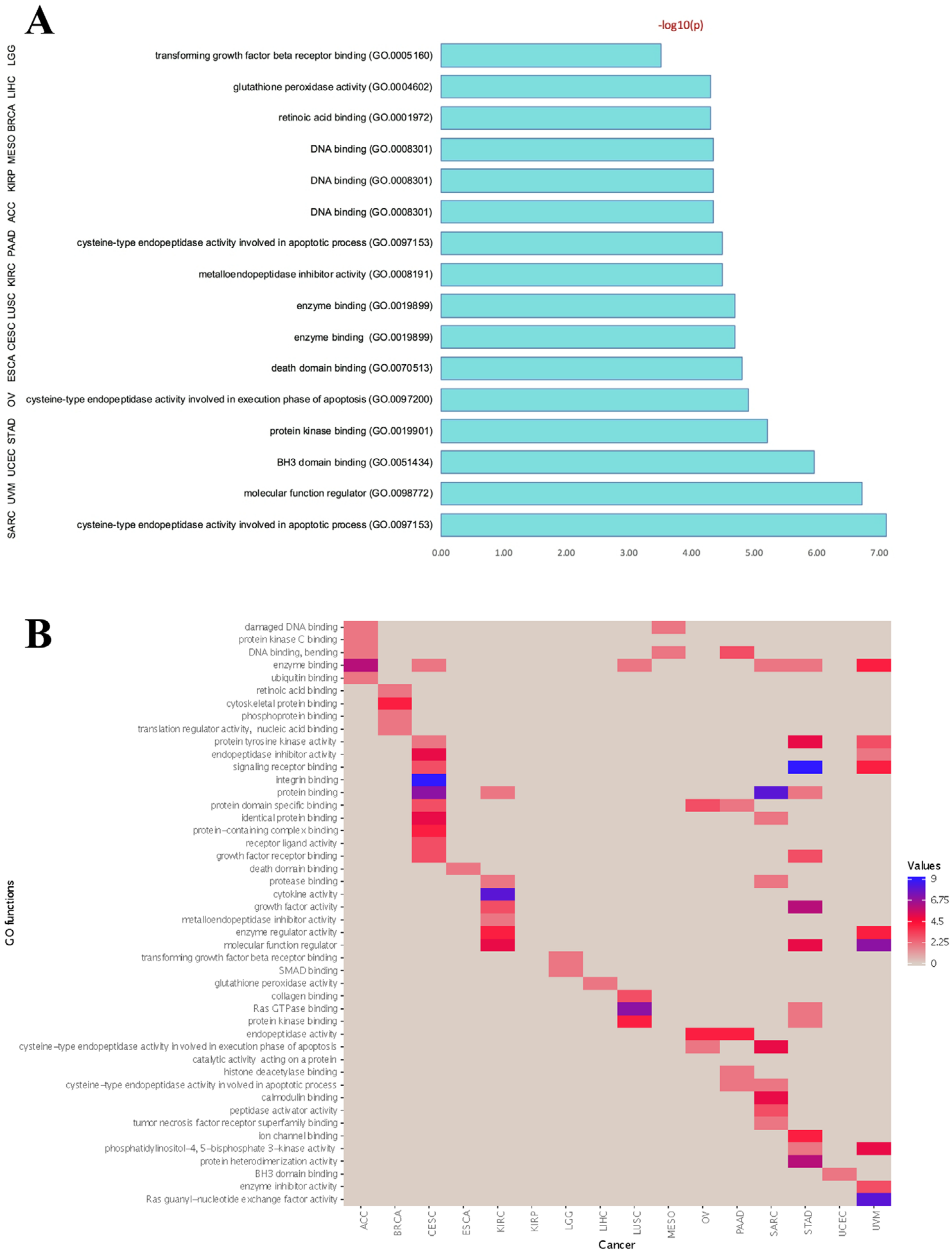


FIGURE 3 GO enrichment analysis in individual cancer cohorts. (A) The top enriched GO molecular function for each cancer corresponding to top genes. x-axis is the -$\log_{10}$ (*p*-value) and y corresponds to the enriched function corresponding to the cancer. (B) Heatmap showing enriched GO molecular functions by top genes for each cancer. Number of genes is encoded by different colours

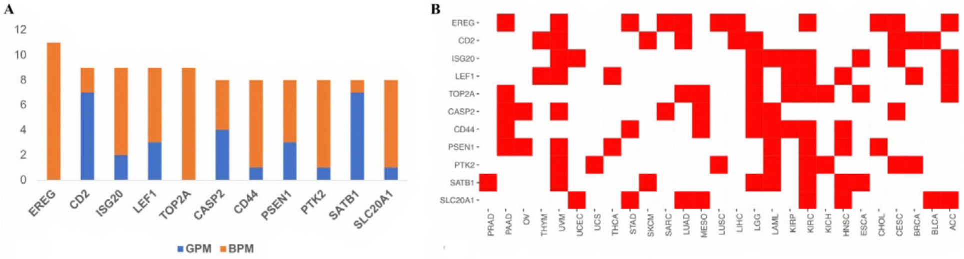


FIGURE 4 Multi-cancer survival genes. (A) Shows the distribution of role of each of these 11 genes across 27 cancers. y-axis shows the number of cancers in which the corresponding gene plays prognostic role. (B) Red blocks indicate that the survival genes associated with the cancer

