ExSurv: A Web Resource for Prognostic Analyses of Exons Across Human Cancers Using Clinical Transcriptomes

Seyedsasan Hashemikhabir[1], Gungor Budak[1] and Sarath Chandra Janga[1–3]

[1]Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, IN, USA. [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 5021 Health Information and Translational Sciences (HITS), Indianapolis, IN, USA. [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, Indianapolis, IN, USA.

Supplementary Issue: Integrative Analysis of Cancer Genomic Data

ABSTRACT: Survival analysis in biomedical sciences is generally performed by correlating the levels of cellular components with patients' clinical features as a common practice in prognostic biomarker discovery. While the common and primary focus of such analysis in cancer genomics so far has been to identify the potential prognostic genes, alternative splicing – a posttranscriptional regulatory mechanism that affects the functional form of a protein due to inclusion or exclusion of individual exons giving rise to alternative protein products, has increasingly gained attention due to the prevalence of splicing aberrations in cancer transcriptomes. Hence, uncovering the potential prognostic exons can not only help in rationally designing exon-specific therapeutics but also increase specificity toward more personalized treatment options. To address this gap and to provide a platform for rational identification of prognostic exons from cancer transcriptomes, we developed ExSurv (https://exsurv.soic.iupui.edu), a web-based platform for predicting the survival contribution of all annotated exons in the human genome using RNA sequencing-based expression profiles for cancer samples from four cancer types available from The Cancer Genome Atlas. ExSurv enables users to search for a gene of interest and shows survival probabilities for all the exons associated with a gene and found to be significant at the chosen threshold. ExSurv also includes raw expression values across the cancer cohort as well as the survival plots for prognostic exons. Our analysis of the resulting prognostic exons across four cancer types revealed that most of the survival-associated exons are unique to a cancer type with few processes such as cell adhesion, carboxylic, fatty acid metabolism, and regulation of T-cell signaling common across cancer types, possibly suggesting significant differences in the posttranscriptional regulatory pathways contributing to prognosis.

KEYWORDS: survival, splicing, exon expression, cancer, posttranscriptional regulation

SUPPLEMENT: Integrative Analysis of Cancer Genomic Data

CITATION: Hashemikhabir et al. ExSurv: A Web Resource for Prognostic Analyses of Exons Across Human Cancers Using Clinical Transcriptomes. *Cancer Informatics* 2016:15(S2) 17–24 doi: 10.4137/CIN.S39367.

TYPE: Technical Advance

RECEIVED: April 01, 2016. **RESUBMITTED:** May 25, 2016. **ACCEPTED FOR PUBLICATION:** May 30, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,425 words, excluding any confidential comments to the academic editor.

FUNDING: SCJ acknowledges support from the School of Informatics and Computing at IUPUI in the form of startup funds. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: scjanga@iupui.edu

Introduction

Survival analysis is a statistical approach to evaluate an event or study where the outcome is based on the vital status of the samples.[1] Cancer treatment and clinical trials are two common use cases for analyzing the survival of the samples based on the given treatment. However, over the recent years, several major web servers that calculate the survival contribution of mRNAs by employing publicly available clinically annotated breast cancer microarray data have become available including Kaplan–Meier plotter,[2] GOBO,[3] RecurrenceOnline,[4] and bc-GenExMiner.[5] PrognoScan[6] is one of the first studies integrating microarray expression data from Gene Expression Omnibus (GEO) for multiple cancer types to predict the prognostic mRNAs. As a result of increase in the availability of sequencing-based genomic data along with clinical annotations across cancer types from various consortiums such as The Cancer Genome Atlas (TCGA)[7–9] (http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (ICGC) (https://icgc.org/), cancer prognostic biomarker identification and comparison among them has become the major outcomes of various multi-omic studies. SurvExpress[10] and PROGgene[11] have integrated multiple RNA sequencing (RNA-seq) expression profiles from GEO datasets and report the contribution of an mRNA expression as a prognostic biomarker. SurvNet[12] is an online service for identifying the network-based biomarkers associated with clinical information. It is preloaded with TCGA data and reports the survival significance in mRNA and protein levels. cBioPortal[13] is a comprehensive web-based platform that integrates several cancer genomic datasets including those from TCGA. However, all of the listed services only report the prognostic properties and survival contribution at mRNA level. Alternative splicing is a posttranscriptional gene

regulatory mechanism that contributes to different protein products due to alterations in the combination of exons originating from the same gene.[14] Exon inclusion/skipping is a class of a splicing event, wherein an exon is included or missed in the final isoform product.[14] Hence, identifying the potential prognostically involved/missed exons in cancer genomes would be beneficial for a better understanding of not only the involvement of posttranscriptional regulatory alterations in cancers but would also provide novel insights into potential exon-level functions in tumor evolution. For example, SRSF2, an RNA-binding protein, mutations drive recurrent mis-splicing of key hematopoietic regulators such as EZH2 in patients with myelodysplastic syndromes.[15] Similarly, mutant U2AF1, an RNA-binding factor, alters downstream gene isoform expression by altering the splicing mechanism.[15] A recent study[16] analyzed genome-wide patterns of RNA splicing across 805 matched tumor and normal control samples from 16 distinct cancer types to identify signals of abnormal cancer-associated splicing. This study reports intron retention, a category of alternative splicing, to be common across cancers even in the absence of mutations directly affecting the RNA splicing machinery. In light of several recent studies providing support for extensive rewiring of posttranscriptional regulatory networks in multiple cancer types,[16–19] there is an immediate need for resources that can enable rapid mining and validation of high-confident prognostic alternative splicing events in cancer transcriptomes.

