OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–12
https://doi.org/10.1093/bib/bbac090
Problem Solving Protocol

cSurvival: a web resource for biomarker interactions in cancer outcomes and in cell lines

Xuanjin Cheng, Yongxing Liu, Jiahe Wang, Yujie Chen, Andrew Gordon Robertson, Xuekui Zhang, Steven J. M. Jones and Stefan Taubert

Corresponding authors: Xuanjin Cheng, Centre for Molecular Medicine and Therapeutics; British Columbia Children's Hospital Research Institute; Department of Medical Genetics, The University of British Columbia, Vancouver, British Columbia, V5Z 4H4, Canada. Tel.: +1-604-875-3860; Fax: +1 604-875-3819; E-mail: jcheng@cmmt.ubc.ca; Stefan Taubert, Centre for Molecular Medicine and Therapeutics; British Columbia Children's Hospital Research Institute; Department of Medical Genetics, The University of British Columbia, Vancouver, British Columbia, V5Z 4H4, Canada. Tel.: +1-604-875-3860; Fax: +1 604-875-3819; E-mail: taubert@cmmt.ubc.ca

Abstract

Survival analysis is a technique for identifying prognostic biomarkers and genetic vulnerabilities in cancer studies. Large-scale consortium-based projects have profiled >11 000 adult and >4000 pediatric tumor cases with clinical outcomes and multiomics approaches. This provides a resource for investigating molecular-level cancer etiologies using clinical correlations. Although cancers often arise from multiple genetic vulnerabilities and have deregulated gene sets (GSs), existing survival analysis protocols can report only on individual genes. Additionally, there is no systematic method to connect clinical outcomes with experimental (cell line) data. To address these gaps, we developed cSurvival (https://tau.cmmt.ubc.ca/cSurvival). cSurvival provides a user-adjustable analytical pipeline with a curated, integrated database and offers three main advances: (i) joint analysis with two genomic predictors to identify interacting biomarkers, including new algorithms to identify optimal cutoffs for two continuous predictors; (ii) survival analysis not only at the gene, but also the GS level; and (iii) integration of clinical and experimental cell line studies to generate synergistic biological insights. To demonstrate these advances, we report three case studies. We confirmed findings of autophagy-dependent survival in colorectal cancers and of synergistic negative effects between high expression of *SLC7A11* and *SLC2A1* on outcomes in several cancers. We further used cSurvival to identify high expression of the Nrf2-antioxidant response element pathway as a main indicator for lung cancer prognosis and for cellular resistance to oxidative stress-inducing drugs. Altogether, these analyses demonstrate cSurvival's ability to support biomarker prognosis and interaction analysis via gene- and GS-level approaches and to integrate clinical and experimental biomedical studies.

Keywords: genetic interaction, survival analysis, TCGA, TARGET, DepMap, biomarker

Introduction

Survival analysis, or, more broadly, time-to-event analysis, assesses the statistical association between potential risk factors and the time to an event such as death or disease recurrence [1, 2]. In both basic and clinical cancer biology studies, survival analysis is an important technique for identifying prognostic biomarkers and genetic vulnerabilities. Experimentally, it is useful for hypothesis generation and mechanistic inference. Clinically, it may help stratify patients into subgroups with distinct risk profiles and guide therapeutic decisions [3, 4].

Since 2006, consortium-based projects, such as The Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET), have gathered clinicopathologic data along with multiomics molecular profiles of more than 15 000 adult and pediatric human tumors across diverse cancer types [5–7]. Such large data resources allow exploration

Xuanjin Cheng is a researcher in the Department of Medical Genetics at The University of British Columbia. Her research combines bioinformatics and molecular biology to study gene regulation in development and disease.
Yongxing Liu is a fourth-year Mathematics student at The University of British Columbia. He is interested in data science, machine learning and bioinformatics.
Jiahe Wang is currently a Computer Science master student at Simon Fraser University with a concentration on big data and machine learning.
Yujie Chen is an undergraduate student majoring in Statistics at The University of British Columbia, and her research interest is machine learning and bioinformatics.
Andrew Gordon Robertson was an analyst with BC Cancer's Canada's Michael Smith Genome Sciences Centre in Vancouver, Canada. He has contributed to many The Cancer Genome Atlas, PanCancer Atlas and GDAN projects and publications. He is currently an analyst with Dxige Research Inc., in Courtenay, BC, Canada.
Xuekui Zhang is a Tier 2 Canada Research Chair in biostatistics and bioinformatics and an assistant professor in the Department of Mathematics and Statistics at the University of Victoria.
Steven J.M. Jones is a codirector and the head of bioinformatics at the Genome Sciences Centre, BC Cancer. He is also a professor in the Department of Medical Genetics at The University of British Columbia.
Stefan Taubert is an associate professor in the Department of Medical Genetics at The University of British Columbia who studies gene regulation in health and disease.

Received: October 5, 2021. **Revised:** February 2, 2022. **Accepted:** February 24, 2022

into cancers at the molecular level, using clinical correlations at an unprecedented scale [6].

Despite the importance of survival analysis and the richness of tumor molecular datasets, we find that currently available tools do not fully exploit the potential of survival analysis (Supplementary Table 1, see Supplementary Data available online at *Briefings in Bioinformatics* online). First, existing tools can only analyze one genomic predictor at a time, typically mutation or expression of an individual gene [8–24]. However, cancer often occurs due to multiple (epi)genomic alterations, creating highly connected molecular networks where different alterations synergize to cause malignancy. Cross talk can happen between different factors in cancers, for example microRNA (miRNA) expression and deoxyribonucleic acid (DNA) methylation [25]. In survival analysis, incorporating more than one predictor could identify interactions between molecular alterations. Such interaction analysis could also be used to screen for synthetic lethality or to identify compensatory targets for nontargetable drivers [26–29]. This, in turn, could facilitate the development of combination therapies (e.g. drug cocktails), which have higher efficacy and milder adverse effects than monotherapies (aka one-gene-one-drug) approaches [30, 31]. Second, existing tools support analysis only at the single-gene level [8–24]; however, molecular dysregulations in cancers may involve GSs [32], in which a collection of genes act in concert. For example, a GS may represent a specific pathway (e.g. transforming growth factor-β-mediated SMAD signaling), biological process (e.g. cell cycle), disease (e.g. hereditary nonpolyposis colorectal cancer) or treatment (e.g. chemotherapy) [33, 34]. Given this, analysis of prognostic biomarkers at the GS level rather than at the single-gene level should be informative [35–37]. Third, experimentally derived cancer cell line viability data [38–40] and multiomics profiling [41] provide valuable *in vitro* information on genetic dependencies and interactions; however, no existing tool connects clinical data to such experimental studies. This hinders identifying suitable preclinical cell line tools to investigate molecular mechanisms underpinning poor prognosis.

Motivated by the lack of suitable tools to address the above challenges, we developed cSurvival (Figure 1). Its major advances are as follows.

(i) Joint analysis with two genomic predictors on a wide range of individual cancer types or combinations of cancer types, including new algorithms to search for optimal cutoffs in combinations of two continuous predictors, in order to stratify patients into risk groups. The two predictors can include combinations of diverse data types that can include gene or GS expression, somatic mutation, miRNA expression, DNA methylation and protein expression.

(ii) Survival analysis at the GS level with comprehensive and up-to-date GS libraries from the easy Visualization and Inference Toolbox for Transcriptome Analysis (eVITTA) project, a webserver dedicated to analyzing, comparing and visualizing transcriptome patterns [42].

(iii) A pipeline to integrate clinical outcomes and experimental cancer cell line data.

We have combined a curated cancer outcomes database with a refined analytical pipeline and customizable visualizations into the cSurvival webserver so that nonprogrammers can use it. In the work described here, we demonstrate cSurvival's capabilities with three case studies and more application cases in our user guide (https://tau.cmmt.ubc.ca/cSurvival/help.html#9_Application_cases). We not only recapitulated reported cancer biomarkers and their interactions but also identified genetic regulations consistent with published studies, demonstrating that cSurvival's advanced pipeline facilitates cancer biomarker studies.

