Computational and Structural Biotechnology Journal 21 (2023) 1978–1988

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Splicing signature database development to delineate cancer pathways using literature mining and transcriptome machine learning

Kyubin Lee [a,b,1], Daejin Hyung [a,1], Soo Young Cho [c,1], Namhee Yu [a], Sewha Hong [a], Jihyun Kim [a,d], Sunshin Kim [a], Ji-Youn Han [a], Charny Park [a,*]

[a] Research Institute, National Cancer Center, 232 Ilsan-ro, Goyang-si, Gyeonggi-do 10408, Republic of Korea
[b] Center for Public Health Genomics, School of Medicine, University of Virginia, Charlottesville, VA 22908, USA
[c] Department of Molecular & Life Science, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea
[d] Department of Precision Medicine, National Institute of Health, Korea Disease Control and Prevention Agency, Osong Health Technology Administration Complex, 187, Osongsaengmyeong 2-ro, Osong-eup, Heungdeok-gu, Cheongju-si, Chungcheongbuk-do 28159, Republic of Korea

ARTICLE INFO

Article history:
Received 11 January 2023
Received in revised form 28 February 2023
Accepted 28 February 2023
Available online 2 March 2023

Keywords:
Text-mining
Machine-learning
Alternative splicing
Tumor transcriptome
Database
Gene signature

ABSTRACT

Alternative splicing (AS) events modulate certain pathways and phenotypic plasticity in cancer. Although previous studies have computationally analyzed splicing events, it is still a challenge to uncover biological functions induced by reliable AS events from tremendous candidates. To provide essential splicing event signatures to assess pathway regulation, we developed a database by collecting two datasets: (i) reported literature and (ii) cancer transcriptome profile. The former includes knowledge-based splicing signatures collected from 63,229 PubMed abstracts using natural language processing, extracted for 202 pathways. The latter is the machine learning-based splicing signatures identified from pan-cancer transcriptome for 16 cancer types and 42 pathways. We established six different learning models to classify pathway activities from splicing profiles as a learning dataset. Top-ranked AS events by learning model feature importance became the signature for each pathway. To validate our learning results, we performed evaluations by (i) performance metrics, (ii) differential AS sets acquired from external datasets, and (iii) our knowledge-based signatures. The area under the receiver operating characteristic values of the learning models did not exhibit any drastic difference. However, random-forest distinctly presented the best performance to compare with the AS sets identified from external datasets and our knowledge-based signatures. Therefore, we used the signatures obtained from the random-forest model. Our database provided the clinical characteristics of the AS signatures, including survival test, molecular subtype, and tumor microenvironment. The regulation by splicing factors was additionally investigated. Our database for developed signatures supported retrieval and visualization system.

1. Introduction

Alternative splicing (AS) expands isoform diversity and is implicated in phenotypic plasticity [1]. Each tissue type revealed distinct differential alternative splicing (DAS); dysregulated AS events promote oncogenic signaling and aggressiveness in several diseases, such as cancer [2]. Cell cycle, epithelial-mesencymal transition (EMT), and apoptosis are well-known biological functions modulated by splicing [1–4]. The advent of high-throughput datasets and the accumulation of reliable knowledge have facilitated the rapid development of AS analysis methods. Despite increasing information, the bona-fide splicing event identification remains a challenge for recapitulating biological process regulation.

Abbreviations: AS, Alternative splicing; DAS, differential alternative splicing; EMT, epithelial mesenchymal transition; ML, machine learning; NER, named entity recognition; NLP, natural language process; PSI, percent spliced in index; PCA, principal component analysis; RF, random-forest; SF, splicing factor; AUCPR, the area under the precision-recall curve; AUROC, the area under the receiver operating characteristic; TCGA, The Cancer Genome Atlas

* Correspondence to: 323 Ilsan-ro, Ilsandonggu, Goyang-si, Gyeonggi-do 10408, Republic of Korea.
E-mail address: charn78@ncc.re.kr (C. Park).
[1] To whom correspondence should be addressed.

https://doi.org/10.1016/j.csbj.2023.02.052

Literature on splicing has been extensively published over the past decades while studying biological functions, but its essential summary profile is insufficient. In the biomedical field, diverse associations like those between immune cells, miRNA and gene, gene and disease, and drug and gene were already well collected from previously published literature using the natural language process (NLP) [5–7]. However, text mining for splicing was performed for restricted information extraction. Previous sentence learning system for eukaryote splicing was used for extracting widespread feature sets, so it was obscure to recognize essential splicing gene candidates [8]. The catalogs for fusion gene junctions were well developed but restricted to only gene IDs extraction for aberrant transcriptome [9,10]. Moreover, the demonstration of current mining results was restricted to computational performance assessment in a simulated or benchmark environment. Existing results of splicing analysis remain insufficient to delineate the functional regulation in the complex human transcriptome. Text-mining is useful technology to summarize dispersed biological associations of splicing genes.

Cancer transcriptome accompanies the heterogeneous molecular subtypes regulated by oncogenic or tumor suppression [11,12]. With respect to splicing, previous studies have mostly interrogated aberrant splicing using single trans-acting elements like splicing factor (SF) mutation [13–15]. However, the aberrant regulation by splicing in cancer cells was found to be implicated in multiple factors, such as cooperative activation of multiple SFs, variants of trans-acting elements, polymorphic or somatic cis-element variants, and epigenetic regulatory changes [2,16–18]. The complex splicing regulation was revealed by identifying AS events involved with biological processes or disease [2,16]. In summary, various splicing events simultaneously regulate certain cancer pathways, and multiple factors of both trans and cis-element variants could disrupt the pathway regulation. The complicated splicing regulation was difficult to determine by single evidence. Well-organized splicing signatures can be helpful to understand novel regulatory mechanisms and pathways of cancer.

Here, we identified the splicing gene signatures based on both knowledge-based contents extracted from the literature and transcriptome-based evidence acquired from the pan-cancer exon usage profile. To obtain knowledge-based splicing signatures associated with pathways, text-mining was performed from published abstracts and identified gene and pathway entities. The association between splicing genes and pathways was inferred from the connection within a sentence. Next, transcriptome-based signatures associated with cancer pathway activity were obtained by machine learning (ML) models using The Cancer Genome Atlas (TCGA) splicing profile. We compared our results from multiple learning methods to extract AS events following the importance of the features. Multiple cross-evaluations demonstrated the predictive performance of the ML models. Additionally, the clinical relevance of splicing was investigated from molecular subtype, survival, and tumor microenvironment. Next, the association of the splicing factor and its binding site was investigated for evaluating the AS signature. Our obtained resource is provided in the form of a database and is available on the web browser system.