In this study, we present ExSurv (https://exsurv.soic.iupui.edu/), which to our knowledge is the first online web server that provides exon-level survival significance by using the RNA-seq expression datasets and the associated clinical metadata for four cancer types from the TCGA project[7,9] (http://cancergenome.nih.gov/). We precalculated the prognostic significance of more than 600,000 annotated exons in Ensembl[20] using survival package[21] in R across the four cancer types. We stored the TCGA clinical data, exon survival *P*-values, and the expression of significant exons, for visualizing the survival curves per exon in a MySQL database. A PHP/R backend and a JavaScript frontend were employed for designing the interface and calling R visualization libraries. We also compare the overlap in the prognostic exons and their functional overlap across cancer types studied here.

Materials and Methods

Cancer types and the corresponding number of cancer samples analyzed in this study from the TCGA project. We selected four cancer types, namely, breast invasive carcinoma (BRCA), liver hepatocellular carcinoma (LIHC), glioblastoma (GBM), and kidney renal papillary cell carcinoma (KIRP), and quantified the exon expression levels for patients with accessible raw RNA-seq data. It is important to note that TCGA is a collaborative effort; hence, the generated data could be from different dates and platforms. However, it is not suggested to do batch effect correction since it may bring additional noises to the analysis. The majority of TCGA cancer patients are *White* males (except breast cancer patients) and aged from 40 to 80 years at the time of first time cancer diagnosis (see Table 1). While there are several clinical features available for each cancer type, to simplify the calculations and increase the performance of the web server given the large number of annotated human exons employed in the analysis, we only considered the prognostic effect of exons among all the samples of a cancer type without controlling for covariates such as gender, age, or grade for a given cancer type. It is important to note that most existing web servers performing survival analysis for multiple cancer types also employ a similar approach to avoid sample size artifacts when multiple covariates are considered.

Workflow for processing RNA-seq data and quantification of exonic expression levels. Hierarchical indexing for spliced alignment of transcripts (HISAT)[22] is a highly efficient alignment tool for aligning short reads from RNA-seq experiments onto reference genome. HISAT claims to be the fastest alignment system currently available, with better accuracy than most other methods. We ran HISAT with – dta parameter options to align each of the RNA-seq samples to generate corresponding Binary Alignment/Map (BAM) files using the human genome reference 38 annotations obtained from Ensembl.[23] We then quantified the expression of exons using StringTie,[24] a computational method that applies a network flow algorithm together with optional de novo assembly, and estimated the multimap corrected number of reads for every annotated exon in human genome build 38. To normalize the number of reads mapped to genomic regions, transcript per million is proposed by Li and Dewey,[25] which corresponds to normalized read count value based on the length of the transcript isoform under consideration and the total number of mapped reads in the whole genome. We employed a similar approach to quantify the exon expression, normalized by the length of exon and the total number of the reads mapped to the whole genome. Here, $E(x_i)$ is the expression of ith exon. C and L functions for an exon represent the number of multimap corrected reads mapping to the exon and length of the exon, respectively, across the complete transcriptome of a sample.

$$E\left(x_i\right)=\frac{C\left(x_i\right)}{L\left(x_i\right)}\left(\frac{1}{\sum_j \frac{C\left(x_j\right)}{L\left(x_j\right)}}\right)\times 10^6$$

Clinical metadata from TCGA samples employed for survival analysis in ExSurv. Clinical data are crucial for survival analysis. TCGA (http://cancergenome.nih.gov/) provides both sequencing data and the corresponding clinical information for most of the patients for each cancer type included in ExSurv. The clinical data included the vital status of a patient, number of days from the first day of medication

Table 1. Number of cancer patient samples employed for each cancer type in ExSurv.

CANCER	NUMBER OF ANALYZED SAMPLES	NUMBER OF PATIENTS WITH CLINICAL DATA	GENDER (M/F)	RACE	AGE
Breast invasive carcinoma (BRCA)	1040	1099	12/1087	W:758 B:183 A:61	>80: 55 60–80: 438 40–60: 508 <40: 98
Glioblastoma (GBM)	174	594	365/229	W:505 B:51 A:13	>80: 22 60–80: 246 40–60: 249 <40: 70
Kidney renal papillary cell carcinoma (KIRP)	287	287	211/76	W:207 B:61 A:5	>80: 14 60–80: 138 40–60: 119 <40: 16
Liver hepatocellular carcinoma (LIHC)	368	368	250/118	W:180 B:17 A:160	>80: 10 60–80: 184 40–60: 143 <40: 27

Notes: Other clinical parameters and the corresponding composition of the cancer cohorts are shown. We also categorized the age of patients into four categories with intervals of 20 years for reference.
Abbreviations: (Race column represents) W, White; B, Black; A, Asian.

till the last follow-up, gender, cancer subtypes, and other cancer-specific information. We included vital status and the number of days since last follow-up in our survival analysis.