Materials and methods

Data extraction and processing

The Cancer Genome Atlas

We extracted curated clinical outcome endpoints data from the TCGA Pan-Cancer Clinical Data Resource [6] and multiomics molecular data from [43]. We removed 2614 low-quality samples (Do_not_use = True) and flagged additional 507 problematic cases based on comments in the merged_sample_quality_annotations.tsv file (https://gdc.cancer.gov/node/977). While cSurvival removes these 507 cases (https://tau.cmmt.ubc.ca/cSurvival/project_data/977/flagged_cases.tsv) by default, a user can choose to include them via the web interface. Next, we used TCGA sample type codes (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes) to extract tumor samples: for solid tumors, we extracted primary solid tumor (01) samples; for acute myeloid leukemia, we extracted primary blood-derived tumors (03 and 09); for skin cutaneous melanoma, we extracted both primary solid (01) and metastatic tumors (06). Then, from Ref. [43], we used batch-corrected, upper quartile-normalized RNA-Seq by Expectation Maximization (RSEM) data; merged somatic mutation calls from the Multi-Center Mutation Calling in Multiple Cancers project [44]; purity- and ploidy-corrected, gene-level, thresholded somatic copy number (CN) data; batch-corrected, reads per million data for expressed miRNA mature strands; beta values from Illumina HumanMethylation27 (HM27) and HumanMethylation450 (HM450) arrays; and batch-corrected reverse-phase protein array (RPPA) data. For duplicated tumor samples, for gene expression, miRNA expression, DNA methylation and RPPA data, we calculated the geometric means as the final readouts. Statistics in log scale are commonly used to summarize the characteristics of genomic data, since their original values are usually not normally distributed (such as log-fold changes reported by Model-based Analysis of Single-cell Transcriptomics (MAST) [45] and edgeR [46] in differential expression analysis). Hence, we use the geometric mean to represent the average, which is equivalent to the exponential of the average of log-transformed data. We further used the annotations in [47] to map HM27 and HM450 probe IDs with chromosomal coordinates and adjacent genes.

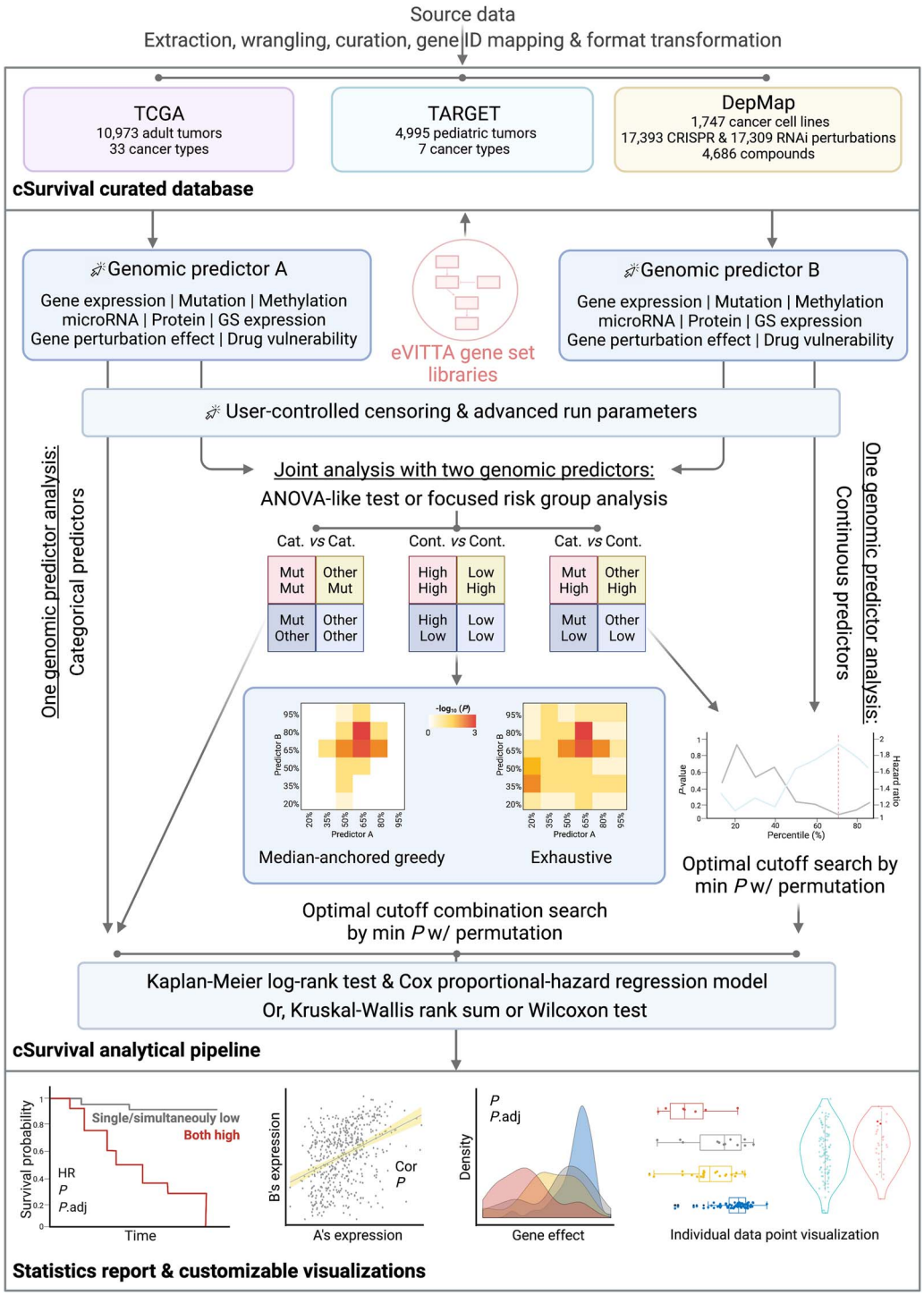

Figure 1. Overview of the cSurvival analytical framework. DepMap, Dependency Map; eVITTA, easy Visualization and Inference Toolbox for Transcriptome Analysis; Cat., categorical predictor; Cont., continuous predictor; Mut, mutated; HR, hazard ratio; P, *P*-value; P.adj, adjusted *P*-value; Cor, correlation coefficient.

Therapeutically Applicable Research to Generate Effective Treatment

We extracted clinical and multiomics data from the NCI Genomic Data Commons (GDC) [48] (https://gdc.cancer.gov/) with TCGAbiolinks v2.16.4 [49–51]. As above, we extracted primary tumor samples (01 for solid tumors, 03 and 09 for blood-derived). We used Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-UQ) data for gene expression analysis, and open-access somatic mutation calls for mutation analysis. We converted Ensembl gene IDs into Human Genome Organisation (HUGO) symbols and Entrez IDs using org.Hs.eg.db v3.11.4 [52].

Dependency Map

We extracted cell line annotation, mutation, gene expression (transcripts per million [TPM]), CN, clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9, RNA interference (RNAi) and drug sensitivity data from Dependency Map (DepMap) 21Q3 [39–41, 53],

normalized protein expression levels from the CCLE proteomics (TS2) database (accessed on 5/19/2021) [54] and drug information from the Drug Repurposing Hub v3/24/2020 [55].

Calculating GS expression

After a user selects a cancer type or combinations of cancer types, we first transform the normalized gene expression values (upper quartile-normalized RSEM in TCGA, FPKM-UQ in TARGET, TPM in DepMap) of all samples in the selected cancer type(s) into z-scores. Specifically, for a given gene, its z-score in a sample (patient in TCGA and TARGET, cell line in DepMap) is calculated as $z = \frac{(x-\mu)}{\sigma}$, where x is the expression value of the gene in the sample, μ is the mean count of the gene across samples and σ is the standard deviation of all expression values of the gene across samples. Then, for each sample, we computed the expression of a GS as the average expression z-score of all genes within the GS [36].