2. Material and methods

Splicing signatures to elucidate pathways were collected from two different data sources. The first was obtained from PubMed abstracts (Fig. 1). Entity recognition methods for genes and pathway names were combined with a dictionary and rule-based mining. The gene-pathway associations were ranked for co-occurrence reliability. The second was obtained from the splicing profile of TCGA pan-cancer RNA-seq (Fig. 1) [13]. ML algorithms were used to classify pathway activity, and splicing events were ranked in order of importance of the learned model features. These two different signatures referred to equal pathway terms and gene sets collected from MSigDB [19–21], causing a match between knowledge and transcriptome-based signatures. We used knowledge-based signatures to evaluate transcriptome signatures for each pathway. Additionally, computational, and external dataset comparisons were also performed to verify our ML results, which are described in the subsequent section.

2.1. Text mining to extract splicing genes involved in pathway regulation

To prepare AS genes associated with a specific pathway for evaluation, we collected pathway-related splicing genes using different data sources and text mining. First, we collected 63,229 abstracts published in PubMed related to human alternative splicing from January 1970 to October 2019 (Fig. 2A) using its retrieval query with search keywords 'splicing,' 'splice,' 'spliced,' 'exon skip,' 'exon skipping' or 'skipped exon,' and excluded abstracts containing model organism keywords like 'mouse,' 'Drosophila,' '*C. elegans*,' and 'animal.' We expected that search condition, word 'splicing' could explore wide-range splicing event types, like novel exon skipping in disease or splicing on polyadenylation site. We tried to transform the terminology of complex phrases into an entity recognized as genes or pathways. To recognize the entities, both dictionary and rule-based approaches were generally applied. More specific details were described below. Integrative usage of dictionary and rule-based approaches improves the collection of novel terms from diverse texts [22]. Next, the association between gene and pathway was ranked in the order of reliability measured from the co-occurrence in a single sentence [7]. Therefore, the functional association of splicing genes was acquired from accumulated publications.

Next, to recognize pathway names, we employed both rule-based and dictionary-based approaches. To unify various pathway terminologies to denote one pathway, the rule-based approach using ML technique was applied [22]. For example, the rule-based method identifies one entity 'T-cell receptor pathway' from several phrases like 'T-cell receptor signaling pathway,' 'T-cell receptor signaling,' or 'TCR signaling.' First, the pathway name dictionary (n = 2508) was prepared from MSigDB: curated (C2), ontology (C5), and hallmark (H) gene sets [19–21]. The ontologies of GO terms have hierarchical structures, and upper-layer GO terms cover expansive meanings including lower-layer terms like 'metabolic process,' 'RNA processing,' or 'cell cycle.' Therefore, we used terms from GO level 5 for the dictionary to avoid ambiguous pathways, and additionally filter out pathways with too small (< 20) or large gene set size (> 500). First, simple pathway terms were collected using a dictionary-based approach while referring to the pathway name dictionary [22]. Next, novel phases were recognized using a rule-based NER package, pathNER, to acquire 4046 pathway terms [22]. We manually removed false pathway terms and merged those with duplicate meanings into a single pathway entity. Next, we tested pathway redundancy to confirm the pathway entity phase similarity, using the Jaccard index between two gene sets for each pathway pair (Jaccard index > 0.5) merged into a single pathway. Consequently, we determined the final reference pathway entities (n = 762).

In the final step, we measured gene-pathway associations. The association between gene and pathway entity was determined by co-occurrence of gene-pathway pairs within a single sentence (Fig. 2B). The reliability of gene-pathway co-occurrence was assessed from several measures extensively used in text-mining, namely Pearson correlation, Bayesian probability, and log-likelihood. These three measures were transformed into ranks and merged into one representative rank using the Monte Carlo method of the TopKLists package [26]. Finally, we obtained the ranking of splicing gene-pathway associations.

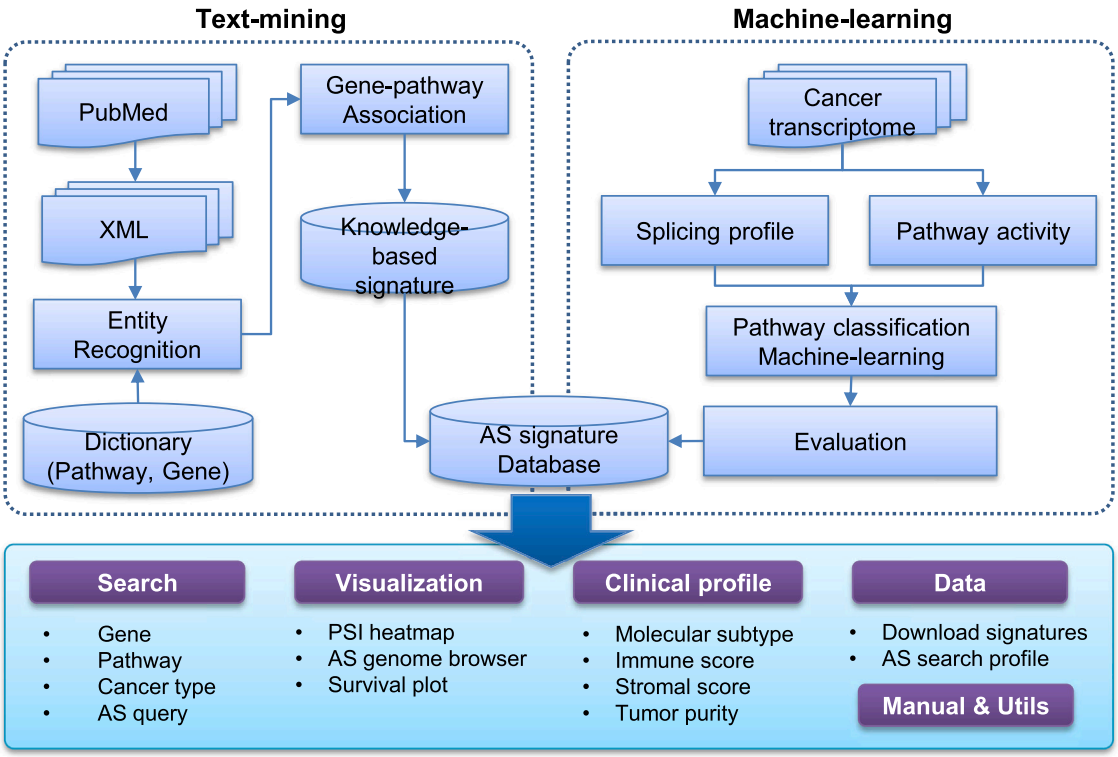

Fig. 1. The workflow of splicing signature development. Knowledge-based signature development process from literature collection to gene-pathway association ranking. Transcriptome-based signature was obtained based on the splicing profile learnings and used for multiple evaluations. Collected database was supported by web server functions, namely, search, visualization, clinical profile summary, and data download.