Survival analysis for exons to identify prognostic markers in a cancer type. We employed survival package in R[26,27] to estimate the contribution of a given exon to survival of patients in a cancer type based on their expression. We divided the cancer patients based on the expression of a given exon into two groups (ie, High and Low). The group that was labeled as *High* are the patients where the expression of the exon is above the median expression of the same exon among all the specific cancer type patients. Similarly, we labeled the patients *Low* if the expression of the exon is below the median expression level of that exon, among all the patients studied in a cancer type. We applied Kaplan–Meier estimate of survival with log-rank test and Cox proportional regression model to measure the significance of difference between the High and Low groups in a given cancer type. To increase the stringency and to reduce the potential false positives, we introduced an additional approach by redefining the High group as patients where the expression of an exon is more than the third quartile expression of the same exon, among all the specific cancer types, and Low group if the expression of the exon is less than the first quartile expression level of that exon, among all the patients studied in a cancer type. We applied similar statistical models to evaluate the survival contribution of the exons in every cancer type. We corrected the results for the false discovery rate (FDR) using Benjamini–Hochberg procedure. The exons that exhibited a survival probability of at least 0.05 as a result of the test were considered as significant prognostic biomarkers for downstream analysis, and the corresponding FDR values are also provided as output on the web server for each exon and the same results are available in the MySQL dump for local use.

Construction of the database backend for serving the user interface of ExSurv. We aligned the raw RNA-seq datasets from the cancer samples against HG38 and used the exon annotations from Ensembl database,[23] as discussed above. The exon survival information and metadata associated with cancer patients were organized into separated tables. Likewise, the genomic annotations were organized into separate tables (Supplementary Fig. 1 for a schema of the ExSurv database). The design of the database enables updating either the annotations or precomputed datasets or both without extensive dependencies, thereby speeding up updates to the ExSurv when there are changes in the datasets without the need for a structural change in the database. We stored the data required for our service in a MySQL database. We also developed PHP scripts to handle user queries and generate exon-level survival plots by integrating preprocessed data from the database to perform dynamic survival analysis using the queried data in R with survival package.[26]

Framework employed for building the user interface of ExSurv. We implemented a simple interface for users to search for their gene of interest and visualize the exon-level survival analysis results in the selected cancer type. We use JavaScript integrated with PHP scripts to query the database and retrieve the expression results that are then provided as input for survival analysis in R. ExSurv provides the users to not only search for genes of interest and naturally all the exons associated with it but to also limit to certain level of significance. The results of the search are listed as survival plots for each exon corresponding to the gene that matches the query (eg, a gene symbol or Ensembl gene ID). The metadata related to that exon in the context of the cancer samples analyzed is also given next to its survival plot. The users can export the exon-level survival plots for storing them locally. We also implemented an export functionality for downloading