Survival analysis

Censoring

A user chooses a censoring time in days, months (30.4375 days) or years (365.25 days) (default: 10 years). For a selected clinical endpoint [e.g. overall survival (OS), progression-free survival], if the time-to-event is larger than the defined time, we set censoring status to 0 and time to the defined time; if the time-to-event is smaller than or equal to the defined time, we set censoring status to 1 and time stays unchanged.

Survival analysis

We apply Kaplan–Meier (KM) log-rank tests (default) and Cox proportional-hazards (PH) regression models to assess the association with prognosis, using survival v3.2.11 [56]. In joint analysis with two predictors, we use the KM log-rank test (default) or Cox PH likelihood ratio test to assess the overall significance of any difference between the four subgroup combinations of two predictors (Supplementary Figure 1, see Supplementary Data available online at *Briefings in Bioinformatics* online). In addition, we apply Cox PH regression models to assess how two predictors jointly impact outcomes by calculating the effect sizes (hazard ratios, HRs) and the statistical significances of the two predictors and their interaction, from the fitted regression model. Alternatively, users select a risk subgroup of interest, and we then apply a KM log-rank test (default) or Cox PH likelihood ratio test to assess the difference between the selected subgroup and the rest of the cases.

Determining optimal cutoffs for continuous predictors

By default, for analysis with a single continuous predictor (gene expression, miRNA expression, DNA methylation, protein expression and cell line unthresholded CN), and for joint analysis with combinations of continuous and categorical predictors, we determine optimal cutoffs using the minimum *P*-value (default: KM log-rank) method [57–61] by testing from the lowest (default 0.2) to the highest (default 0.8) percentile with a defined step (default 0.1). In joint analysis with combinations of two continuous predictors, we determine optimal cutoffs using a median-anchored greedy (default) or an exhaustive search (described below, and in Figure 1). Because multiple tests are conducted in searching for optimal cutoffs, we apply a *P*-value correction method to control for false positive probability (described below).

(i) Median-anchored greedy search: we construct a 2D grid using percentiles of both predictors (Figure 1). Next, we determine the starting point for a greedy search by locating the minimum *P*-value computed from testing each percentile in predictor B against the median percentile in predictor A. Then, we test the nearest three unexplored points; if a lower *P*-value is found, we move the search to that newly found minimum *P*-value point and test the nearest unexplored points until no lower *P*-value can be found. We test only percentile combinations that have at least 10% (default) of total cases in each subgroup or subgroup combination.

(ii) Exhaustive search: We construct a 2D grid using percentiles of both predictors (Figure 1). Next, we determine the optimal percentile combination by locating the minimum *P*-value computed from testing each percentile in predictor B against each percentile in predictor A. As for the greedy search above, we only test percentile combinations giving at least 10% (default) of total cases in each subgroup or subgroup combination.

We also offer customizable analysis with user-selected percentile cutoffs. However, we recommend using the default minimum *P*-value search because: (i) studies have shown its advantages over an arbitrarily prespecified cutoff [57–61], and (ii) if a percentile (e.g. median) or percentile combination (e.g. median-median) was optimal to stratify the patient samples, the search would detect it and use it.

In joint analysis with two continuous genomic predictors, we recommend 'median-anchored greedy search' over 'exhaustive search'. Exhaustive search can find the best cutoff values if there is enough statistical power (i.e. a large enough sample size). However, this approach involves many tests, which can lead to a heavy penalty in adjusting for multiple testing, i.e. can result in insufficient statistical power. In practice, most comparisons only have limited sample sizes; hence, we propose a balanced choice, the 'median-anchored greedy search'. This smart search strategy involves many fewer tests to address the issue of losing statistical power due to multiple testing, while minimally sacrificing the opportunity to investigate good candidate cutoff values.

Permutation-based multiple testing adjustment for optimally selected cutoffs

We use permutations to correct the multiple testing that arises from assessing a sequence of candidate cutoffs

with the minimum *P*-value method [62]. Briefly, we randomly permute outcomes values (in TCGA and TARGET, survival days and censoring status; in DepMap, gene perturbation effects or drug sensitivity scores) over the samples, and determine the new optimal cutoff. We repeat this a defined number of times (default: $n = 100$) to generate a null distribution of minimum *P*-values, i.e. the empirical distribution for the minimum *P*-values when there is no association between the biomarkers and the survival outcomes. Then, we calculate an empirically adjusted *P*-value (*P*.adj) by comparing the observed minimum *P*-value to this empirical null distribution. To speed up the calculation, we use mclapply v4.0.3 [63] for parallel processing.

Differential dependency and cell viability analysis

For two-group comparisons, we used a two-tailed two-sample Wilcoxon test (wilcox.test [63]) to assess the differences in dependency scores (CRISPR-Cas9, RNAi) or cell viabilities (drug sensitivity assays). Comparisons between more than two groups are assessed with a Kruskal–Wallis rank sum test (kruskal.test [63]) to test the overall significance of any difference between subgroups. Because dependency scores and cell viability data are skewed to the left (Supplementary Figure 2A–C, see Supplementary Data available online at *Briefings in Bioinformatics* online), we use nonparametric Wilcoxon and Kruskal–Wallis rank sum tests, which require no distribution assumption. For continuous genomic predictors, we determine optimal cutoffs and apply a multiple testing adjustment, as described above.

Customizable and interactive visualizations

We generate survival curves and forest plots with survminer v0.4.9 [64]. We also create interactive visualizations for further analysis with ggplot2 v3.3.5 [65] and plotly v4.9.4.1 (https://plotly.com/) (Figure 1): (i) density and box plots showing the distribution of dependency scores (DepMap); (ii) line plots showing *P*-values and HRs tracked over percentiles; (iii) heatmaps showing *P*-values and HRs searched over percentile combinations; (iv) bar plots showing the distribution of somatic mutations; (v) scatter plots analyzing correlations between two continuous predictors and (vi) violin plots assessing differences in values of a continuous predictor between two categories (e.g. expression differences of a pathway between mutated *versus* nonmutated groups). Each visualization is customizable with its own plotting parameters (e.g. colors, time intervals on the *x*-axis), and data points of interest are searchable and highlightable in box, scatter and violin plots.

Correlation analysis in scatter plots

We apply Pearson's product–moment correlation (default), Kendall's rank correlation tau and Spearman's rank correlation rho (cor.test [63]) to measure correlations between two continuous predictors.

Group mean analysis in violin plots

We apply a two-tailed two-sample Wilcoxon test (wilcox.test [63]) to assess the differences in a continuous predictor (e.g. gene expression) between two subcategories of a categorical predictor (e.g. loss-of-function mutations versus other). As above, we chose a nonparametric Wilcoxon test because it requires no distribution assumption (examples of nonnormalities are provided in Supplementary Figure 2D and E, see Supplementary Data available online at *Briefings in Bioinformatics* online).

Web interface implementation

We implement the web interface of cSurvival using Apache (v2.4.29, https://httpd.apache.org), R (v4.0.3, https://www.r-project.org/), R Shiny (v1.5.0, https://CRAN.R-project.org/package=shiny) and R Shiny Server (v1.5.14.948, https://rstudio.com/products/shiny/download-server/). We use plumber (v1.1.0, https://CRAN.R-project.org/package=plumber) and pm2 (v5.1.1, https://pm2.keymetrics.io/) to host cSurvival's API.

Results

In cSurvival v1.0.0, we have aggregated the following data. From the TCGA and the TARGET projects, clinical and multiomics data of 10 973 adult and 4995 pediatric tumors across 40 cancer types (33 adult, 7 pediatric). From the DepMap project [38]: (i) multiomics data of 1747 cell lines; (ii) cell viability data from 4686 drug compounds screened in 578 cell lines; and (iii) genetic perturbation data from 17 393 and 17 309 genes screened via CRISPR-Cas9 and RNAi in 1032 and 712 cell lines, respectively. From the eVITTA project v1.2.13 [42]: 120 953 GSs (Figure 2).