2.2. Knowledge-based splicing signature's performance evaluation

To evaluate the splicing signature obtained from text-mining, we compared signatures with DAS results derived from external RNA-seq datasets. We collected 11 published results to identify three pathway regulations (EMT, cell cycle, and apoptosis), and the previous studies were validated via biological experiments (Supplementary Table S1). Our signatures for the three pathways were compared with the external dataset DAS results. Chi-squared test was performed for DAS sets and gene sets were randomly chosen (n = 100, 200, 500).

2.3. Establishing a machine learning-based model for obtaining transcriptome-based splicing signature

To identify the AS signature events regulating cancer pathways, we collected TCGA pan-cancer transcriptome and pathway gene sets. Cancer transcriptome was acquired from the TCGA project RNA-seq dataset of 16 cancer types [27]. Splicing profile and percent spliced in index (PSI) matrix were assessed from the TCGA level 3 transcriptome profile using SUPPA2 [28]. AS events were filtered out, revealing low variance of exon usage (PSI standard deviation < 0.1), low gene expression (average RPKM < 1), and dominant frequency (> 95%) of extreme value outliers or missing values in PSI profile. PSI profile was imputed after filtering and normalized for each tumor type. We estimated sample-level pathway activity scores using Gene Set Variation Analysis (GSVA) for MSigDB pathways to classify cancer samples by pathway activities [21,29]. The pathways exhibiting low standard deviations (< 0.2) were excluded with respect to the biological function of non-delineating tumor heterogeneity. Finally, 42 pathways were selected (Supplementary Fig. S1).

ML methods were then used to learn models based on the association between AS events and cancer pathway activity. We trained separate classifiers for each cancer type per pathway where the binary response variable indicated pathway activation, using median GSVA scores as the threshold values. The normalized PSI profiles of the filtered AS events (n = 11,116–14,977 across tumors) were used to train ML models. We investigated six different learning methods representing two broad types of models, which learn non-linear (random-forest (RF), and gradient Boosting Machines (XGBoost)) and linear functions (Naïve Bayes (NB), Lasso, Ridge, and Elastic nets). For each cancer type, training and test samples were randomly divided in a 9:1 ratio. All learning models were trained with five-fold cross-validation, in which the best set of hyperparameters was selected via grid search based on the validation error, as summarized in Supplementary Table S2.

2.4. Transcriptome signature performance evaluation and splicing signature index scoring

The learned ML models were evaluated using three approaches: 1) assessment of predictive accuracy on held-aside test sets of TCGA samples, 2) enrichment test of the knowledge-based signatures and 3) comparison with previous studies using external datasets. The learned models' predictive accuracy was assessed using the area under the receiver operating characteristic (AUROC) and the precision-recall curve (AUCPR), computed with the R PRROC package [30]. Next, we extracted top-ranked AS events from each learned model (n = 100, 300, 500, and 1000) based on the importance of features and tested the degree of enrichment with the knowledge-based signatures. Additionally, the enrichment test was performed against previous publications to validate functionality by splicing regulation from wet-lab experiments. The external datasets for three pathways (EMT, cell cycle, and apoptosis) were utilized (Supplementary Table S1).

RF model was selected based on its predictive power and superior performance than that of other methods in detecting functional splicing events to determine the final AS event signatures. The