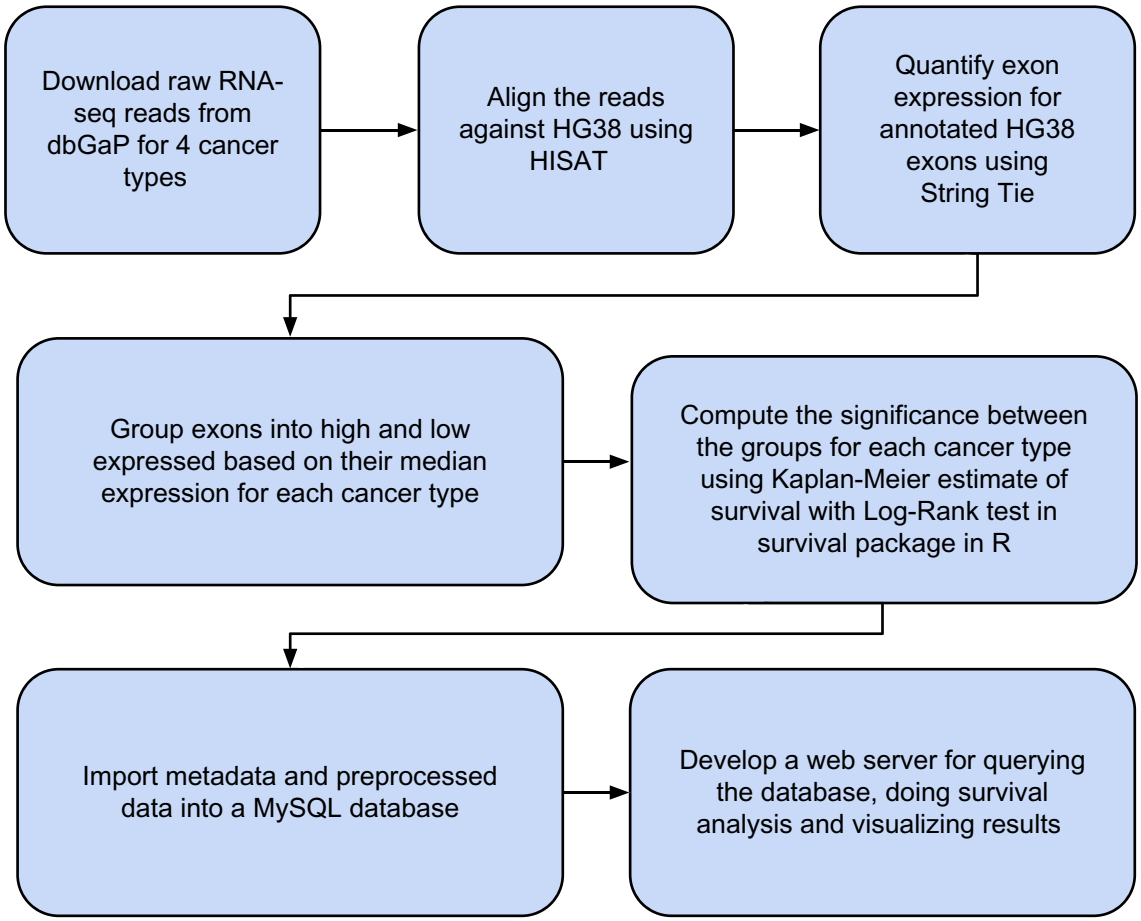

Figure 1. Major steps involved in the generation of exon-level survival estimates in the ExSurv database and visualization engine. As outlined in the flowchart, our analysis pipeline comprises obtaining the raw RNA-seq datasets for more than 1800 patients spanning four different cancer types along with their clinical metadata. Raw RNA-seq was processed using HISAT[22] aligner to generate BAM formatted files, and StringTie[24] was used for exon-level expression quantification. Exon expression level across patients in a given cancer type along with clinical information on the follow-up time periods were used to generate Kaplan–Meier plots and corresponding *P*-values from log-rank tests. A MySQL database was developed to store the exon-level expression as well as the results of the survival analysis to facilitate dynamic querying and dissemination of the data. A web server was developed to show the Kaplan–Meier plots and corresponding expression data employed for generating them.

raw expression data as well as the associated metadata across cancer samples analyzed for each exon. ExSurv does not require users to log in, and it is freely available on https://exsurv.soic.iupui.edu/.

Results and Discussion

Overview of datasets and approach employed for building ExSurv. Figure 1 shows the major steps involved in the generation of exon-level survival estimates as implemented in ExSurv database and visualization engine. Briefly, we obtained the raw RNA-seq datasets for individual patients from CGHub portal (https://cghub.ucsc.edu/) after an approval process from the dbGaP (http://dbgap.ncbi.nlm.nih.gov/) to access the raw datasets for various cancer types. Grouping of the patients based on exon levels together with the vital information of the patients and the time-to-event information for each cancer type were used for survival modeling by employing the Kaplan–Meier estimate with log-rank test (see "Materials and methods" section). Resulting datasets were stored in a MySQL database as described below and a visualization engine built for dynamic access to the underlying data.

Four tables were designed to store the gene–transcript–exon relationship (Supplementary Fig. 1). "Gene" table has Ensembl gene IDs, their corresponding symbols, and gene synonyms to facilitate searches when users query for alternate gene names in addition to those reported as standard HUGO Gene Nomenclature Committee (HGNC) gene symbols (http://www.genenames.org/). "Transcript" and "Transcript_Exon" tables show the transcript IDs associated with each gene and exons that are associated with each transcript, respectively. Since exons are the building blocks of ExSurv, their genomic coordinates, associated chromosomes, and strand information are organized in the "Exon_Info" table that is connected with the "Transcript_Exon" table. "Cancer_Patient_Info" table stores the clinical information retrieved from TCGA portal such as vital status, number of days from the first day of medication till the last follow-up, and gender for all the patients analyzed in this study for each cancer type. Table 1 shows the composition and the