TABLE 3 The performance of prognostic gene-voting models for each cancer

Cancer	HR	*p*-value	logrank-*p*	C	%95 CI L	%95 CI U
THCA	41.59	3.36×10^{-4}	3.81×10^{-8}	0.84	5.42	319.17
UVM	40.50	5.32×10^{-4}	5.12×10^{-7}	0.85	4.99	328.82
KICH	25.61	2.27×10^{-3}	3.53×10^{-5}	0.83	3.19	205.6
ACC	22.68	7.95×10^{-7}	1.63×10^{-10}	0.81	6.57	78.31
THYM	12.53	2.42×10^{-2}	6.98×10^{-3}	0.79	1.39	112.93
UCEC	10.42	4.51×10^{-4}	1.13×10^{-4}	0.7	2.81	38.6
CHOL	8.72	4.75×10^{-4}	2.45×10^{-4}	0.77	2.59	29.4
PRAD	8.42	4.41×10^{-3}	4.20×10^{-3}	0.65	1.94	36.5
READ[a]	7.45	6.50×10^{-2}	2.56×10^{-2}	0.72	0.88	62.93
KIRP	5.10	6.64×10^{-5}	1.27×10^{-5}	0.72	2.29	11.37
LGG	4.99	2.88×10^{-12}	1.54×10^{-13}	0.72	3.18	7.83
CESC	4.92	2.14×10^{-8}	2.98×10^{-9}	0.71	2.82	8.6
LIHC	4.58	7.91×10^{-11}	2.24×10^{-11}	0.7	2.89	7.24
PAAD	4.41	4.23×10^{-7}	1.72×10^{-7}	0.69	2.48	7.85
COAD	4.08	5.05×10^{-5}	2.42×10^{-5}	0.67	2.07	8.05
MESO	3.99	1.67×10^{-6}	2.00×10^{-6}	0.68	2.26	7.03
KIRC	3.96	5.41×10^{-16}	3.03×10^{-17}	0.68	2.84	5.53
LAML	3.96	3.92×10^{-12}	5.07×10^{-12}	0.67	2.68	5.84
ESCA	3.80	2.19×10^{-6}	3.32×10^{-6}	0.65	2.19	6.61
UCS	3.61	8.77×10^{-4}	6.13×10^{-4}	0.68	1.69	7.67
BRCA	3.45	2.36×10^{-9}	6.76×10^{-10}	0.67	2.3	5.18
BLCA	3.41	6.35×10^{-10}	3.51×10^{-10}	0.66	2.31	5.02
STAD	3.35	2.78×10^{-7}	1.39×10^{-7}	0.64	2.11	5.31
SARC	2.81	1.32×10^{-5}	1.03×10^{-5}	0.67	1.77	4.48
LUAD	2.76	6.94×10^{-8}	4.82×10^{-8}	0.63	1.91	3.99
HNSC	2.36	9.24×10^{-8}	5.80×10^{-8}	0.62	1.72	3.24
LUSC	2.21	1.26×10^{-6}	1.30×10^{-6}	0.61	1.6	3.04
OV	2.19	1.38×10^{-6}	1.16×10^{-6}	0.61	1.59	3
GBM	2.07	3.73×10^{-4}	3.22×10^{-4}	0.61	1.38	3.09
SKCM	1.99	2.18×10^{-5}	2.55×10^{-5}	0.59	1.45	2.75

[a] Statistically insignificant, HR: Hazard Ratio, C: Concordance Index.

best for prostate cancer (HR = 5.88) which is in corroboration with its performance on TCGA PAAD dataset (HR = 4.49). The model is also seen to perform significantly in a variety of cancer types such as kidney cancer, ovarian cancer, lung cancer and so on. thereby strengthening its employability as a multi-cancer risk prediction model.

3.5 | Clustering of different type of cancers

It is interesting to find out which genes are shared across cancers in the context of their association with patient overall survival. Firstly, it will help in development of prognostic biomarkers applicable across cancers. Further, it will also elucidate the key apoptotic genes whose altered expression is linked with carcinogenesis in those cancers. The identification of such genes could motivate future experimental efforts in the direction of understanding their fundamental biological roles as well as designing better therapeutic strategies. To do this, we find out pairwise similarity between cancers c1 and c2 using Jaccard similarity index defined as:

$$J\ (c1,\ c2) = \frac{|c1\ \cap c2|}{|c1\ \cup\ c2|} \tag{3}$$

where c1and c2 represent the set of genes that are associated with survival in cancer c1 and cancer c2, respectively. The similarity matrix thus obtained is converted to distance matrix, following which UPGMA (unweighted pair group method with arithmetic mean) is used for hierarchical clustering of cancers. Figure 5 shows the dendrograms

TABLE 4 Risk stratification using 11 gene voting based universal prognostic model

Cancer	HR	*p*-value	logrank-*p*	C	%95 CI L	%95 CI U
UVM	11.74	1.80×10^{-3}	1.77×10^{-4}	0.71	2.50	55.17
THYM	10.12	4.07×10^{-2}	1.48×10^{-2}	0.77	1.10	92.91
PRAD	8.94	4.07×10^{-2}	1.01×10^{-2}	0.62	1.10	72.80
KICH	7.41	1.27×10^{-2}	4.98×10^{-3}	0.72	1.53	35.75
ACC	7.37	3.09×10^{-5}	3.77×10^{-6}	0.73	2.88	18.86
THCA	4.81	4.94×10^{-3}	3.93×10^{-3}	0.74	1.61	14.37
UCEC	4.49	1.04×10^{-2}	1.07×10^{-2}	0.64	1.42	14.18
PAAD	4.17	4.21×10^{-6}	7.61×10^{-7}	0.69	2.27	7.65
MESO	3.45	1.29×10^{-5}	1.75×10^{-5}	0.65	1.98	6.02
CHOL	3.22	4.15×10^{-2}	4.89×10^{-2}	0.65	1.05	9.91
CESC	2.93	1.96×10^{-4}	8.11×10^{-5}	0.65	1.67	5.17
KIRP	2.93	2.85×10^{-3}	2.58×10^{-3}	0.65	1.45	5.95
LIHC	2.92	6.27×10^{-6}	2.38×10^{-5}	0.61	1.83	4.65
KIRC	2.87	1.14×10^{-9}	1.55×10^{-10}	0.63	2.04	4.03
LGG	2.75	6.37×10^{-6}	2.69×10^{-6}	0.66	1.77	4.26
LUAD	2.47	9.02×10^{-7}	1.09×10^{-6}	0.63	1.72	3.54
BLCA	2.38	2.11×10^{-5}	5.74×10^{-5}	0.59	1.59	3.54
STAD	2.28	1.06×10^{-3}	5.75×10^{-4}	0.61	1.39	3.73
UCS	2.19	4.14×10^{-2}	3.72×10^{-2}	0.58	1.03	4.66
SKCM	2.07	4.19×10^{-5}	9.45×10^{-5}	0.58	1.46	2.93
ESCA	2.03	9.00×10^{-3}	8.56×10^{-3}	0.60	1.19	3.45
HNSC	1.95	3.19×10^{-5}	2.49×10^{-5}	0.60	1.42	2.67
BRCA	1.90	2.08×10^{-3}	1.57×10^{-3}	0.61	1.26	2.86
SARC	1.72	3.16×10^{-2}	3.86×10^{-2}	0.56	1.05	2.81
LAML	1.68	6.55×10^{-3}	7.46×10^{-3}	0.58	1.16	2.45
LUSC	1.59	8.86×10^{-3}	1.12×10^{-2}	0.54	1.12	2.25
OV	1.53	1.51×10^{-2}	1.83×10^{-2}	0.53	1.09	2.17