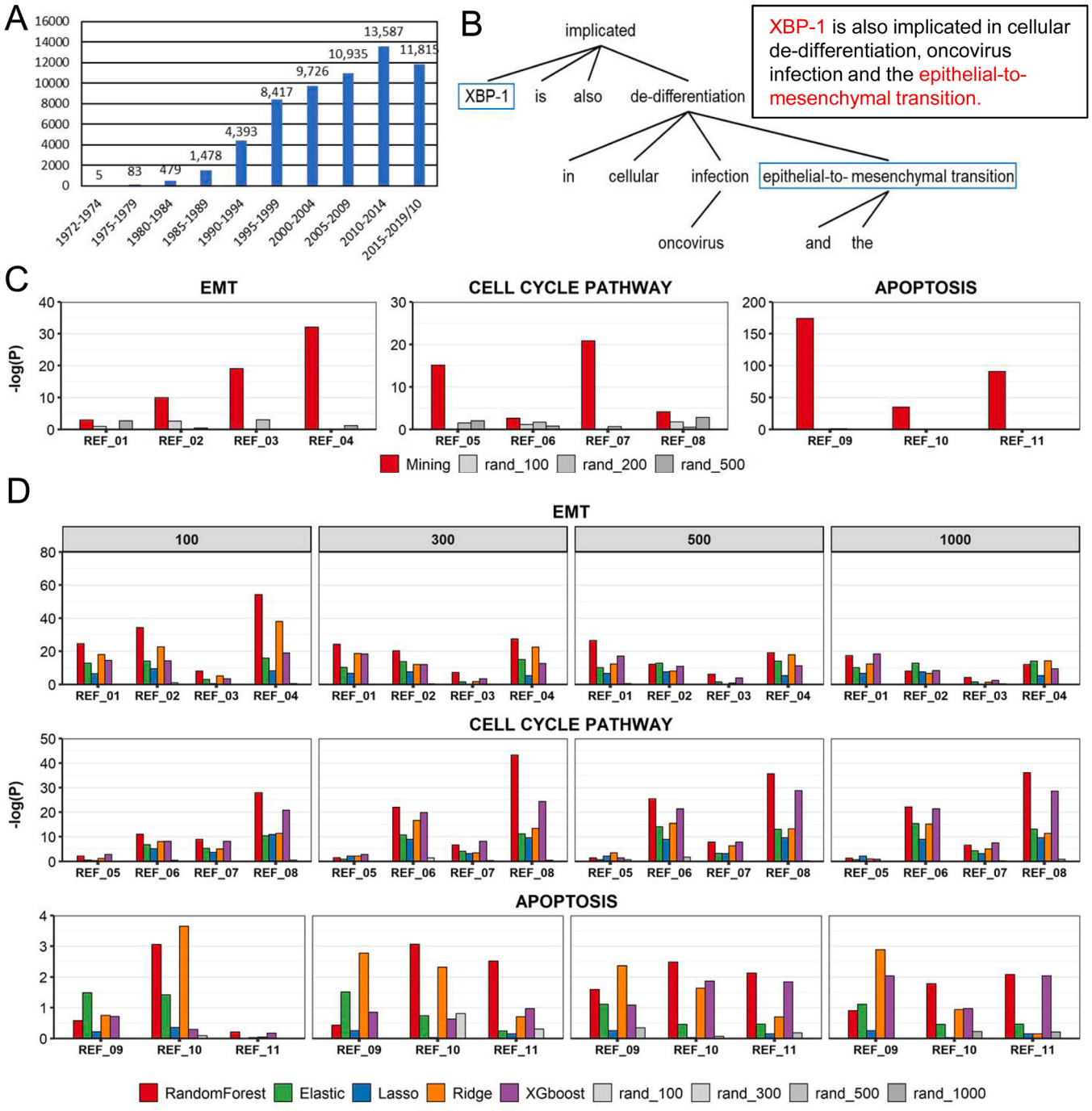

Fig. 2. Splicing signature extraction using text-mining and evaluation of pathways with machine learning approaches. A) The number of literatures published from 1972 to 2019 about text-mining to include splicing keywords. B) An example of entity recognition and co-occurrence investigation from a sentence in text-mining. C) Comparison with previous differential splicing analysis results for three pathways: epithelial-to-mesenchymal transition (EMT), cell cycle, and apoptosis. Splicing events of each study were compared with our obtained mining results and random sets. Pubmed IDs and accession numbers for 11 references are described in Supplementary Table S1. D) Comparison of ML splicing signatures of five methods and random sets with references. Splicing signatures were chosen by feature importance rank (n=100, 300, 500, and 1000).Pre-processing of abstract texts was generally performed using Stanford core NLP parser to sequentially proceed with tokenization, segmentation, part-of-speech tanning, and lemmatization [23]. Named entity recognition (NER) was used to identify gene name entities using BANNER for reference in the gene name dictionary [24], (n = 60,864) which included official symbols, synonyms, and full names collected from the NCBI Gene database [25]. Our aims to interrogate not up-stream splicing regulator, but splicing events. Therefore, we attempted to additionally eliminate SF gene names referring to previously collected RNA-binding protein list (n = 1350) [14], which were referred to in the subsequent evaluation step. In the final step, we manually removed the false positive gene names, such as those with short names (length ≤ 2) or general nouns of low frequency.

top 100 AS events of the highest positive feature importance were obtained from the learned RF model. Next, the AS signatures of compact size were augmented with our knowledge-based splicing gene signatures revealing high recurrence (> 50%) of positive feature importance. Additionally, we re-trained RF models with the same learning procedure but using only the final AS signatures as features and evaluated their predictive accuracy on the previously used test sets. Lastly, we generated a splicing-signature index (splicing-SI) for each cancer type per pathway to determine a representative measure for the identified splicing signatures. Splicing-SI was estimated using principal component analysis (PCA) for the normalized PSI

Table 1
Test set and knowledge-based evaluations for six machine learning models. SD, standard deviation.

Methods	Test set evaluation (value ± SD)			Knowledge-based evaluation (-log P-value)			
	AUROC	AUCPR	Accuracy	Top 100	Top 300	Top 500	Top 1000
RF	0.87 ± 0.08	0.88 ± 0.09	0.78 ± 0.08	1.74	2.30	1.06	1.63
Elastic	0.89 ± 0.07	0.90 ± 0.08	0.80 ± 0.08	1.50	1.09	1.01	1.06
Lasso	0.88 ± 0.08	0.88 ± 0.09	0.80 ± 0.08	0.99	1.28	1.28	1.28
NB	0.71 ± 0.1	0.67 ± 0.12	0.68 ± 0.09	NA	NA	NA	NA
Ridge	0.90 ± 0.06	0.90 ± 0.07	0.81 ± 0.07	1.29	1.60	1.00	0.71
XGBoost	0.90 ± 0.07	0.91 ± 0.08	0.82 ± 0.08	0.71	1.10	1.07	1.02

profiles of the identified AS signatures and quantified as the first principal component values.

2.5. Comparison analysis with cancer clinical data

We assessed the performance of our splicing-SI across cancer types and pathways. Survival analysis was conducted using the R package 'survival'. The Kaplan–Meier model was used to plot the patient's overall survival, and the log-rank test was performed for statistical comparison between two subgroups of patients, using Q1 and Q3 of 1) pathway scores or 2) PSI values for an AS event of interest. Cox proportional hazard regression model was used to estimate hazard ratios. Moreover, the comparison analysis was performed between molecular subtypes of a given cancer type, wherein subtypes of samples with a size smaller than ten were excluded (Supplementary Table S3). Additionally, levels of infiltrating immune and stromal cells, and tumor purity were calculated using the ESTIMATE method [31]. The significance of difference between the groups of interest was measured using the Wilcoxon-rank sum test.

2.6. Association between the activity of splicing factor and splicing signature set

SFs and trans-acting elements directly contributed to step-wise alternative splicing, which consequently modulates specific pathway regulation. We demonstrate that the splicing signatures that we obtained imply SF regulation. To evaluate the regulation, we collected SF genes (n = 165) from reference and classified known pathway-associated SFs (n = 113) from initial text-mining results including all gene-pathway associations before filtering [14]. We categorically excluded known SFs from each pathway just as unknown SF.