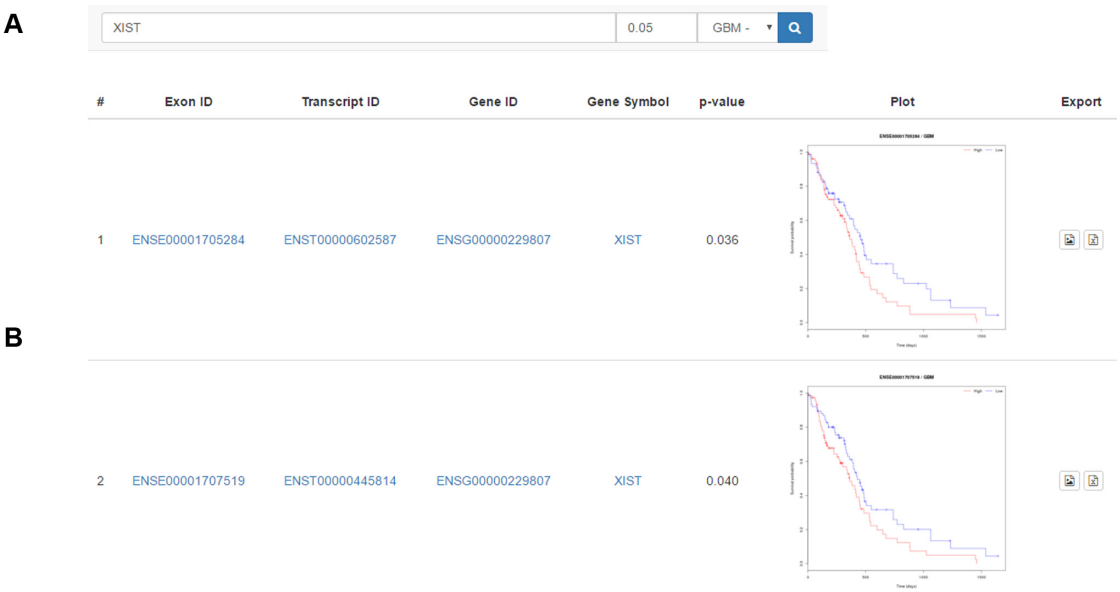

Figure 2. Screenshots from ExSurv user interface. (**A**) The search box with a text box for gene symbol or Ensembl gene ID entry, *P*-value text box that defaults to 0.05 and cancer name select box. (**B**) A sample results table that shows the first two exons obtained. Identifiers are linked to their sources such as Ensembl and GeneCards, and the survival plot can be zoomed in by clicking it. The first button under "Export" column for exporting the survival plot as SVG file and the second one is for exporting a matrix of expression levels of the exons across the patient cohort of a cancer along with several available metadata variables for the patient group as TSV file.

number of patients with various clinical attributes for the four cancer cohorts employed in this study. We stored the precalculated survival probability of each exon across every cancer type in "Cancer_Exon_Survival" table. "Exon_Survival_Data" contains the generated expression data for each exon along with its corresponding expression group, needed for survival analysis and for generating the survival plots (Supplementary Fig. 1 for the complete database schema).

ExSurv provides a rapid means of visualizing the exon-level survival contributions for all exons in a given gene. ExSurv has a user friendly web interface for querying and visualizing the exon-level survival contributions for all exons associated with a given gene symbol or Ensembl gene ID. It has a search form placed next to its main navigation bar to make it accessible from everywhere, and the form includes *P*-value threshold and cancer name selection for narrowing down the query (Fig. 2A). When a search request is received by ExSurv, it queries the database for matching exons to the given gene and the exons that are significant under a given *P*-value threshold and are detected in a given cancer type. If any exon is available for visualization, the survival plots for each of them are prepared using survival package in R and returned to the user as portable network graphics (PNG) formatted images, which are supported by all browsers. Generated plots along with the identifiers such as exon ID, transcript ID, gene ID, and gene symbol, as well as *P*-value, are provided as a resulting table (Fig. 2B). Each plot can be exported as resolution-independent scalable vector graphics (SVG) images. Moreover, the raw data used to generate the visualization of the survival plots can be exported as a table in Tab-Separated Values (TSV) format. These export options are given in the results table under "Export" column.

Most prognostic exons are cancer specific. We extracted the high-confident prognostic exon biomarkers across each of the cancer types studied here at a survival probability <0.01 to study their overlap across cancer types. A comparison of the prognostic exons revealed that while more than 40,000 exons were significant for BRCA, only 4000–7000 significant exons were found for other cancer types. This can be due to a number of reasons including the variability in the sample sizes such as the higher coverage for BRCA compared to the other cancer types and heterogeneity in the cancer samples for all cancer types may not be the same. Nevertheless, a comparison of the number of exons, which were found to be significant across two or more cancer types, clearly revealed a surprising trend (Fig. 3). We calculated the pairwise significance of the overlap among the cancers using hypergeometric test and found that the reported overlap values between all the pairs are indeed significant (P-value < 0.01). We found that only a small fraction of the exons were shared among different cancer types. One might argue that the primary reason for observing cancer-specific prognostic exons is due to the difference in tissue of origin. To further validate this hypothesis, we analyzed the overlap between KIRP and kidney renal clear cell carcinoma (KIRC) to verify that cancers originated from the same anatomical context would also exhibit small overlap; hence, the origin of the dissimilarity between the set of prognostic exons is indeed not associated with tissue of origin rather it is related to cancer-specific mechanism. The analysis suggests that there is indeed small overlap between the two cancer types (hypergeometric,