HR: Hazard Ratio, C: Concordance Index.

TABLE 5 External validation of the top BPM (EREG) and GPM (SATB1) genes in external cohorts

S.no.	Dataset	Cancer	HR	*p*-value	C	%95CI	logrank-*p*
	EREG						
1	PACA-AU - ICGC	Pancreatic	1.47	0.04	55.7	1–2.15	0.04
2	Bild Nevins lung GSE3141	Lung	1.79	0.02	58.66	1.06–3.04	0.02
3	Rousseaux GSE30219	Lung	1.45	0.01	53.77	1.06–1.97	0.01
4	Zhao renal kidney GSE3538	Kidney	1.61	0.02	56.6	1.05–2.46	0.02
5	Chibon F Sarcoma GSE21050	Sarcoma	1.04	0.84	51.06	0.72–1.49	0.83
	SATB1						
1	Zhao renal kidney GSE3538	Kidney	0.63	0.03	57.73	0.41–0.97	0.03
2	Rao Giddings esophagus GSE11595	Esophageal	0.30	0.01	65.96	0.11–0.79	0.01

HR: Hazard Ratio, C: Concordance Index.

TABLE 6 External validation of Universal prognostic model in eight cancer cohorts

S.no.	Dataset/GEO accession	HR'	HR	*p*-value	C	%95CI	logrank-p
1	Zhao Renal Kidney GSE3538	2.87	3.03	1.84×10^{-6}	0.69	1.92–4.79	4.82×10^{-7}
2	Tothill Bowtell Survival Ovarian GSE9891	1.53	3.97	2.17×10^{-10}	0.76	2.6–6.09	6.15×10^{-12}
3	OV-AU - ICGC Ovarian Cancer - Serous cystadenocarcinoma	1.53	2.4	6.01×10^{-4}	0.65	1.46–3.96	4.20×10^{-4}
4	Gulzar-Prostate-GSE40272	8.94	5.88	1.20×10^{-3}	0.84	2.01–17.24	1.80×10^{-4}
5	Tomida lung GSE13213	2.47	3.97	2.41×10^{-5}	0.75	2.09–7.54	5.43×10^{-6}
6	Hoshida Golub liver GSE10186	2.92	2.46	1.70×10^{-2}	0.65	1.17–5.15	1.40×10^{-2}
7	PACA-AU - ICGC - Pancreatic cancer - Ductal adenocarcinoma	4.17	2.59	2.05×10^{-6}	0.67	1.75–3.84	8.67×10^{-7}
8	Peters C.Fitzgerald Esophagus GSE19417	2.03	2.8	1.00×10^{-4}	0.64	1.67–4.72	5.59×10^{-5}

HR: Hazard Ratio, C: Concordance Index, β: Cox regression coefficient.
The reference column HR' denotes the model's performance in TCGA datasets for respective cancer type followed by the columns showing validation performance.

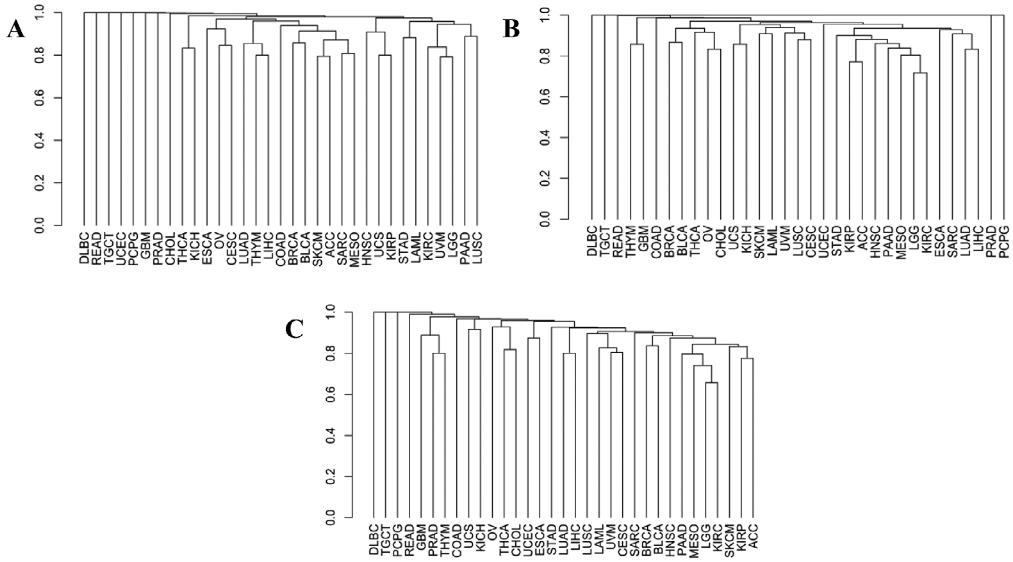


FIGURE 5 Hierarchical clustering of cancers based on (A) shared GPM genes, (B) shared BPM genes, and (C) all shared survival related genes

representing hierarchical clustering plots on the basis of shared GPM genes, shared BPM genes and shared total survival genes (both BPM and GPM). S1 Tables 6–8 highlight the number of shared genes in the form of matrices and Table S9 shows the Jaccard indices J_{GPM}, J_{BPM} and J_{all} evaluated to create dendrograms.