We assumed that splicing-SI scores presented the difference by pathway regulation SFs; therefore, we performed the Wilcoxson-rank sum method to determine whether splicing-SI scores exhibited the difference between low (≤50%) and high (> 50%) SF expression groups for all SFs. Additionally, P-values were adjusted using the Bonferroni correction. We assumed that known pathway-associated SFs presented a more significant difference than unknown SFs. To demonstrate this, we evaluated the probability of P-values of known SFs being smaller than those of randomly chosen SFs (iteration = 1000).

Among SF-associated cancer pathways, we picked up the 'T-cell activation' pathway, where splicing-SI exhibited strong performance in molecular and clinical features. CELF2 and SRSF6 were known SFs to regulate the T-cell activation pathway. We focused on binding genes of CELF2, the most highly correlated SF across cancer types, and evaluated whether the best candidate binds to our splicing signature events. CELF2 CLIP-seq in human JSL1 T-cells could be acquired (GSE71264) for cases (CELF2-positive, stimulated) and controls (CELF2-negative, unstimulated) [32]. Its binding sites were annotated using bedtools. Further, we tested whether these distinct binding sites in stimulated cases were enriched to our splicing signatures of T-cell activation using the chi-squared test and odds ratio.

3. Results

3.1. Knowledge-based signature extraction and performance evaluation

To obtain spliced gene-pathway association knowledge, we collected 63,229 PubMed abstracts that included splicing terms (Fig. 2A). As a text-mining result, pathway-associated spliced gene sets were acquired by workflow (Fig. 2B). A total of 442 pathway entities were recognized from literature, of which, we finalized 202 pathways with gene set size of over 10. To evaluate the performance of our knowledge-based signatures, we compared them with published RNA-seq differential AS (DAS) results. Three pathways were found to be frequently published in the literature that included RNA-seq datasets and wet-lab experiments. To demonstrate the performance of our splicing signatures to elucidate pathways, we compared these DAS sets with our mining sets and randomly chosen background gene sets (Fig. 2 C, Supplementary Fig. S2). Expectedly, our mining splicing sets exhibited significant enrichment with published DAS results, in contrast to random gene sets (Fig. 2 C). Additionally, our database provides the knowledge-based splicing sets for each pathway including literature PubMed IDs and ranking by each gene-pathway co-occurrence. The finalized signature dataset is freely downloadable as a text format file on our website.

3.2. Machine learning model establishment and performance evaluation

To select the best performance model, we determined the learning result by (1) the performance metric, (2) the developed knowledge-based signature, and (3) the comparison with external datasets. First, the performance was assessed for six ML methods from test datasets. AUROC values were calculated from each pathway and cancer type; their medians have been summarized in Table 1. All methods achieved high AUROC (> 0.85) values, except NB, which showed the inferior-most performance. XGBoost exhibited the best performance. However, computational measurement was insufficient to decide ML model performance. Particularly, AS candidates extracted for each ML model should be demonstrated via cross-evaluation using external biological datasets. Therefore, we additionally assessed the performance by comparison with the knowledge-based splicing gene set.

Next, we considered the obtained knowledge-based splicing signatures to match with pathways (n = 42) of cancer transcriptome analysis. Next, we compared matched AS signatures of learning models with knowledge-based signatures (Table 1). To compare with known signatures, we selected the events ranked highest by feature importance for each model. Except top 500, RF exhibited the best performance.

Table 2
The performance AUROC values of all AS profiles, ML-ranked top 100, and adjusted final signature for each cancer type. AUCPR has been summarized in Supplementary Table S4.

Cancer type	All (AUROC)	Top 100 (AUROC)	Signature (AUROC)	All vs Top 100 (-log P-value)	All vs Signature (-log P-value)
Pan-cancer	**0.86 ± 0.08**	**0.87 ± 0.08**	**0.87 ± 0.08**	**9.99E-18**	**3.96E-20**
BLCA	0.85 ± 0.09	0.86 ± 0.09	0.86 ± 0.09	7.89E-03	6.50E-03
BRCA	0.9 ± 0.06	0.91 ± 0.06	0.91 ± 0.05	2.39E-05	3.58E-05
CESC	0.79 ± 0.1	0.83 ± 0.08	0.83 ± 0.08	6.85E-06	3.54E-06
COAD	0.89 ± 0.06	0.89 ± 0.07	0.9 ± 0.07	7.44E-02	9.54E-03
ESCA	0.85 ± 0.08	0.83 ± 0.09	0.84 ± 0.08	9.66E-01	8.69E-01
HNSC	0.85 ± 0.06	0.87 ± 0.06	0.87 ± 0.06	8.67E-04	4.34E-04
KIRC	0.85 ± 0.08	0.87 ± 0.08	0.87 ± 0.08	3.45E-03	6.69E-03
KIRP	0.87 ± 0.09	0.87 ± 0.1	0.87 ± 0.1	1.34E-01	2.00E-01
LIHC	0.83 ± 0.08	0.85 ± 0.08	0.85 ± 0.08	2.07E-06	9.61E-07
LUAD	0.86 ± 0.07	0.87 ± 0.06	0.87 ± 0.06	1.56E-03	5.70E-04
LUSC	0.84 ± 0.08	0.85 ± 0.07	0.85 ± 0.07	3.88E-02	1.18E-02
OV	0.85 ± 0.08	0.85 ± 0.08	0.85 ± 0.08	6.70E-01	7.07E-01
PRAD	0.85 ± 0.06	0.89 ± 0.05	0.88 ± 0.05	1.40E-06	2.68E-07
SKCM	0.86 ± 0.06	0.86 ± 0.06	0.86 ± 0.06	4.33E-01	4.78E-01
STAD	0.91 ± 0.05	0.92 ± 0.05	0.92 ± 0.05	3.88E-02	4.25E-02
THCA	0.88 ± 0.07	0.89 ± 0.06	0.89 ± 0.06	7.92E-04	1.66E-03

Third evaluation was performed using an external dataset. We selected three pathway signatures (EMT, cell cycle pathway, and apoptosis) evaluated from external datasets and their wet-lab experiments (Fig. 2). In addition, randomly chosen sets (n = 100, 300, 500, 1000) exhibited the least favorable result (Fig. 2D). In both the pathways, RF exhibited the best performance. Even though apoptosis presented inconsistent performance across all methods, RF exhibited favorable results in 300–1000 sets as compared to other methods.