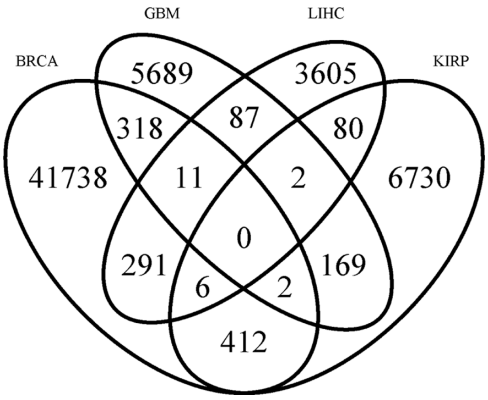

Figure 3. Venn diagram showing the number of prognostic exons in four different cancer types. Significant majority of the exons contributing to patients' survival were found to be cancer specific. Only exons that were found to be significant at a *P*-value threshold of 0.05 from the log-rank test were included in this plot. At these thresholds, no exons were found to be significant for prognosis in all the cancer types studied here.

P-value <0.001) (Supplementary Fig. 2). This observation suggests that most prognostic exons are likely to be cancer specific, indicating a potential for extensive differences in the post-transcriptional alterations seen across cancer types.

Functional analysis of genes associated with survival-associated exons reveals unique as well core functional signatures of cancer types. Performing a functional analysis of the prognostic exons against the current literature is not yet possible as very few studies have attempted to analyze the exon-level contributions to cell growth phenotypes. However, it is possible to analyze the genes associated with the prognostic exons, and hence, we extracted the list of top 500 genes with highest number of prognostic exons, contributing to patients' survival in each cancer type to understand the biological processes associated with such exons. Figure 4 shows the analysis of the enriched biological processes associated with the extracted genes for each cancer type using ClueGO[28] against Gene Ontology database[29] with a *P*-value <0.01. Extracted biological processes suggest that exons