3.6 | Case study: Cross-cancer prognostic biomarkers (LGG vs. KIRC)

Based on the Jaccard similarity index, $J_{all} = 0.34$, LGG-KIRC pair was found to be most similar in the context of survival related genes. The list of genes is shown in S1 Table 10 along with HR and *p*-values obtained from Cox univariate analyses. An intersection between the set of top 20 genes (based on *p*-values) of both the cancers was used to develop risk stratification models. The conjoined set consisted of 15 genes viz. BTG3, CDK2, SOD2, TOP2A, HMGB2, TIMP1, ISG20, TNFRSF12A, AFNB1, ADD1, CASP8, CDC25B, IFITM3, CD44, and GPX1. PI models were developed for both the cancers as follows:

$$\begin{aligned} PI_{LGG} = {} & 1.19 \times BTG3 + 1.07 \times CDK2 + 0.99 \times SOD2 + 1.07 \\ & \times TOP2A + 0.99 \times HMGB2 + 0.98 \times TIMP1 + 0.89 \\ & \times ISG20 + 0.91 \times TNFRSF12A + 0.91 \times IFNB1 - 0.81 \\ & \times ADD1 + 0.79 \times CASP8 + 0.77 \times CDC25B + 0.76 \\ & \times IFITM3 + 0.74 \times CD44 + 0.74 \times GPX1 \end{aligned} \tag{4}$$

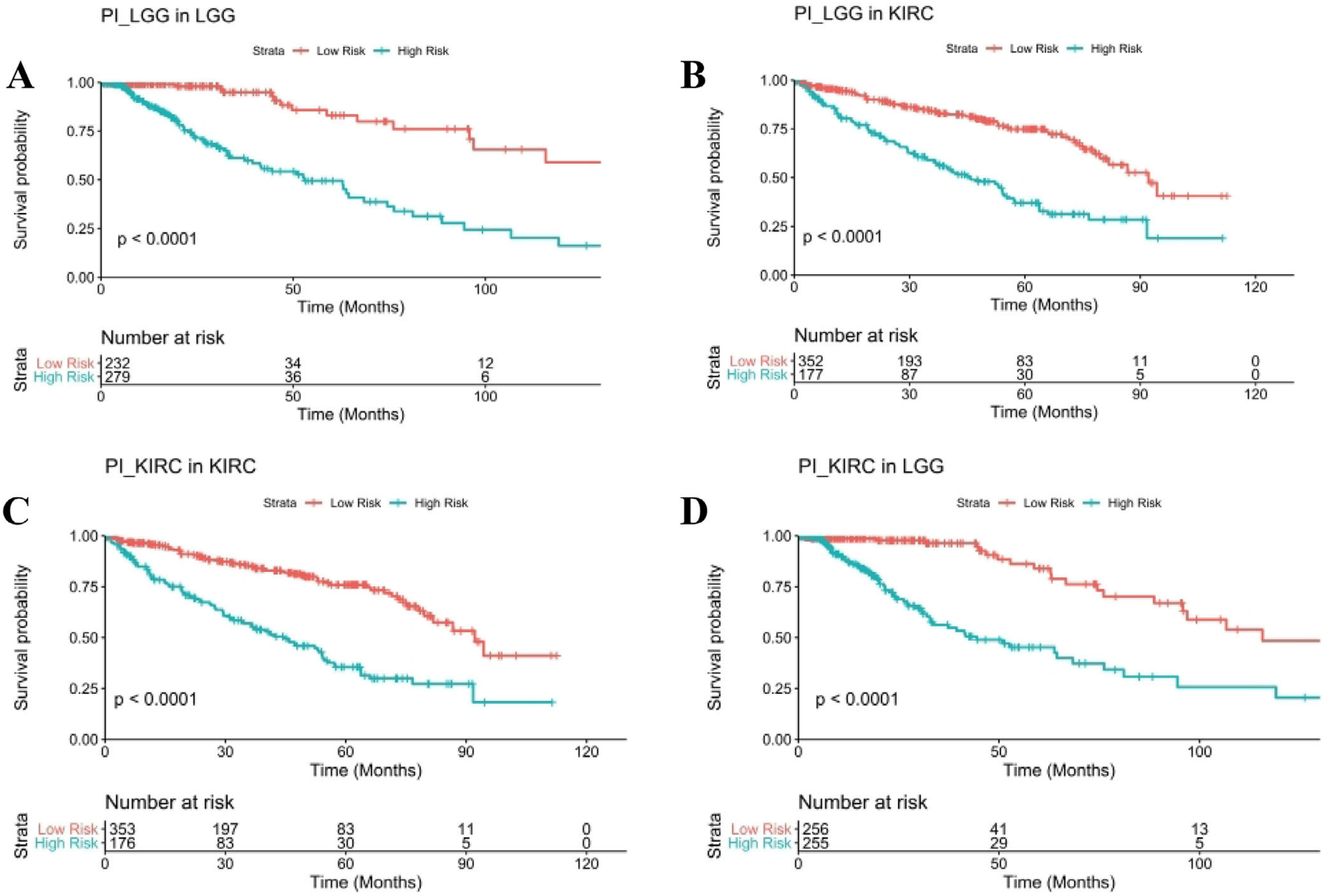


FIGURE 6 Development of cross-cancer prognostic models: LGG-KIRC. (A) KM plots representing the segregation of risk groups by $\mathrm{PI}_{\mathrm{LGG}}$ in LGG cohort and in (B) KIRC cohort. (C) KM plots representing the segregation of risk groups by $\mathrm{PI}_{\mathrm{KIRC}}$ in KIRC cohort and in (D) LGG cohort

$$\begin{aligned}\mathrm{PI}_{\mathrm{KIRC}} &= 0.6 \times \mathit{BTG3} + 0.7 \times \mathit{CDK2} + 0.56 \times \mathit{SOD2} + 0.54\\ &\times \mathit{TOP2A} + 0.6 \times \mathit{HMGB2} + 0.9 \times \mathit{TIMP1} + 0.55\\ &\times \mathit{ISG20} + 0.61 \times \mathit{TNFRSF12A} + 0.57 \times \mathit{IFNB1} - 0.79\\ &\times \mathit{ADD1} + 0.54 \times \mathit{CASP8} + 0.82 \times \mathit{CDC25B} + 0.53\\ &\times \mathit{IFITM3} + 0.52 \times \mathit{CD44} + 0.57 \times \mathit{GPX1} \qquad (5)\end{aligned}$$