3.3. Transcriptome signature extraction from pan-cancer splicing profile

To finalize splicing signatures for cancer transcriptome, we selected RF model learning results. AS events were ranked by learning models' feature importance. To construct the essential AS signature, AS events ranked in the top 100 were chosen as the signature for each pathway. Loss of knowledge-based splicing genes in transcriptome's top 100 signatures is inevitable. Therefore, we refined our AS signature to include knowledge-based splicing genes. Therefore, we added the reliable AS events to be (1) observed in knowledge database, (2) recurrent (> 50%) across cancer types, and (3) positive in feature importance. The adjusted AS signature is described in Table 2. AUROC of adjusted signatures was maintained in the top 100, and showed minor increase as compared to original dataset. Performance (AUROC) of adjusted signatures exhibited better (median -log P-value = 3.96E-20) than that of top 100 (median -log P-value = 9.99E-18). Our final cancer transcriptome signatures are available in the database.

3.4. Splicing signature indexes to present the clinical characteristics

Transcriptome profile of each cancer type has classified tumor patients into molecular subtypes to elucidate the distinct biological processes. Moreover, transcriptomic evidence acquired by molecular classification is regarded as prognosis predictors and diagnosis markers [33–35]. Here, we investigated whether our splicing signatures delineate the tumor regulatory mechanism following molecular subtypes. Previously identified subtypes were obtained from TCGA analysis results (Supplementary Table S3) [36]. Each subtype was mostly selected by multi-omics analysis to integrate methylation, mutation, copy number, and gene expression profile. We assessed that our splicing-SIs correlated with tumor pathway scores and molecular subtypes. Additionally, survival was investigated according to splicing-SI.

Among the 42 tumor pathways, we extracted 15 splicing signatures to be highly correlative with pathway scores and to classify molecular subtypes (Fig. 3 A). Splicing signatures partially determined the prognosis across multiple cancer types. We selected stomach cancer (STAD) as a case study from the results. STAD was classified into four subtypes: chromosomal instable (CIN), Epstein-Barr virus (EBV) infection, genomic stable (GS), and microsatellite instability (MSI) [37]. The original study revealed that GS exhibited cell migration activation to be most likely EMT characteristics [37]. In the EMT pathway, our splicing-SI highly correlated with the pathway scores acquired from the gene expression profile (Fig. 3B). Moreover, splicing-SI showed distinctly distinguished GS subtype ($P < 0.001$) than pathway scores ($P < 0.01$), and patients with poor prognosis were predictable from high splicing-SI ($P < 0.001$) like pathway scores. We extracted top-ranked EMT splicing signatures (Fig. 3 C). Among them, top-ranked ML feature importance skipping exon (SE) events in well-known genes modulating EMT: *SEC31A*, *ATP5C1*, and *ARID1B* [38–40]. We selected the *ENAH* SE event as a candidate diagnostic marker (Fig. 3 C). *ENAH* contained 14 exons and exon 11a was skipped in EMT at the EVH2 domain location (Fig. 3D) [41]. EVH2 domain region binds and bundles F-actin is localized to stress fibers due to its importance in cytoskeleton organization and actin-based cell motility [42]. *ENAH* SE was dramatically presented in the GS subtype ($P < 0.001$) and considered as a prognosis marker in overall survival ($P = 0.043$; Fig. 3D).

EMT is a biological process that facilitates cancer cell invasion and metastasis abundantly in fibroblasts and mesenchymal stroma cells. Therefore, we cross-evaluated using an additional metric to dissect tumor microenvironments from bulk gene expression profiles. We utilized immune and stromal cell scores [43]. Stromal development was discovered in the GS subtype, and the stromal cell abundance strongly correlated with the EMT pathway score (Fig. 3E). Finally, our EMT splicing-SI also showed an association with the stromal score. Exon usage of the single marker candidate *ENHA* SE exhibited significant results. In summary, EMT splicing-SI was pivotal in distinguishing EMT and stromal status. In our splicing signature events, a single candidate *ENAH* SE was found to be a diagnostic marker to classify subtypes and predict prognosis.

3.5. Splicing signatures indicate splicing factor regulations and their binding sites demonstrated in T-cell activation splicing signature

SFs involved in selective exon usage of mRNA to consequently proceed AS, and its alteration exhibited tumor-associated splicing machinery [13,44]. We demonstrated that the splicing signatures we