Here, we report three case studies to demonstrate cSurvival's unique abilities in:

(i) GS-level predictor analysis;
(ii) joint analysis with two genomic predictors;
(iii) integration of clinical and laboratory data to generate biological insights.

First, we tested cSurvival's analytical pipeline at the level of GSs. We recapitulated the finding that high expression of an autophagy signature [Gene Ontology (GO): 0010506] is associated with poor OS in colorectal cancers [colon adenocarcinoma (TCGA-COAD) and rectum adenocarcinoma (TCGA-READ)] (Figure 3A and B; percentile tracking in Supplementary Figure 3A, see Supplementary Data available online at *Briefings in Bioinformatics* online) [36].

Second, we performed joint analysis with gene expression data of solute carrier family 7 member 11 (*SLC7A11*) and solute carrier family 2 member 1 (*SLC2A1*, also known as *glucose transporter 1 (GLUT1)*) in liver hepatocellular carcinoma (TCGA-LIHC). We found that *SLC7A11* and *SLC2A1* showed a moderate correlation in their expressions (Figure 3C), and that patients with higher expression of both *SLC7A11* and *SLC2A1* showed

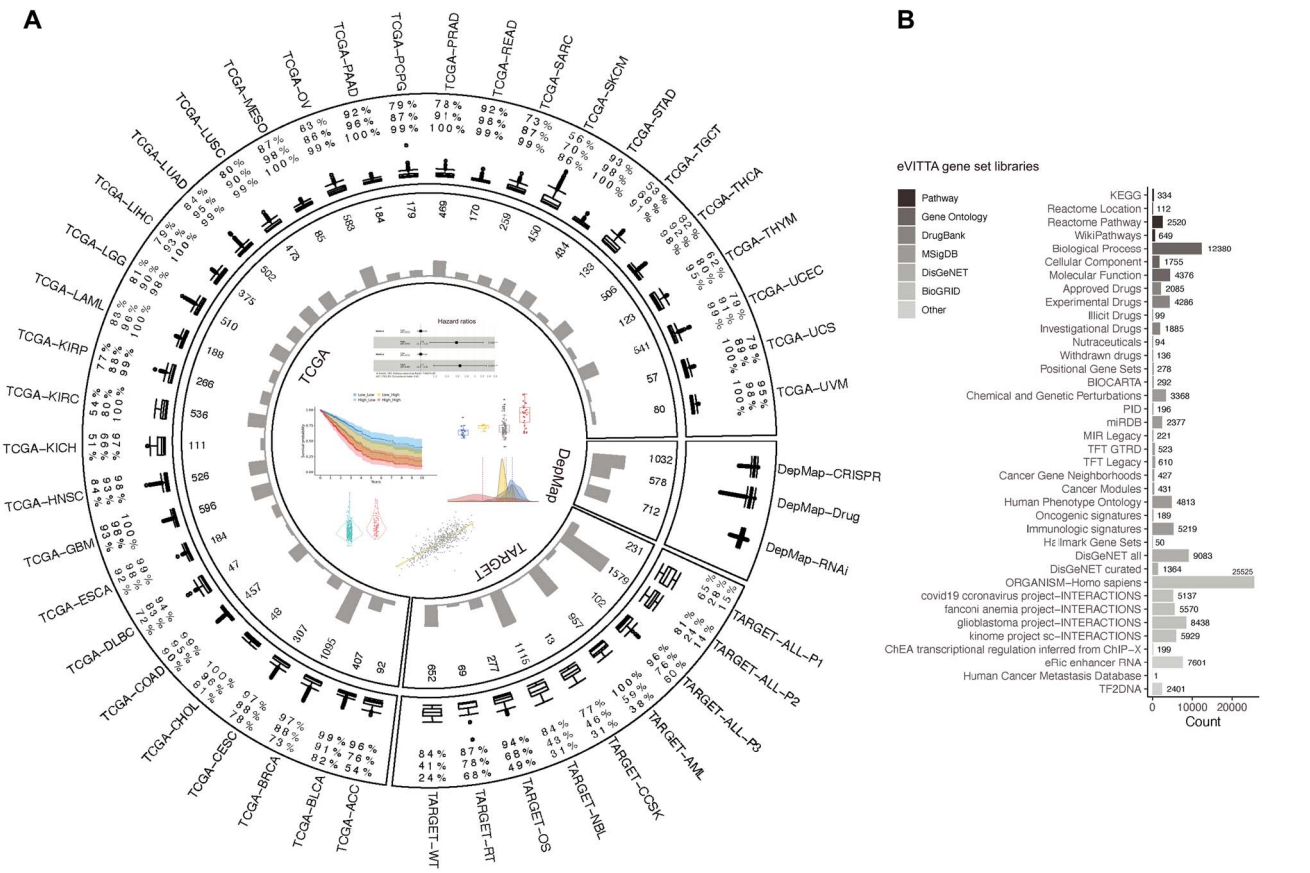

Figure 2. Overview of the cSurvival database. (**A**) The circos plot (rendered with circularize [82]) shows the distribution of tumor and cell line datasets: Inner to outer: example outputs; project names; histograms showing the total number of cases per study; box plots showing the distribution of survival days (TCGA, TARGET) or dependency scores/cell viabilities (DepMap), numbers denoting 3-, 5- and 10-year survival rates from inner to outer; study names. (**B**) The histogram shows the distribution of GS libraries from eVITTA; numbers denote the number of GSs per library. DepMap, Dependency Map; eVITTA, easy Visualization and Inference Toolbox for Transcriptome Analysis.

significantly lower survival rates than patients with low expression of *SLC7A11* and/or *SLC2A1* (Figure 3D–F; percentile tracking in Supplementary Figure 3B and C, see Supplementary Data available online at *Briefings in Bioinformatics* online). The synergistic negative effects between high expressions of *SLC7A11* and *SLC2A1* on outcomes were also observed in several other cancer types (Supplementary Figure 4, see Supplementary Data available online at *Briefings in Bioinformatics* online). These results are consistent with the finding that cotargeting the L-cystine importer *SLC7A11* and the glucose transporter *SLC2A1* induces synthetic lethal cell death in glucose-deprived cell lines [66].

Third, to illustrate cSurvival's integrated workflow, and to show how cSurvival can bridge clinical and cell line studies, we assessed the Nrf2 [nuclear factor erythroid 2 related factor 2 (*NFE2L2*)]-Keap1 [Kelch-like erythroid cell-derived protein with CNC homology-associated protein 1 (*KEAP1*)] signaling pathway. Nrf2 is a master orchestrator of oxidative homeostasis and is primarily regulated by Keap1 [67]. In cancers, *KEAP1* is frequently mutated, resulting in constitutively active Nrf2 that protects cancer cells from chemotherapeutic agents and facilitates cancer progression [67]. For example, Nrf2 is aberrantly activated in ~30% of human lung cancers [68]. Using cSurvival, we found that expression or mutation of *NFE2L2* and *KEAP1* themselves showed no association with patient OS (Supplementary Figure 5, see Supplementary Data available online at *Briefings in Bioinformatics* online). However, high expression of genes in the Nrf2-antioxidant response element (ARE) pathway (WikiPathways: WP4357) correlated strongly with poor prognosis in lung adenocarcinoma patients (Figure 4A and B). Consistent with this clinical finding, in DepMap's experimental genetic perturbation screens, *KEAP1*-mutated lung cancer cell lines were more sensitive to *NFE2L2* knockout and knockdown (Figure 4C–F), consistent with *KEAP1* mutation being the main driver for oncogenic Nrf2 activation [69, 70]. Likewise, cell lines with higher expression of Nrf2–ARE pathway genes showed greater resistance to a potent oxidative stress inducer, menadione (BRD-K78126613-001-28-5) (Figure 5A and B) [71]. Moreover, cSurvival analysis showed that the male A549 and the female H2172 cell lines both harbor deactivating *KEAP1* mutations, manifest high Nrf2–ARE pathway expression, and show relatively high resistance to menadione (Figure 5). These findings are consistent with published reports that the A549 cell line is an excellent tool for studies on Nrf2 regulation and activity