Using these, risk stratification was performed in the respective cancer as well as another cancer. While $\mathrm{PI}_{\mathrm{LGG}}$ in LGG segregated the risk groups with HR = 4.77, *p*-value = 3.51×10^{-9}, C = 0.68, %95CI 2.84–8.01 and logrank-*p* = 3.41×10^{-11}; it showed a performance of HR = 2.95, *p*-value = 1.44×10^{-11}, C = 0.64, %95CI 2.15–4.04 and logrank-*p* = 1.37×10^{-11} in KIRC. Similarly, $\mathrm{PI}_{\mathrm{KIRC}}$ in KIRC stratified high and low risk patients with HR = 3.27, *p*-value = 1.82×10^{-13}, C = 0.66, %95CI 2.39–4.49 and logrank-*p* = 1.31×10^{-13} and in LGG with HR = 4.23, *p*-value = 1.88×10^{-9}, C = 0.69, %95CI 2.64–6.77 and logrank-*p* = 1.07 × 10^{-10}. KM plots corresponding to these are shown in Figure 6. It is also interesting to observe the same nature of these genes in both the cancers, as evident from the β values.

3.7 | Potential drug molecules for LGG and KIRC

We further utilized the Cmap2 database and screened the potential drug molecules which could modulate the expression profile of genes and reduce risk of death associated with high-risk groups in LGG and KIRC. After querying the list of 15 genes above, we obtained the ranked therapeutic molecules as shown in -S1 Table 11. Top two enriched candidates were Genistein (enrichment = 0.592, p = 0) and Hexestrol (enrichment = 0.918, p = 0.00004). Amongst these, Genistein continues to be a focus of attention in the scientific community for its anti-cancer effects.

4 | DISCUSSION

Risk assessment and management is a major challenge in treatment of cancer patients. In current clinical practice, patients are characterized in different risk groups based on several clinico-pathological features. Thereafter, different therapy regimens and treatment procedures are employed. However, this risk assessment process is plagued with a number of limitations which are mainly attributed to the process of obtaining the clinical features. To curb these issues and provide more reliable and efficient prognosis, a variety of molecular prognostic biomarkers have been proposed. Identification of majority of these biomarkers is based on the recent high-throughput sequencing data and modern bioinformatics tools. These biomarkers also offer a better understanding of the fundamental processes involved in the tumorigenesis and reveal novel therapeutic targets. Dysregulation of the apoptotic machinery of the cell is one of the most prominent feature of tumour cells. As a result of which, it has been widely studied in the past and the information gained from those studies has been successfully translated into development of several potential anti-cancer drugs. However, the past studies were limited in scope and confined to a small number of apoptosis related regulatory molecules and cancers. Also, while many of these studies have suggested many genes/proteins as potential prognostic biomarkers in specific cancers, only a few of these have identified the biomarkers which could be applicable across multiple cancers. These handful of pan-cancer biomarkers, however, have a limited performance. Thus, a comprehensive study involving all possible apoptosis related molecules across the multitude of cancers is needed.

In this study, we analysed the expression pattern of 165 genes of the apoptosis pathway across 33 different cancer cohorts. Using the pan-cancer dataset, we performed a correlation analysis wherein we found out that expression of the genes in the dataset was poorly correlated with OS of the patients. The risk prediction models constructed via utilizing the top correlated genes, thus performed poorly in discriminating high and low risk groups. Further, by utilizing Cox univariate survival analysis, we were able to find out genes that were related to the prognosis of patients in pan-cancer cohort in the context of OS. A universal biomarker based on voting model of top ten genes viz. EREG, IL1A, IL18, BAK1, BID, CDC25B, IL1B, ANXA1, TOP2A, and BRCA1 was able to provide a two-fold discrimination between the low and high-risk groups. Many of the previous studies highlight the prognostic relevance of these genes, thereby, strengthening our finding and providing additional value to our universal biomarker voting model. However due to the low stratification ability of the model ($C = 0.62$), we performed a further analysis to develop better multi-cancer risk prediction models. We estimated the prognostic potential of each of the 165 genes in individual cancers by means of univariate survival analyses. For each cancer, we developed risk stratification models using the top genes of the cancer and were able to achieve significant stratification in 29 out of 33 cancers. Next, we filtered out 11 genes whose expression was related to prognosis in at least eight cancers. All of these genes were found to be associated with one or more cancers as found out by the Cancer Gene Database. EREG and TOP2A, two of the top ten genes in previous pan-cancer analysis, were also amongst these 11 genes. Universal risk prediction models based on PI and voting were then constructed for these 11 genes and used to stratify patients in 27 cancer cohorts. Risk stratification was significant in all the cancers, though performance varied from $HR = 11.74$ in UVM (8 out of 11 genes were significantly related to OS) to $HR = 1.73$ in OV (2 out of 11 genes were significantly related to OS). This 11-gene model, thus can be used across multiple cancers for patient prognosis as corroborated from the validation performance using external datasets belonging to 7 cancer types. We also proposed a strategy for construction of cross-cancer prognostic biomarkers. We clustered the cancers in a hierarchical order by utilizing the number of shared survival associated genes, thereby finding the cancer pairs closest to each other. For this, a number of similarity indices were used (results not shown here) and Jaccard similarity index was found to be the best measure. We display the efficacy of this strategy by developing a PI based 15-gene prognostic biomarker for LGG-KIRC pair. The expression profile of 15 genes in LGG-KIRC pair, surprisingly, were observed to have a similar prognostic nature. 10 out of 11 genes were associated with a poor prognosis while ADD1 gene was associated with a good prognosis in both the cancers. While many of these genes have been associated with carcinogenesis in the past [47–53], these results demand that future efforts should be made to elucidate the functional roles of these genes in such two distant cancers. Further, in order to provide a therapeutic approach for reducing the mortality rate associated with high risk LGG and KIRC patients, we screened potential molecules which could modulate the gene expression of these 15 genes. The top screened molecule, Genistein, is an isoflavone found in soy products which has recently drawn attention of the scientific community due to its potential use in treatment of cancer. Genistein is well known to induce apoptosis and prevent metastasis and has been shown to benefit colorectal and breast cancer patients [54,55]. Another top enriched molecule, Hexestrol, is a synthetic oestrogen which was previously used for treatment of prostate and breast cancer but has been discontinued in most of the countries.