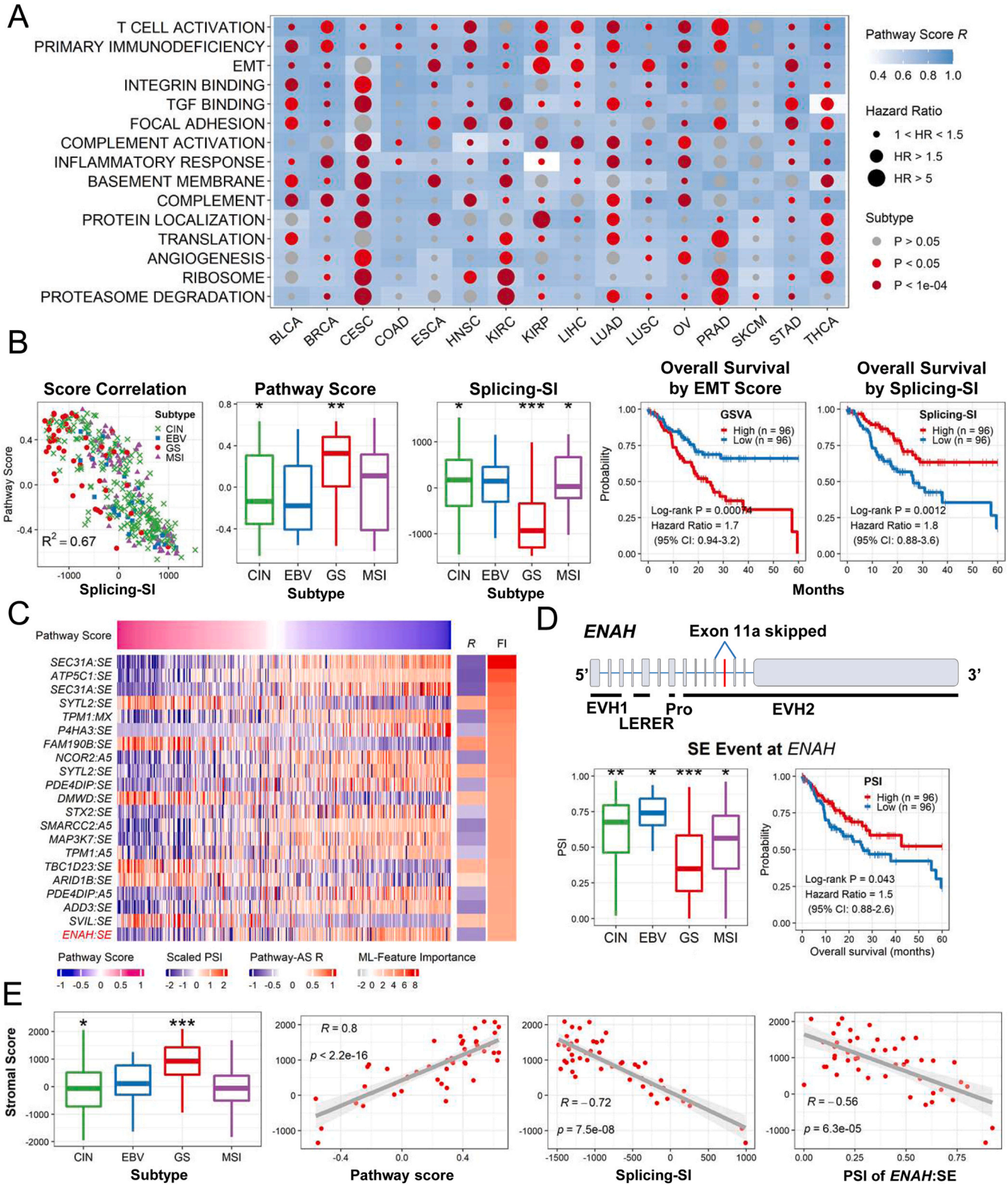

Fig. 3. The clinical relevance of splicing signatures, and a case study of epithelial-to-mesenchymal (EMT) pathway in stomach cancer (STAD). A) Heatmap summarizing the 15 pathway splicing signatures correlated with tumor subtypes and survivals. Other pathways have been summarized in Supplementary Fig. S3. B) EMT pathway score and splicing-signature index (SI) status following four subtypes: chromosomal instable (CIN), Epstein-Barr virus (EBV), genomic stable (GS), and microsatellite instability (MSI). This includes a scatter plot between the EMT pathway score and splicing-signature index (SI), two boxplots of pathway scores and splicing-SI according to four subtypes, and two overall survival plots by pathway score and splicing-SI. C) Splicing event percent-spliced in (PSI) heatmap. Columns were sorted by EMT pathway scores and AS events by machine-learning feature importance. D) Gene structure of ENAH skipping exon (SE), and a PSI boxplot of four subtypes, showing a survival curve of *ENAH* PSI in high and low patient groups. E) Evaluation using stromal scores (y-axis). This includes a boxplot of stromal cell scores in four subtypes, and three scatter plots of stromal scores (Y-axis) with pathway, splicing-SI, and *ENAH* PSI (X-axis). Survival tests were performed using the log-rank test. The Cox regression model was used to acquire hazard ratio (HR) and 90% confidence interval (CI). In boxplots, asterisks indicated P-values (* P < 0.1, ** P < 0.01, *** P < 0.001) by the Wilcoxon-rank sum test. In scatter plots, Pearson correlation coefficients (*R*) and their P-values were specified.