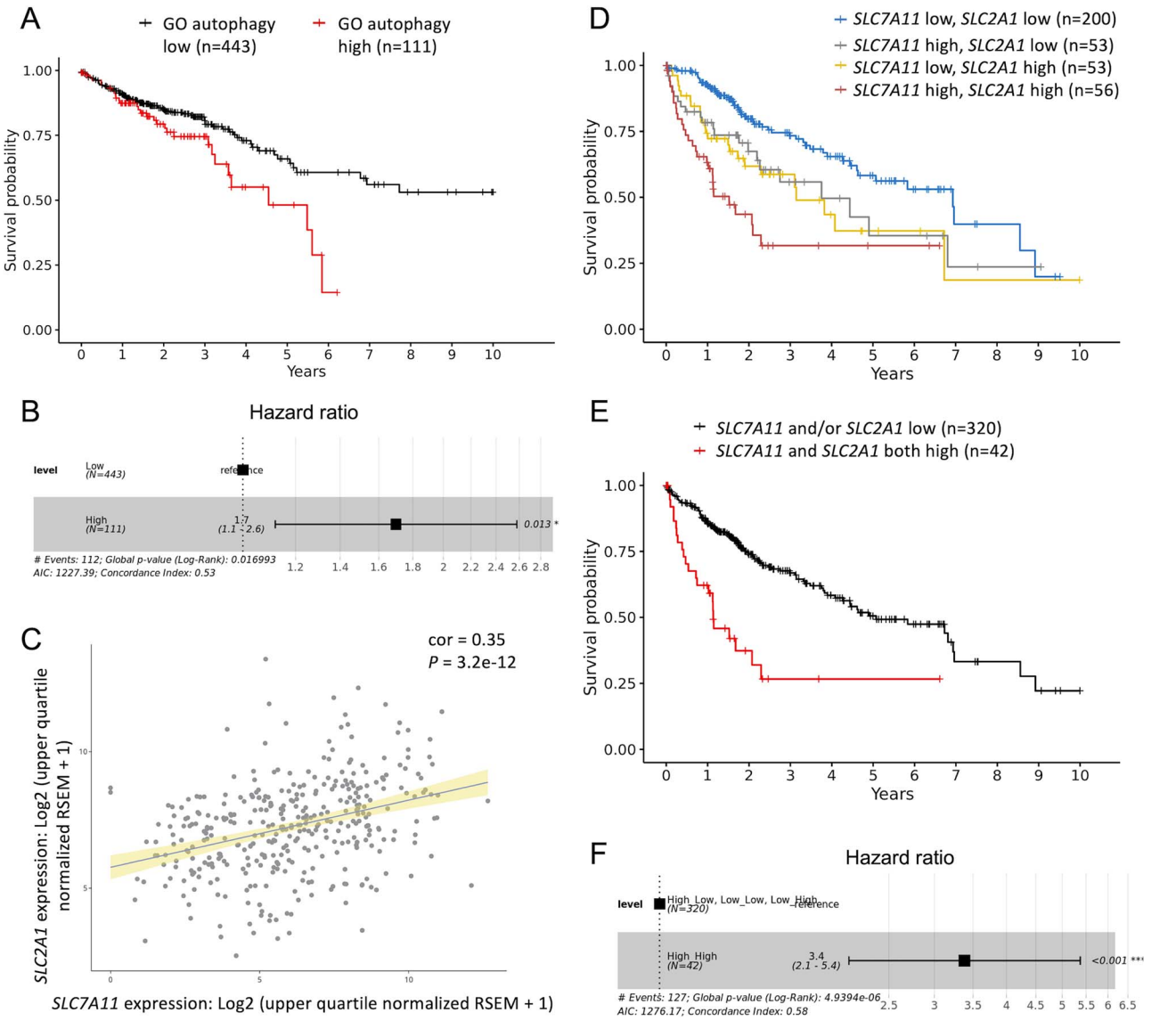

Figure 3. Case studies on autophagy-dependent survival in colorectal cancer and synergistic effects between high expression of *SLC7A11* and *SLC2A1* in liver cancer. The survival curves (**A**) and forest plots (**B**), censored at 10 years, show correlation between GO autophagy signature (GO: 0010506) and OS in colorectal cancers (TCGA-COAD and READ) ($P = 0.012$, P.adj = 0.03, KM log-rank; HR = 1.7, $P = 0.017$, P.adj = 0.05, Cox PH likelihood ratio). The scatter plot (**C**) shows a moderate correlation between expression of *SLC7A11* and *SLC2A1* (Pearson's correlation coefficient = 0.35, $P = 3.2e{-}12$). The survival curves (**D**), censored at 10 years, show significant differences in OS rates among patients with liver hepatocellular carcinoma (TCGA-LIHC) stratified by *SLC7A11* and *SLC2A1* expression levels ($P = 1e{-}07$, P.adj < 0.01, KM log-rank; $P = 1.7e{-}06$, P.adj < 0.01, Cox PH likelihood ratio). The survival curves (**E**) and forest plots (**F**), censored at 10 years, show a lower OS rate in TCGA-LIHC patients with high expression of both *SLC7A11* and *SLC2A1* than in patients with low expression of *SLC7A11* and/or *SLC2A1* ($P = 4.2e{-}08$, P.adj < 0.01, KM log-rank; HR = 3.38, $P = 4.9e{-}06$, P.adj < 0.01, Cox PH likelihood ratio). GO, gene ontology; COAD, colon adenocarcinoma; READ, rectum adenocarcinoma; *P*, *P*-value; P.adj, adjusted *P*-value; KM, Kaplan–Meier; PH, proportional-hazard; HR, hazard ratio; LIHC, liver hepatocellular carcinoma.

[72–74]; extending this, cSurvival analysis suggests that the female H2172 cell line is another excellent model to study Nrf2 function, and can be used in conjunction with A549 to study sex-specific effects (Figure 5).

Together, these analyses demonstrate cSurvival's unique ability to support biomarker prognosis and interaction analysis via gene- and GS-level approaches, and to facilitate integrating clinical and experimental biomedical studies. See the cSurvival user guide (https://tau.cmmt.ubc.ca/cSurvival/help.html) for:

(i) detailed steps for the reported case studies;

(ii) application cases for single biomarker analysis with each type of molecular data;

(iii) application cases for biomarker interaction analysis with two different types of molecular data, including DNA methylation and miRNA expression, mutation and miRNA expression, mutation and CN variation, and GS expression and gene expression.