To conclude, our analysis first resulted in the development of cancer specific biomarkers, many of which show better stratification performances than the previously suggested biomarkers. Secondly, we developed universal or multi-cancer biomarkers which can be utilized across a larger number of cancers in contrast to the biomarkers suggested in previous studies. Thirdly, we proposed a new strategy for development of cross-cancer biomarkers. Overall this study sheds light on the prognostic power of apoptotic pathway genes in various cancers, implying their association with cancer related risk. The exploitation of this information can be used to modulate the expression of key apoptotic genes, thereby reducing the associated risk of death. Therefore, these genes could serve as potential therapeutic targets across multiple cancers.

5 | LIMITATIONS OF THE STUDY

Although, our study shows some promising results, we were unable to devise a universal biomarker which has a high stratification ability in all the cancers chosen here. Another limitation of the study is the lack of experimental validation, which is necessary for a clinical realization of the biomarkers. We were able to validate our universal model in eight external cancer cohorts but ideally it is expected to be tested through-out all the cancer cohorts mentioned in this study. Future efforts through public availability of more datasets would help realize the complete potential of this study. Moreover, addition of other omics data could further improve the performance of expression based prognostic models.

ACKNOWLEDGMENTS

The authors would like to acknowledge Indraprastha Institute of Information Technology-Delhi for providing infrastructure and fellowship support. No specific funding was available to carry out this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets used in this study are freely available at TCGA-GDC portal (https://portal.gdc.cancer.gov).