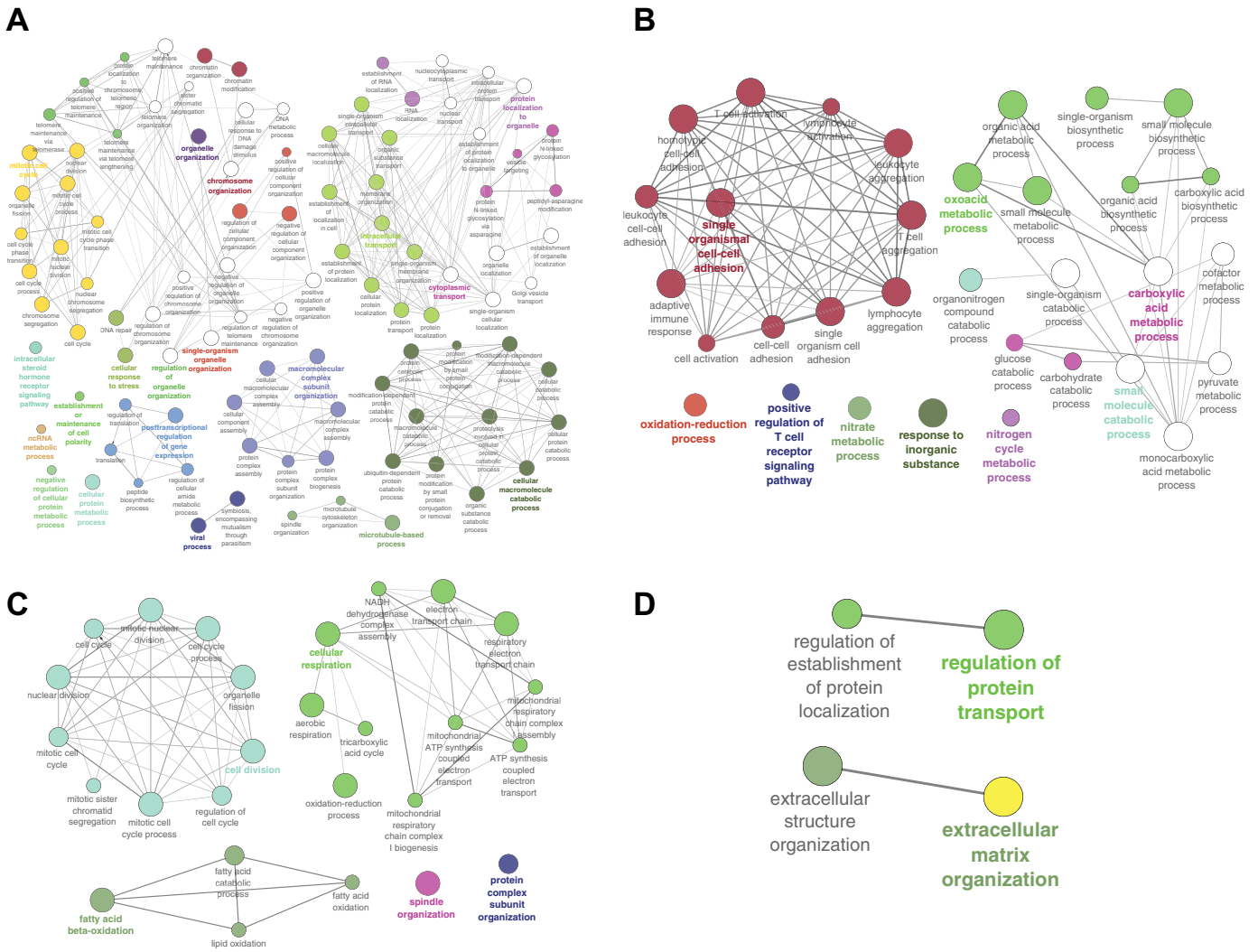

Figure 4. Functional enrichment using ClueGO for genes associated with exons found to be significant in (**A**) BRCA, (**B**) LIHC, (**C**) KIRP, and (**D**) GBM. Biological terms are represented as nodes, and the size of a node is defined based on the number of overlapping genes and the significance of enrichment. Nodes with same color are from the same functional group. An edge between two nodes indicates that genes are associated with multiple biological terms.

contributing to GBM patients' survival are involved in extracellular matrix organization and regulation of protein transport (Fig. 4D). In contrast, prognostic exons in breast cancer patients are primarily active in critical cell functions such as intracellular and cytoplasmic transport, organelle and chromosome organization, mitotic cell cycle, response to stress, protein localization to organelle, and posttranscriptional regulation of gene expression (Fig. 4A and Supplementary Fig. 3). In KIRP patients, survival-associated exons were found to be enriched for cellular respiration, oxidation–reduction process, and fatty acid beta-oxidation, perhaps indicating kidney-specific alterations. Similarly, oxoacid and organic acid metabolic processes together with cell–cell adhesion processes were enriched in liver cancer with prominent contribution of immune response-related pathways such as T-cell activation and aggregation, which are already documented to be implicated in liver disorders (see Fig. 4, Supplementary Fig. 3, and Supplementary Table 1).[30] We also performed functional enrichment analysis of genes whose exons are significantly contributing to patients' survival in at least two cancer types. These results suggest that cell adhesion, carboxylic and fatty acid metabolic process, and positive regulation of T-cell signaling pathway are few major biological processes that were enriched among the significant exons associated with KIRP, LIHC, and BRCA (P-value <0.05; Supplementary Fig. 4).

Conclusions

Cancer is a complex multifactorial disease with our understanding of the posttranscriptional mechanisms contributing to or causal to the cancer phenotypes being very limited. Most approaches currently focus on identifying prognostic markers at the individual gene or transcript level within a cancer type or across cancer types; however, our understanding of the posttranscriptional mechanisms altered in cancer transcriptomes or the resulting splicing biomarkers are limited. Hence, identifying and understanding the function of cancer type-specific prognostic biomarkers based on such poorly characterized layers of regulation is challenging and has recently gained lot of attention in precision medicine field. Alternative splicing is a posttranscriptional mechanism that might change the expression level of the resulting mRNA isoform and hence the final protein product consequently, thereby causing aberrations in the downstream interactome and disease phenotypes. In particular, exon inclusion/exclusion is one of the well-studied alternative splice forms that results in different protein products. While a complete and comprehensive understanding of the functions of exons in the context of their corresponding functional transcripts is still premature, with the availability of novel CRISPR/Cas9 genome editing screens, it might be increasingly possible to study the impact of individual exons on cancer phenotypes in order to rationally design exon-specific therapeutics to decrease the off-target effects and to increase specificity towards more personalized treatment options for cancers. To address this gap and to provide a platform for rational identification of prognostic exons, in this study, we developed ExSurv, a database and web server for exon-level survival significance for four cancer types. ExSurv is a web-based platform where a user can query a gene of interest in a cancer type, resulting in the expression levels of all the exons associated with the gene along with their survival significance and associated plots to be visualized online or for local use.

Our analysis of the resulting prognostic exons across four cancer types clearly revealed that most of the survival-associated exons are unique to a cancer type with few associated processes common across cancer types, possibly suggesting significant differences in the posttranscriptional regulatory pathways contributing to prognosis. ExSurv is a fully functional proof-of-concept platform that will be improved by adding additional cancer types and other functionalities in future versions. Current version of ExSurv performs the survival analysis of exons among the patients of a selected cancer type without employing additional clinical features such as gender and subtype. We will also improve the current platform to facilitate exon survival analysis only among patients who match specific clinical features. Currently, such filtering frequently limits the number of samples, thereby decreasing the power of the analysis for most cancer types. In addition, RNA-seq data for most of the cancer types from TCGA project have been sequenced from multiple sequencing centers or platforms, which could potentially add batch effects to the downstream analysis. However, since such postprocessing and stratification can further reduce the power of the data, careful considerations should be included in the pipelines. This becomes especially important as the number, source, and heterogeneity of the samples increase due to contributions from initiatives such as the ICGC.[31] We anticipate that future versions of ExSurv, which can accommodate such stratifications and controls, could serve to become very powerful for studying the impact of prognostic exons even in subtypes of cancers.