Discussion

Cancer arises from accumulated genetic and epigenetic alterations, creating interactions that endow cancer cells with growth and survival advantages. Such functional relationships happen not only between genes, but also between GSs, or between genes and GSs. Correspondingly,

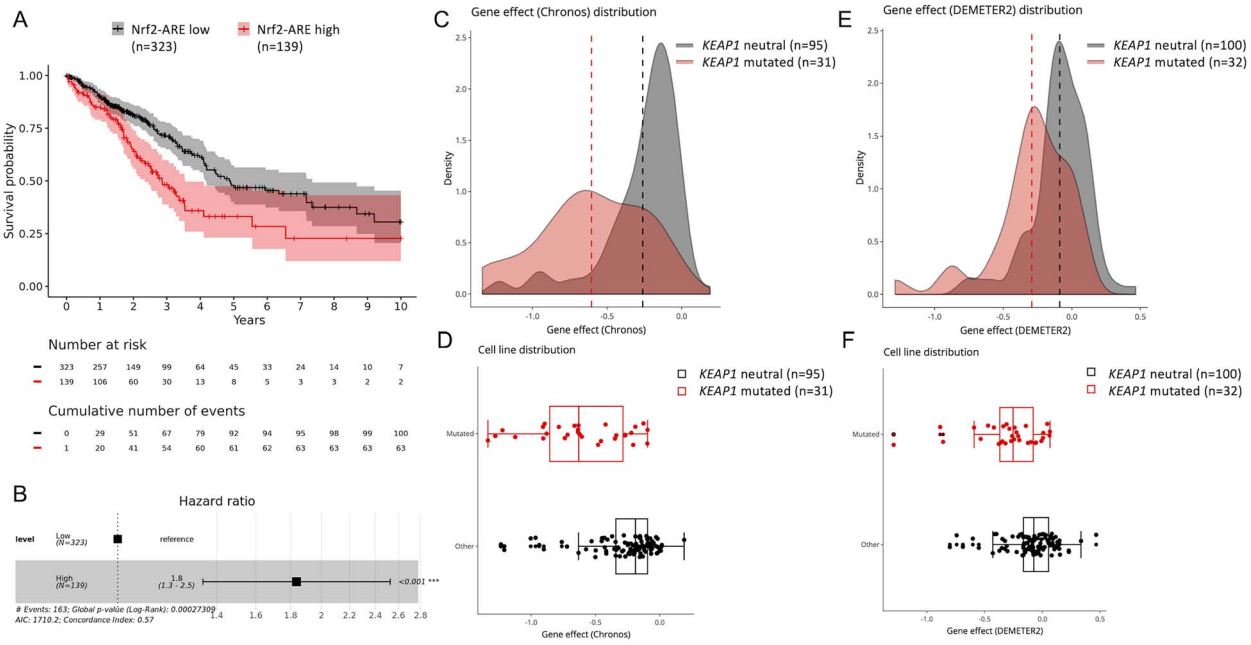

Figure 4. High expression of the Nrf2–ARE pathway correlates with poor prognosis and *KEAP1* mutation in lung cancers. The survival curves (**A**) and forest plot (**B**), censored at 10 years, show correlation between high expression of the Nrf2–ARE pathway (WikiPathways: WP4357) and poor prognosis in TCGA-LUAD cases ($P = 0.00015$, $P.adj < 0.01$, KM log-rank; HR = 1.84, $P = 0.00027$, $P.adj < 0.01$, Cox PH likelihood ratio). Shades reflect 95% confidence intervals in survival curves in (**A**). The density and box plots show lung cancer cell lines with an inactivating *KEAP1* mutation (red) being more sensitive to *NFE2L2* CRISPR-Cas9 knockout ($P = 4.1e{-}07$, two-tailed Wilcoxon test) (**C**, **D**) and RNAi knockdown ($P = 8.4e{-}05$, two-tailed Wilcoxon test) (**E, F**). ARE, antioxidant response element; LUAD, lung adenocarcinoma; *P*, *P*-value; *P.adj*, adjusted *P*-value; KM, Kaplan–Meier; HR, hazard ratio; PH, proportional-hazard; RNAi, RNA interference; Chronos, an algorithm for inferring gene knockout fitness effects; DEMETER2, gene dependency estimates for RNAi datasets.

regimens that combine multiple drugs that target different genes/GSs have emerged as more effective and less toxic than monotherapy approaches [30, 31]. Pinpointing genetic interactions in cancers is thus an important research goal.

Here, we developed cSurvival, an open-source framework to identify potential genetic interactions and survey preclinical cell line tools. cSurvival offers innovative algorithms that can assess interactions between many types of cancer biomarkers, including GSs, as well as a curated database that combines clinical and experimental data to generate synergistic biological insights. As shown by the three case studies, cSurvival sheds light on genetic interactions in cancers and facilitates the identification of preclinical cell line tools for mechanistic studies.

For long-term sustainability, we have automated data extraction from the NCI GDC (https://gdc.cancer.gov) [48] and will continuously follow consortium efforts/GDC evolution, so that cSurvival uses up-to-date processing pipelines, human reference genomes and gene annotations.

Despite the advances described herein, cSurvival has certain limitations. For example, the current version is designed for one query (a gene or a gene combination) at a time. For continuous genomic predictors, we have corrected the inflated false positive rate due to exploring multiple cutoff values for each gene/gene combination. However, when testing many genes/gene combinations, users should apply another layer of multiple testing adjustment. This is similar to a *t*-test, which is designed to compare two groups; when using a *t*-test to test many pairs of groups, a multiple testing adjustment needs to be applied outside of these *t*-tests. For example, when using cSurvival to test ten genes/gene combinations, and each test returns a *P*-value (or adjusted *P*-value for continuous genomic predictors), one can do the outer multiple testing adjustment using the *P. adjust* function in R [63]. Note that a Bonferroni correction may be more appropriate for a smaller number of genes/gene combinations, while a false discovery rate (FDR) correction may be more appropriate for a larger number. The *P. adjust* function offers a range of correction methods, including Bonferroni and FDR. Future versions of cSurvival may extend to batch analysis with corrections for multiple testing. In addition, cohorts from TCGA and TARGET were largely from North America. To address ethnicity [75] heterogeneities in human populations, we plan to expand the cSurvival database, incorporating resources such as the International Cancer Genome Consortium [76] and cohorts from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). Future versions of cSurvival may also address the challenges of immunogenomics [77] and may integrate cellular signatures from the Connectivity Map [78, 79].

Although we built cSurvival as a web resource to analyze molecular biomarkers in published cancer datasets, its source code could be adapted to analyze unpublished or protected data locally. Its algorithms for searching