ORCID

Gajendra P. S. Raghava https://orcid.org/0000-0002-8902-2876

REFERENCES

1. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians 70*, 7–30.
2. Sever, R., & Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine 5*(4), a006098.
3. Zhao, N., Guo, M., Wang, K., Zhang, C., & Liu, X. (2020). Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. *Frontiers in Bioengineering & Biotechnology 8*, 268.
4. Li, B., Zhang, B., Wang, X., Zeng, Z., Huang, Z., Zhang, L., Wei, F., Ren, X., & Yang, L. (2020). Expression signature, prognosis value, and immune characteristics of Siglec-15 identified by pan-cancer analysis. *Oncoimmunology 9*(1), 1807291.
5. Wang, D., Zhang, H., Fang, X., Cao, D., & Liu, H. (2020). Pan-cancer analysis reveals the role of long non-coding RNA LINC01614 as a highly cancer-dependent oncogene and biomarker. *Oncology Letters 20*, 1383–1399.
6. Yuan, Q., Sun, Na, Zheng, J., Wang, Y., Yan, X., Mai, W., Liao, Y., & Chen, X. (2019). Prognostic and immunological role of FUN14 domain containing 1 in pan-cancer: Friend or foe? *Frontiers in Oncology 9*, 1502.
7. Chen, W., Li, G., Peng, J., Dai, W., Su, Q., & He, Y. (2020). Transcriptomic analysis reveals that heat shock protein 90α is a potential diagnostic and prognostic biomarker for cancer. *European Journal of Cancer Prevention Off European Journal of Cancer Prevention Organ 29*, 357–364.
8. Wu, H.-X., Wang, Z.-X., Zhao, Qi, Chen, D.-L., He, M.-M., Yang, L.-P., Wang, Y.-N., Jin, Y., Ren, C., Luo, H.-Y., Wang, Z.-Q., & Wang, F. (2019). Tumor mutational and indel burden: A systematic pan-cancer evaluation as prognostic biomarkers. *Annals of Translational Medicine 7*(22), 640.
9. Wong, R. S. Y. (2011). Apoptosis in cancer: From pathogenesis to treatment. *Journal of Experimental & Clinical Cancer Research 30*(1), 87.
10. Bauer, J. H., & Helfand, S. L. (2006). New tricks of an old molecule: Lifespan regulation by p53. *Aging Cell 5*, 437–440.
11. Frenzel, A., Grespi, F., Chmelewskij, W., & Villunger, A. (2009). Bcl2 family proteins in carcinogenesis and the treatment of cancer. *Apoptosis 14*, 584–596.
12. Pfeffer, C. M., & Singh, A. T. K. (2018). Apoptosis: A target for anticancer therapy. *International Journal of Molecular Sciences 19*(2), 448.
13. Charles, E. M., & Rehm, M. (2014). Key regulators of apoptosis execution as biomarker candidates in melanoma. *Molecular Cell Oncology 1*(3), e964037.
14. Zeestraten, E. C. M., Benard, A., Reimers, M. S., Schouten, P. C., Liefers, G. J., van de Velde, C. J. H, & Kuppen, P. J. K. (2013). The prognostic value of the apoptosis pathway in colorectal cancer: A review of the literature on biomarkers identified by immunohistochemistry. *Biomarkers Cancer 5*, 13–29.
15. Lathwal, A., Arora, C., & Raghava, G. P. S. (2019). Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway. *Plos One 14*(9), e0217527.
16. Lindner, A. U., Salvucci, M., Morgan, C., Monsefi, N., Resler, A. J., Cremona, M., Curry, S., Toomey, S., O'Byrne, R., Bacon, O., Stühler, M., Flanagan, L., Wilson, R., Johnston, P. G., Salto-Tellez, M., Camilleri-Broët, S., McNamara, D. A., Kay, E. W., Hennessy, B. T., Laurent-Puig, P., & Schaeybroeck, S. V. (2017). BCL-2 system analysis identifies high-risk colorectal cancer patients. *Gut 66*, 2141–2148.
17. Bai, Z., Ye, Y., Liang, B., Xu, F., Zhang, H., Zhang, Y., Peng, J., Shen, D., Cui, Z., Zhang, Z., & Wang, S. (2011). Proteomics-based identification of a group of apoptosis-related proteins and biomarkers in gastric cancer. *International Journal of Oncology 38*, 375–383.
18. Ding, L., Li, B., Yu, X., Li, Z., Li, X., Dang, S., Lv, Q., Wei, J., Sun, H., & Chen, H. (2020). KIF15 facilitates gastric cancer via enhancing proliferation, inhibiting apoptosis, and predict poor prognosis. *Cancer Cell International 20*(1), 125.
19. Pandya, V., Githaka, J. M., Patel, N., Veldhoen, R., Hugh, J., Damaraju, S., & McMullen, T. (2020). BIK drives an aggressive breast cancer phenotype through sublethal apoptosis and predicts poor prognosis of ER-positive breast cancer. *Cell Death & Disease 11*(6), 448.
20. Nakano, T., Go, T., Nakashima, N., Liu, D., & Yokomise, H. (2020). Overexpression of antiapoptotic MCL-1 predicts worse overall survival of patients with non-small cell lung cancer. *Anticancer Research 40*, 1007–1014.
21. Zeng, S., Liu, A., Dai, L., Yu, X., Zhang, Z., Xiong, Q., Yang, J., Liu, F., Xu, J., Xue, Y., Sun, Y., & Xu, C. (2019). Prognostic value of TOP2A in bladder urothelial carcinoma and potential molecular mechanisms. *Bmc Cancer [Electronic Resource] 19*(1), 604.
22. Liu, Y.-Q., Wu, F., Li, J.-J., Li, Y.-F., Liu, X., Wang, Z., & Chai, R.-C. (2019). Gene expression profiling stratifies IDH-wildtype glioblastoma with distinct prognoses. *Frontiers in Oncology 9*, 1433.
23. Ma, L., Zhang, L., Guo, A., Liu, L. C., Yu, F., Diao, N., Xu, C., Wang, D. (2019). Overexpression of FER1L4 promotes the apoptosis and suppresses epithelial-mesenchymal transition and stemness markers via activating PI3K/AKT signaling pathway in osteosarcoma cells. *Pathology-Research and Practice 215*(6), 152412.
24. Wang, S. H., & Baker, J. R. (2006). *Thyroid Cancer (Second Edition): A Comprehensive Guide to Clinical Management*, (pp. 55–61). Humana Press.
25. Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., & Ji, Y. (2018). TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics 34*, 1615–1617.
26. Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W.-W., Foltz, S. M., Shmulevich, I., Ding, Li, Heins, Z., Ochoa, A., & Gross, B. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell 173*, 321–337 e10.
27. van der Net, J. B., Janssens, A. C., Defesche, J. C., Kastelein, J. J., Sijbrands, E. J. G., Steyerberg, & E. W. (2009). Usefulness of genetic polymorphisms and conventional risk factors to predict coronary heart disease in patients with familial hypercholesterolemia. *American Journal of Cardiology 103*, 375–380.
28. Dyrskjot, L., Reinert, T., Algaba, F., Christensen, E., Nieboer, D., Hermann, G. G., Mogensen, K., Beukers, W., Marquez, M., Segersten, U., Høyer, S., Ulhøi, B. P., Hartmann, A., Stöhr, R., Wach, S., Nawroth, R., Schwamborn, K., Tulic, C., Simic, T., Junker, K., & Harving, N. (2017). Prognostic impact of a 12-gene progression score in non-muscle-invasive bladder cancer: A prospective multicentre validation study. *European Urology 72*, 461–469.
29. Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research 24*, 1248–1259.