Acknowledgment

The authors wish to thank the members of the Janga lab for helpful discussions in the course of the study.

Author Contributions

Conceived and designed the study: SCJ, SH. Downloaded, preprocessed the datasets, and imported the preprocessed data into the database: SH. Developed the web server: GB. All the authors contributed to the generation of the material needed for writing the manuscript. All the authors reviewed and approved the final version of the manuscript.

Supplementary Material

Supplementary Table 1. List of GO terms significantly associated with single or a pair of cancer types (P-value <0.01). GO associations are extracted from the latest Gene Ontology database using ClueGO and results are ordered by P-value.

Supplementary Figure 1. The database schema of Exsurv: Genomic annotations from ENSEMBL are stored

in Gene, Transcript, Exon_Info, and Transcript_Exon tables. This enables the updating of annotations without changing the whole database structure. Patients' clinical information are stored in Cancer_Patient_Info. Precomputed survival significance values and the exon expression values required for survival plots are stored in Cancer_Exon_Survival and Exon_Survival_Data, respectively.

Supplementary Figure 2. Prognostic exons in kidney renal papillary cell carcinoma (KIRP) and kidney renal clear cell carcinoma (KIRC) exhibit small but significant overlap (P-value <0.001).

Supplementary Figure 3. Pie charts for each cancer type shows the relative significance of GO terms among the associated genes. Bigger pie implies a higher significance GO term related to the cancer genes compared to other terms.

Supplementary Figure 4. Functional enrichment analysis using ClueGO for genes associated with exons found to be significant in at least two cancer types. Carboxylic acid metabolic process and positive regulation of T-cell receptor signaling pathway are two major pathways enriched in the gene sets.

REFERENCES

1. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. *Crit Care*. 2004;8(5):389–94.
2. Gyorffy B, Surowiak P, Budczies J, Lanczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*. 2013;8(12):e82241.
3. Ringner M, Fredlund E, Hakkinen J, Borg A, Staaf J. GOBO: gene expression-based outcome for breast cancer online. *PLoS One*. 2011;6(3):e17911.
4. Gyorffy B, Benke Z, Lanczky A, et al. RecurrenceOnline: an online analysis tool to determine breast cancer recurrence and hormone receptor status using microarray data. *Breast Cancer Res Treat*. 2012;132(3):1025–34.
5. Jezequel P, Campone M, Gouraud W, et al. bc-GenExMiner: an easy-to-use online platform for gene prognostic analyses in breast cancer. *Breast Cancer Res Treat*. 2012;131(3):765–75.
6. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics*. 2009;2:18.
7. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68–77.
8. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med*. 2011;17(3):297–303.
9. The future of cancer genomics. *Nat Med*. 2015;21(2):99.
10. Aguirre-Gamboa R, Gomez-Rueda H, Martinez-Ledesma E, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*. 2013;8(9):e74250.
11. Goswami CP, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinforma*. 2013;3(1):22.
12. Li J, Roebuck P, Grunewald S, Liang H. SurvNet: a web server for identifying network-based biomarkers that most correlate with patient survival data. *Nucleic Acids Res*. 2012;40(Web Server issue):W123–6.
13. Cerami 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(5):401–4.
14. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. 2010;11(5):345–55.
15. Shirai CL, Ley JN, White BS, et al. Mutant U2AF1 expression alters hematopoiesis and pre-mRNA splicing *in vivo*. *Cancer Cell*. 2015;27(5):631–43.
16. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*. 2015;7(1):45.
17. Kechavarzi B, Janga SC. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol*. 2014;15(1):R14.
18. Xia Z, Donehower LA, Cooper TA, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun*. 2014;5:5274.
19. Hollander D, Donyo M, Atias N, et al. A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from ELK1. *Genome Res*. 2016;26(4):541–53.
20. Herrero J, Muffato M, Beal K, et al. Ensembl comparative genomics resources. *Database (Oxford)*. 2016;2016:baw053.
21. *Survival Analysis* [Computer Program]. The Comprehensive R Archive Network; 2015.
22. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60.
23. Cunningham F, Amode MR, Barrell D, et al. Ensembl 2015. *Nucleic Acids Res*. 2015;43(Database issue):D662–9.
24. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5.
25. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
26. Therneau T. *A Package for Survival Analysis in S. Version 2.38*. 2015. Available at http://CRAN.R-project.org/package=survival.
27. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
28. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091–3.
29. Gene Ontology C. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43(Database issue):D1049–56.
30. Zhang F, Xu X, Zhang Y, Zhou B, He Z, Zhai Q. Gene expression profile analysis of type 2 diabetic mouse liver. *PLoS One*. 2013;8(3):e57766.
31. ICGC Data Access Compliance Office, ICGC International Data Access Committee. Analysis of five years of controlled access and data sharing compliance at the International Cancer Genome Consortium. *Nat Genet*. 2016;48(3):224–5.