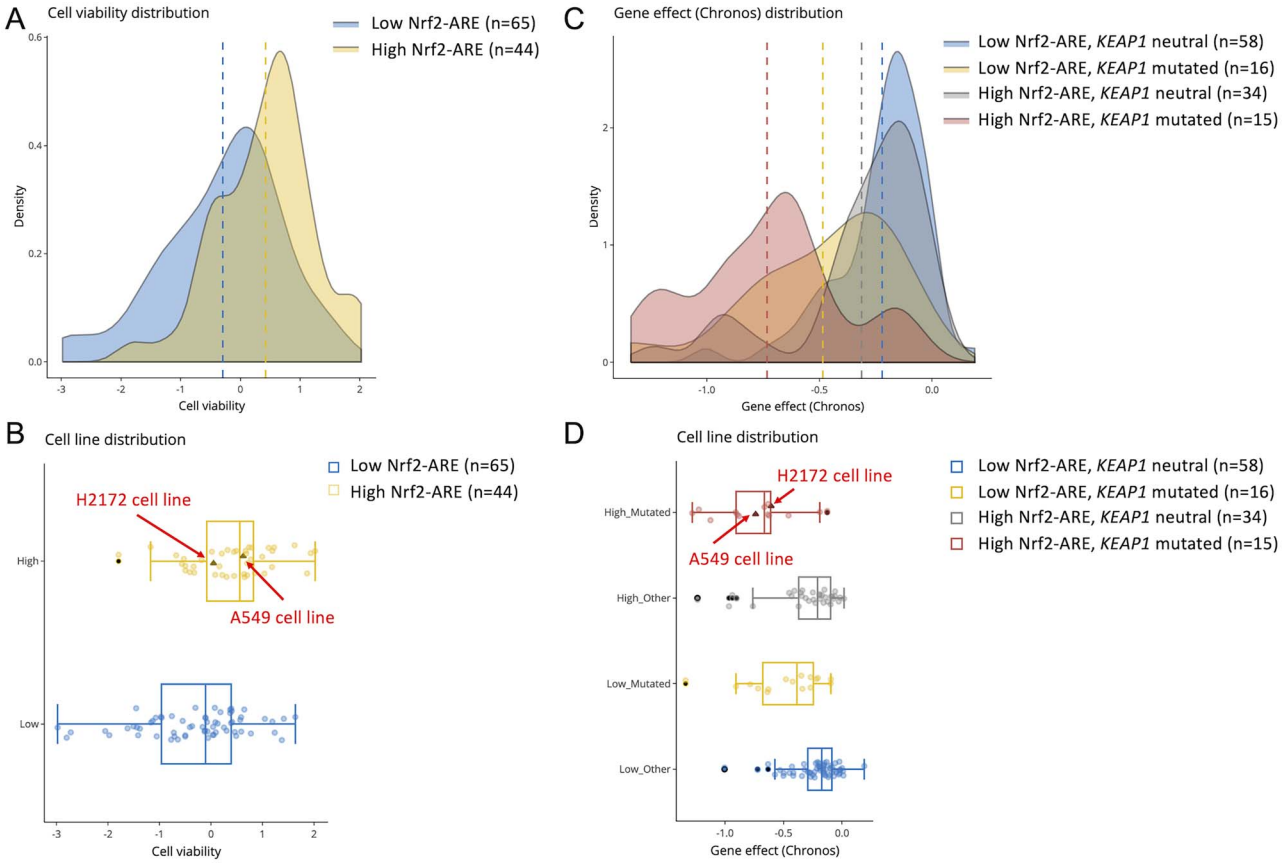

Figure 5. The male A549 and the female H2172 cell lines are top candidates for studies on Nrf2 regulation and activity. The density (**A**) and box (**B**) plots show that lung cancer cell lines with higher expression levels of Nrf2–ARE pathway genes exhibit higher resistance to the oxidative stress inducer menadione (BRD-K78126613-001-28-5, $P = 0.00014$, $P.adj < 0.01$, two-tailed Wilcoxon test). The density (**C**) and box (**D**) plots show that lung cancer cell lines with different *KEAP1* mutation status and expression levels of Nrf2–ARE pathway genes show sensitivity differences to *NFE2L2* knockout ($P = 2.6e{-}06$, $P.adj < 0.01$, Kruskal–Wallis rank sum test). The A549 and the H2172 cell lines are highlighted in a triangle shape with a darker color and labeled in the box plots. ARE, antioxidant response element; *P*, *P*-value; *P*.adj, adjusted *P*-value; Chronos, an algorithm for inferring gene knockout fitness effects.

for optimal cutoffs for interaction analysis could also be applied to other types of (bio)markers and/or other diseases, e.g. to data from biomedical imaging [80] or drug cocktail effect assessment [81], which sometimes involve combinations of continuous and/or categorical variables.

In summary, cSurvival offers a curated database and innovative analytical pipelines to examine cancer biomarkers at high resolution. It complements existing resources such as cBioPortal [9] and enhances mechanistic investigation of malignancy etiologies using clinical correlations. Its intuitive yet flexible web interface makes it a valuable tool for experimental and clinical researchers alike.

Key Points

- We developed cSurvival, an advanced framework using clinical correlations to study biomarker interactions in cancers, with source code and curated datasets freely available for download.
- cSurvival includes new algorithms to identify optimal cutoffs for two continuous predictors to stratify patients into risk groups, enabling, for the first time, joint analysis with two genomic predictors.
- cSurvival allows survival analysis at the gene set (GS) level with comprehensive and up-to-date GS libraries.
- The cSurvival pipeline integrates clinical outcomes data and experimental cancer cell line data to generate synergistic biological insights and to mine for appropriate preclinical cell line tools.
- cSurvival is built on a manually curated cancer outcomes database.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics* online.

Acknowledgement

We thank P.W. Laird (Van Andel Institute, Grand Rapids, MI), T. Lichtenberg (University of Chicago, Chicago, IL), C.A. Maxwell (The University of British Columbia, Vancouver, BC), H. Shen (Van Andel Institute, Grand

Rapids, MI), Z. Zhou (The Chinese University of Hong Kong, Shatin, HK) and Taubert lab members for critical comments on the manuscript. Figure 1 was created with BioRender.com, Toronto, Canada.

Funding

Canadian Institutes of Health Research (CIHR; PJT-153199); Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-05133 to S.T., RGPIN-2017-04722 to X.Z.); Canada Research Chair (No. 950231363 to X.Z.); British Columbia Children's Hospital Research Institute Investigator Grant Award Program award (to S.T.); Canada Research Chairs program (to S.J.M.J.). Funding for open access charges: Canadian Institutes of Health Research (CIHR; PJT-153199).

Data Availability

cSurvival (https://tau.cmmt.ubc.ca/cSurvival) is free and open to all users and has no login requirement. The source code for building cSurvival's framework and database is available at GitHub (https://github.com/easygsea/cSurvival.git).

References

1. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*, 3rd edn, Springer, New York 2012.
2. Schober P, Vetter TR. Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesth Analg* 2018;**127**:792–8.
3. Győrffy B, Lanczky A, Eklund AC, *et al.* An online survival analysis tool to rapidly assess the effect of 22 277 genes on breast cancer prognosis using microarray data of 1809 patients. *Breast Cancer Res Treat* 2010;**123**:725–31.
4. Zheng H, Zhang G, Zhang L, *et al.* Comprehensive review of web servers and bioinformatics tools for cancer prognosis analysis. *Front Oncol* 2020;**10**:68.
5. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer Genome atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* 2015;**19**:A68–77.
6. Liu J, Lichtenberg T, Hoadley KA, *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018;**173**:400, e11–6.
7. Genome OC. Therapeutically applicable research to generate effective treatments. *Off Cancer Genomics* 2013.
8. Cerami1 E, Gao J, Dogrusoz U, *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4.
9. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1.
10. Zhang G, Wang Q, Yang M, *et al.* OSpaad: an online tool to perform survival analysis by integrating gene expression profiling and long-term follow-up data of 1319 pancreatic carcinoma patients. *Mol Carcinog* 2020;**59**:304–10.
11. Park S-J, Yoon B-H, Kim S-K, *et al.* GENT2: an updated gene expression database for normal and tumor tissues. *BMC Med Genomics* 2019;**12**:101.
12. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. *BMC Cancer* 2014;**14**:970.
13. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, *et al.* SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* 2013;**8**:e74250.
14. Mizuno H, Kitada K, Nakai K, *et al.* PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genom* 2009;**2**:18.
15. Nagy Á, Munkácsy G, Győrffy B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep* 2021;**11**:6047.
16. Liu C-J, Hu F-F, Xia M-X, *et al.* GSCALite: a web server for gene set cancer analysis. *Bioinformatics* 2018;**34**:3771–2.
17. Chandrashekar DS, Bashel B, Balasubramanya SAH, *et al.* UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia N Y N* 2017;**19**:649–58.
18. Tang Z, Kang B, Li C, *et al.* GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acid Res* 2019;**47**:W556–60.
19. Han S, Kim D, Kim Y, *et al.* CAS-viewer: web-based tool for splicing-guided integrative analysis of multi-omics cancer data. *BMC Med Genom* 2018;**11**:25.
20. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Comput Sci* 2016;**2**:e67.
21. Goldman MJ, Craft B, Hastie M, *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020;**38**:675–8.
22. Zhang J, Baran J, Cros A, *et al.* International cancer Genome consortium data portal—a one-stop shop for cancer genomics data. *Database J Biol Databases Curation* 2011;**2011**:bar026.
23. Gentles AJ, Newman AM, Liu CL, *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* 2015;**21**:938–45.
24. Wong NW, Chen Y, Chen S, *et al.* OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics* 2018;**34**:713–5.
25. Aure MR, Fleischer T, Bjørklund S, *et al.* Crosstalk between microRNA expression and DNA methylation drives the hormone-dependent phenotype of breast cancer. *Genome Med* 2021;**13**:72.
26. Bradburn MJ, Clark TG, Love SB, *et al.* Survival analysis part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer* 2003;**89**:605–11.
27. Ma X, Huang R, Wu X, *et al.* Dualmarker: a flexible toolset for exploratory analysis of combinatorial dual biomarkers for clinical efficacy. *BMC Bioinformat* 2021;**22**:127.
28. Magen A, Sahu AD, Lee JS, *et al.* Beyond synthetic lethality: charting the landscape of pairwise gene expression states associated with survival in cancer. *Cell Rep* 2019;**28**:938–948.e6.
29. Yu J, Zhou D, Yang X, *et al.* TRIB3-EGFR interaction promotes lung cancer progression and defines a therapeutic target. *Nat Commun* 2020;**11**:3660.
30. Mokhtari RB, Homayouni TS, Baluch N, *et al. Combination therapy in combating cancer*, *Oncotarget* 2017;**8**:38022–43.
31. Ledford H. Cocktails for cancer with a measure of immunotherapy. *Nature* 2016;**532**:162–4.
32. Zhang J, Zhang S. Discovery of cancer common and specific driver gene sets. *Nucleic Acid Res* 2017;**45**:e86.
33. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.