30. Li, P., Ren, H., Zhang, Y., Zhou, Z. (2018). Fifteen-gene expression based model predicts the survival of clear cell renal cell carcinoma. *Medicine 97*(33), e11839.
31. Wang, Y., Ren, F., Chen, P., Liu, S., Song, Z., & Ma, X. (2018). Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. *Cancer Medicine 7*, 5632–5642.
32. Lathwal, A., Kumar, R., Arora, C., Raghava, G. P. S. (2020). Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data. *Journal of Cancer Research and Clinical Oncology 146*(11), 2743–2752.
33. Arora, C., Kaur, D., Lathwal, A., & Raghava, G. P. S. (2020). Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data. *Heliyon 6*(8), e04811.
34. Kaur, D., Arora, C., Raghava, G. P. S. (2021). Prognostic Biomarker-Based Identification of Drugs for Managing the Treatment of Endometrial Cancer. *Molecular Diagnosis & Therapy 25*(5), 629–646.
35. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, Ru, Carr, S. A., Lander, E. S., & Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science 313*, 1929–1935.
36. Musa, A., Ghoraie, L. S., Zhang, S.-D., Glazko, G., Yli-Harja, O., Dehmer, M., Haibe-Kains, B., & Emmert-Streib, F. (2018). A review of connectivity map and computational approaches in pharmacogenomics. *Briefings in Bioinformatics 19*, 506–523.
37. Xia, Q., Zhou, Y., Yong, H., Wang, X., Zhao, W., Ding, G., Zhu, J., Li, X., Feng, Z., & Wang, B. (2019). Elevated epiregulin expression predicts poor prognosis in gastric cancer. *Pathology, Research and Practice 215*, 873–879.
38. Gil, M., & Kim, K. E. (2019). Interleukin-18 is a prognostic biomarker correlated with CD8(+) T cell and natural killer cell infiltration in skin cutaneous melanoma. *Journal of Clinical Medicine 8*(11), 1993.
39. Kubo, T., Kawano, Y., Himuro, N., Sugita, S., Sato, Y., Ishikawa, K., Takada, K., Murase, K., Miyanishi, K., Sato, T., Takimoto, R., Kobune, M., Nobuoka, T., Hirata, K., Takayama, T., Mori, M., Hasegawa, T., & Kato, J. (2016). BAK is a predictive and prognostic biomarker for the therapeutic effect of docetaxel treatment in patients with advanced gastric cancer. *Gastric Cancer Off Journal International Gastric Cancer Association Japanese Gastric Cancer Association 19*, 827–838.
40. Malhotra, S., Costa, C., Eixarch, H., Keller, C. W., Amman, L., Martínez-Banaclocha, H., Midaglia, L., Sarró, E., Machín-Díaz, I., Villar, L. M., Triviño, J. C., Oliver-Martos, B., Parladé, L. N., Calvo-Barreiro, L., Matesanz, F., Vandenbroeck, K., Urcelay, E., Martínez-Ginés, M - L., Tejeda-Velarde, A., Fissolo, N., & Castilló, J. (2020). NLRP3 inflammasome as prognostic factor and therapeutic target in primary progressive multiple sclerosis patients. *Brain 143*, 1414–1430.
41. Raulf, N., Lucarelli, P., Thavaraj, S., Brown, S., Vicencio, J. M., Sauter, T., & Tavassoli, M. (2018). Annexin A1 regulates EGFR activity and alters EGFR-containing tumour-derived exosomes in head and neck cancers. *European Journal of Cancer 102*, 52–68.
42. Stopsack, K. H., Gerke, T., Zareba, P., Pettersson, A., Chowdhury, D., Ebot, E. M., Flavin, R., Finn, S., Kantoff, P. W., Stampfer, M. J., Loda, M., Fiorentino, M., & Mucci, L. A. (2020). Tumor protein expression of the DNA repair gene BRCA1 and lethal prostate cancer. *Carcinogenesis 41*, 904–908.
43. Abbott, K. L., Nyre, E. T., Abrahante, J., Ho, Y.-Y., Vogel, R. I., & Starr, T. K. (2015). The candidate cancer gene database: A database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research 43*, D844-D848.
44. Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., William, A., Spies, N. C., Griffith, O. L., & Griffith, M. (2018). Drib 3.0: A redesign and expansion of the drug-gene interaction database. *Nucleic Acids Research 46*, D1068–D1073.
45. Seligmann, J. F., Elliott, F., Richman, S. D., Jacobs, B., Hemmings, G., Brown, S., Barrett, J. H., Tejpar, S., Quirke, P., & Seymour, M. T. (2016). Combined epiregulin and amphiregulin expression levels as a predictive biomarker for panitumumab therapy benefit or lack of benefit in patients with RAS wild-type advanced colorectal cancer. *JAMA Oncology 2*, 633–642.
46. Aguirre-Gamboa, R., Gomez-Rueda, H., Martínez-Ledesma, E., Martínez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Peña, J. G., Treviño, V. (2013). SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis. *PLoS ONE 8*(9), e74250.
47. Lv, C., Wang, H., Tong, Y., Yin, H., Wang, D., Yan, Z., Liang, Y., Wu, D., & Su, Q. (2018). The function of BTG3 in colorectal cancer cells and its possible signaling pathway. *Journal of Cancer Research and Clinical Oncology 144*, 295–308.
48. Tadesse, S., Anshabo, A. T., Portman, N., Lim, E., Tilley, W., Caldon, C. E., & Wang, S. (2020). Targeting CDK2 in cancer: Challenges and opportunities for therapy. *Drug Discovery Today 25*, 406–413.
49. Kim, Y. S., Gupta Vallur, P., Phaeton, R., & Mythreye, K., Hempel, N. (2017). Insights into the dichotomous regulation of SOD2 in cancer. *Antioxidants (Basel, Switzerland) 6*(4), 86.
50. Grunnet, M., Mau-Sorensen, M., & Brunner, N. (2013). Tissue inhibitor of metalloproteinase 1 (TIMP-1) as a biomarker in gastric cancer: A review. *Scandinavian Journal of Gastroenterology 48*, 899–905.
51. Lavecchia, A., Coluccia, A., Di Giovanni, C., & Novellino, E. (2008). Cdc25B phosphatase inhibitors in cancer therapy: Latest developments, trends and medicinal chemistry perspective. *Anti-Cancer Agents in Medicinal Chemistry 8*, 843–856.
52. Chen, C., Zhao, S., Karnad, A., & Freeman, J. W. (2018). The biology and role of CD44 in cancer progression: Therapeutic implications. *Journal of Hematology & Oncology 11*(1), 64.
53. Cao, M., Mu, X., Jiang, C., Yang, G., Chen, H., & Xue, W. (2014). Single-nucleotide polymorphisms of GPX1 and MnSOD and susceptibility to bladder cancer: A systematic review and meta-analysis. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine 35*, 759–764.
54. Spagnuolo, C., Russo, G. L., Orhan, I. E., Habtemariam, S., Daglia, M., Sureda, A., Nabavi, S. F., Devi, K. D., Loizzo, M. R., Tundis, R., & Nabavi, S. M. (2015). Genistein and cancer: Current status, challenges, and future directions. *Advances in Nutrition 6*, 408–419.
55. Tuli, H. S., Tuorkey, M. J., Thakral, F., Sak, K., Kumar, M., Sharma, A. K., Sharma, U., Jain, A., Aggarwal, V., & Bishayee, A. (2019). Molecular mechanisms of action of genistein in cancer: Recent advances. *Frontiers in Pharmacology 10*, 1336.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202000311 in the Supporting Information section at the end of the article.

How to cite this article: Arora, C., Kaur, D., & Raghava, G. P. S. (2022). Universal and cross-cancer prognostic biomarkers for predicting survival risk of cancer patients from expression profile of apoptotic pathway genes. *Proteomics*, *22*, e2000311. https://doi.org/10.1002/pmic.202000311