34. Maleki F, Ovens K, Hogan DJ, *et al.* Gene set analysis: challenges, opportunities, and future research. *Front Genet* 2020;**11**:654.
35. Zheng X, Amos CI, Frost HR. Comparison of pathway and gene-level models for cancer prognosis prediction. *BMC Bioinformatics* 2020;**21**:76.
36. Rehman SK, Haynes J, Collignon E, *et al.* Colorectal cancer cells enter a diapause-like DTP state to survive chemotherapy. *Cell* 2021;**184**:226, e21–42.
37. Goeman JJ, Oosting J, Cleton-Jansen A-M, *et al.* Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005;**21**:1950–7.
38. Tsherniak A, Vazquez F, Montgomery PG, *et al.* Defining a cancer dependency map. *Cell* 2017;**170**:564, e16–76.
39. Meyers RM, Bryan JG, McFarland JM, *et al.* Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* 2017;**49**:1779–84.
40. Dempster JM, Rossen J, Kazachkova M, *et al.* Extracting biological insights from the project Achilles Genome-scale CRISPR screens in cancer cell lines. *bioRxiv* 2019;720243.
41. Ghandi M, Huang FW, Jané-Valbuena J, *et al.* Next-generation characterization of the cancer cell line Encyclopedia. *Nature* 2019;**569**:503–8.
42. Cheng X, Yan J, Liu Y, *et al.* eVITTA: a web-based visualization and inference toolbox for transcriptome analysis. *Nucleic Acid Res* 2021;**49**:W207–15.
43. Hoadley KA, Yau C, Hinoue T, *et al.* Cell-of-origin patterns dominate the molecular classification of 10 000 Tumors from 33 types of cancer. *Cell* 2018;**173**:291, e6–304.
44. Ellrott K, Bailey MH, Saksena G, *et al.* Scalable Open Science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 2018;**6**:271, e7–81.
45. Finak G, McDavid A, Yajima M, *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;**16**:278.
46. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
47. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acid Res* 2017;**45**:e22.
48. Heath AP, Ferretti V, Agrawal S, *et al.* The NCI genomic data commons. *Nat Genet* 2021;**53**:257–62.
49. Colaprico A, Silva TC, Olsen C, *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;**44**:e71.
50. Silva TC, Colaprico A, Olsen C, *et al.* TCGA workflow: analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* 2016;**5**:1542.
51. Mounir M, Lucchetta M, Silva TC, *et al.* New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput Biol* 2019;**15**:e1006701.
52. Carlson M. *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.11.4, 2020.
53. Corsello SM, Nagari RT, Spangler RD, *et al.* Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* 2020;**1**:235–48.
54. Nusinow DP, Szpyt J, Ghandi M, *et al.* Quantitative proteomics of the cancer cell line Encyclopedia. *Cell* 2020;**180**:387–402.e16.
55. Corsello SM, Bittker JA, Liu Z, *et al.* The drug repurposing hub: a next-generation drug library and information resource. *Nat Med* 2017;**23**:405–8.
56. Therneau TM, Grambsch PM. *Modeling Survival Data. Extending the Cox Model*. Springer, Berlin, 2000.
57. Lausen B, Schumacher M. Maximally selected rank statistics. *Biometrics* 1992;**48**:73–85.
58. Altman DG, Lausen B, Sauerbrei W, *et al.* Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI J Natl Cancer Inst* 1994;**86**:829–35.
59. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints using two ROC based criteria. *Am J Epidemiol* 2006;**163**:670–5.
60. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res Off J Am Assoc Cancer Res* 2004;**10**:7252–9.
61. Budczies J, Klauschen F, Sinn BV, *et al.* Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker Cutoff optimization. *PLoS One* 2012;**7**:e51862.
62. Hilsenbeck SG, Clark GM. Practical P-value adjustment for optimally selected cutpoints. *Stat Med* 1996;**15**:103–12.
63. R Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2020.
64. Kassambara A, Kosinski M, Biecek P. Survminer: drawing survival curves using 'ggplot2'. *R package version* 049 2021. https://rpkgs.datanovia.com/survminer/index.html.
65. Wickham H, Averick M, Bryan J, *et al.* Welcome to the Tidyverse. *J Open Source Softw* 2019;**4**:1686.
66. Joly JH, Delfarah A, Phung PS, *et al.* A synthetic lethal drug combination mimics glucose deprivation–induced cancer cell death in the presence of glucose. *J Biol Chem* 2020;**295**:1350–65.
67. Jaramillo MC, Zhang DD. The emerging role of the Nrf2–Keap1 signaling pathway in cancer. *Genes Dev* 2013;**27**:2179–91.
68. Collisson EA, Campbell JD, Brooks AN, *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;**511**:543–50.
69. de la Vega MR, Chapman E, Zhang DD. NRF2 and the hallmarks of cancer. *Cancer Cell* 2018;**34**:21–43.
70. Singh A, Misra V, Thimmulappa RK, *et al.* Dysfunctional KEAP1–NRF2 interaction in non-small-cell lung cancer. *PLoS Med* 2006;**3**:e420.
71. Thor H, Smith MT, Hartzell P, *et al.* The metabolism of menadione (2-methyl-1,4-naphthoquinone) by isolated hepatocytes. A study of the implications of oxidative stress in intact cells. *J Biol Chem* 1982;**257**:12419–25.
72. Ohta T, Iijima K, Miyamoto M, *et al.* Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. *Cancer Res* 2008;**68**:1303–9.
73. Wang X-J, Sun Z, Villeneuve NF, *et al.* Nrf2 enhances resistance of cancer cells to chemotherapeutic drugs, the dark side of Nrf2. *Carcinogenesis* 2008;**29**:1235–43.
74. Gong M, Li Y, Ye X, *et al.* Loss-of-function mutations in KEAP1 drive lung cancer progression via KEAP1/NRF2 pathway activation. *Cell Commun Signal CCS* 2020;**18**:98.
75. Yuan J, Hu Z, Mahal BA, *et al.* Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer Cell* 2018;**34**:549–560.e9.
76. International network of cancer genome projects. *Nature* 2010;**464**:993–8.
77. Thorsson V, Gibbs DL, Brown SD, *et al.* The immune landscape of cancer. *Immunity* 2018;**48**:812–830.e14.
78. Subramanian A, Narayan R, Corsello SM, *et al.* A next generation connectivity map: L1000 platform and the first 1 000 000 profiles. *Cell* 2017;**171**:1437–1452.e17.

79. Lamb J, Crawford ED, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.
80. Wang JW, Williams M. Repurposing routine imaging for cancer biomarker discovery using machine learning. In: *Handbook of Artificial Intelligence in Healthcare, Vol 1 - Advances and Applications*, Springer International Publishing 2022, 153–76.
81. Gibbons JA, de Vries M, Krauwinkel W, *et al*. Pharmacokinetic drug interaction studies with enzalutamide. *Clin Pharmacokinet* 2015;**54**:1057–69.
82. Gu Z, Gu L, Eils R, *et al*. Circlize implements and enhances circular visualization in R. *Bioinformatics* 2014;**30**: 2811–2.