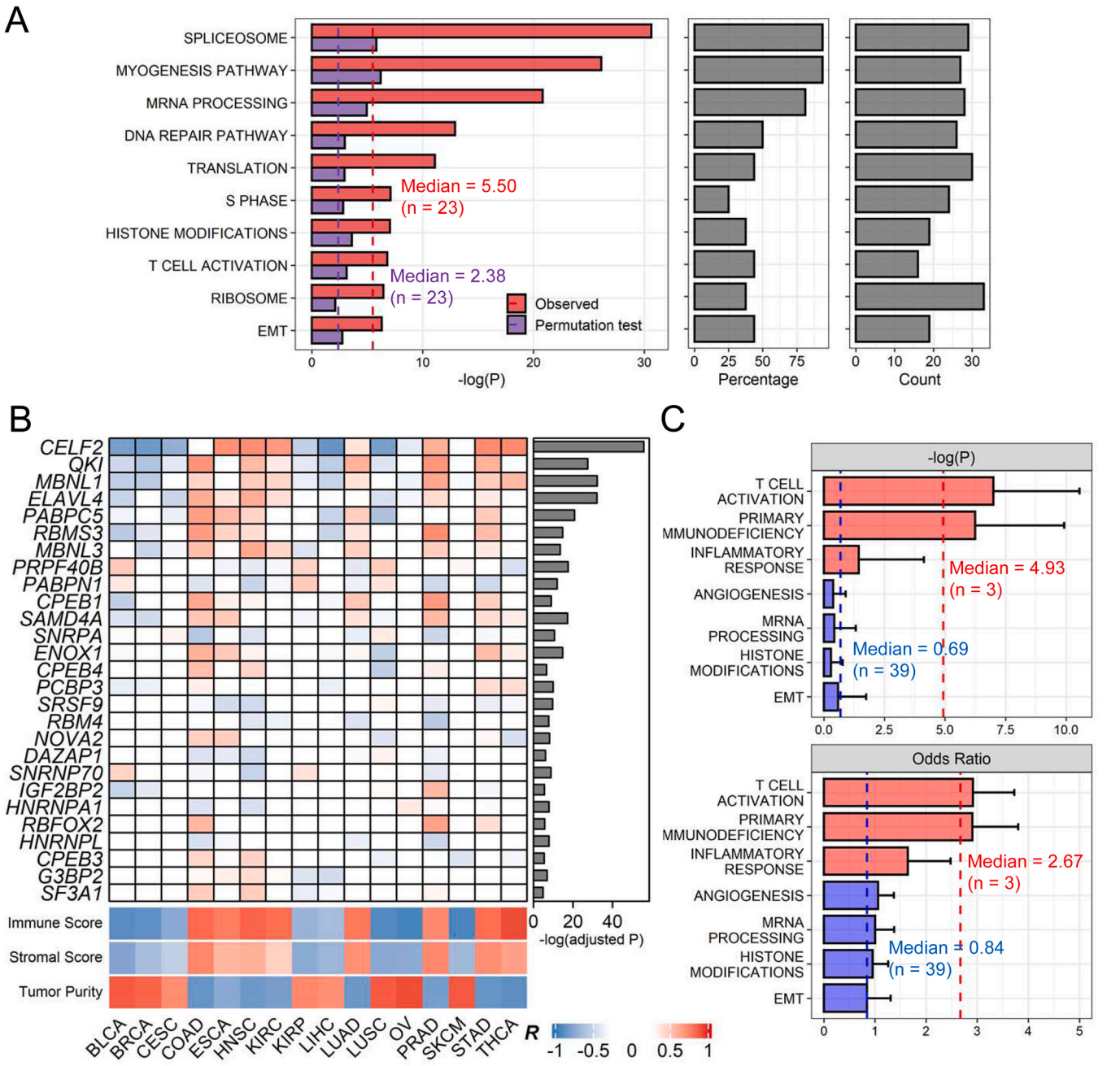

Fig. 4. Activities of splicing factors determine splicing-SI scores. A) Bar plot of log-scale P-values to test the splicing-SI difference for mining SFs (observed), and other SFs (permutation) chosen for permutation test. P-values denote whether mining SFs to regulate certain pathways were significant or not. Observed P-values (red) were obtained by Wilcoxon-test, and permutation test P-values (purple) were performed with iteration 1000. A percentage bar plot shows cancer-type recurrence to pass the test, and the consequent count exhibited detected SFs including novel candidates. Details are summarized in Supplementary Table S5. B) A heatmap of spearman correlation between splicing factor expression and splicing-SI for the T-cell activation pathway. Left panel depicts the median difference in test P-value of splicing-SI between low (≤50%) and high (> 50%) SF groups for each cancer type. Bottom heatmap shows the correlations between immune and stromal scores and tumor purity with T-cell activation pathway splicing-SI score. C) P-value and odds ratio bar plots of enrichment test between CELF2 binding site gene set and the obtained splicing set of T-cell activation pathway. CELF2 binding site was obtained from CLIP-seq to stimulate T-cell activation. Three immune-associated pathways are denoted in red and remaining in blue. Vertical dot line indicates the median P-value and odds ratio for immune-associated pathways and others.

obtained indicate SF regulation. Following previous studies, we identified 165 splicing factor genes, of which 113 were identified as known mining SFs regulating 23 specific pathways acquired from our initial text-mining results (Supplementary Table S5). Our splicing-SI scores indicated that mining SFs regulated pathways were expected to be more significant than other unknown SFs. We determined that the mining SFs maintained their statistical significance as compared to randomly chosen sets via permutation test. Splicing-SI scores of 16 pathways exhibited a significant difference following SF regulation (permutation P-value < 0.1, Fig. 4A). We also tested mining SFs' regulation for cancer types of each pathway. The test pathways that did not pass the permutation test in pathway-level exhibited lower cancer-type recurrence (Supplementary Table S5). Therefore, we speculate that splicing regulation of specific pathways proceeds in a cancer-type specific manner. Meanwhile, our results obtained multiple SFs to participate in a certain pathway from gene expression and text-mining analysis.

Among these, splicing-SI of 'T-cell activation' recurrently determined molecular subtypes across multiple cancer types (n = 14, P < 0.05, Fig. 3A). Therefore, we explored whether our splicing signature events were regulated by specific SF. T-cell activation was chosen for a case study to show clinical relevance. When calculating the correlation between each splicing factor gene expression and splicing-SI, *CELF2* was correlated in multiple cancer types (Fig. 4B) and was known as signal-dependent splicing in T-cells [32]. Immune cell abundance was also highly correlated with *CELF2* expression, while tumor purity exhibited the opposite relation (Fig. 4B).

The stimulation of T-cells was found to be downwardly cascaded with *CELF2*-overexpression [32]. The activation of splicing factor CELF2 facilitated distinct splicing events to participate in T-cell stimulation. Therefore, we investigated the CELF2 binding sites by T-cell stimulation and identified the sites from CLIP-seq under-stimulated and unstimulated T-cell conditions. Unique CELF2-binding genes in T-cell stimulated conditions were extracted after excluding binding sites in unstimulated conditions. Next, we interrogated whether the stimulated condition's binding sites enriched the T-cell activation signature events. When testing for all pathways, three immune response-related pathway signatures (T-cell activation, primary immunodeficiency, and inflammatory response) were found to be enriched on the binding sites (median P < 0.001, median OR = 2.65; Fig. 4B). The remaining pathways exhibited dramatically declined associations (n = 39, median P = 0.84, median OR = 0.84; Supplementary Fig. S4). Although around 100 AS events were extracted for each pathway signature, the events were representative to elucidate SF regulation. Collectively, in the T-cell activation pathway, the obtained splicing signatures were found to accompany the events and bind to specific splicing factors.

3.6. Database browser to retrieve splicing signatures

Knowledge and transcriptome-based splicing signature databases were provided on our database website. The entirety of the contents can be retrieved and downloaded. The use case workflow presented an effective way to understand the splicing profile (Fig. 5). When starting the AS event search from cancer transcriptome signature, users can select cancer type and pathway of interest (Fig. 5A). Top-ranked AS signatures were visualized in a heatmap of PSI profile, and additional search and sorting were provided. When selecting a single splicing event among multiple, its clinical details were presented, namely, overall survival of PSI high and low groups, exon usage difference among TCGA molecular subtypes, PSI distributions according to the tumor microenvironment, immune cell abundance, stromal abundance, and tumor purity. In addition to the cancer profile, pathway-level information was hyperlinked with knowledge-based AS signature for the selected pathway (Fig. 5B). Sequence profile of each selected event was linked with an additional AS sequence profile database ASpedia (Fig. 5C), which is a previously developed database and supports spliced region's multi-omics evidence like variants, RNA-binding protein, protein domain, and protein-protein interaction [41]. Our resources encompassing knowledge-based and transcriptome-based splicing signatures are available in downloadable format on the website.

4. Discussion

Transcriptomic signatures provide advantageous knowledge-based evidence to objectively understand biological processes. Well-defined gene-level signatures have been utilized to estimate the pathway scores, cell decomposition, and so on [45]. The applications potentiated the various transcriptome-based studies across bulk and single-cell analysis. Although AS plays a major role in conferring functional diversity and phenotypic plasticity, splicing evidence is relatively insufficient to compare gene-level databases and depict biological function. The determining scoring approaches and resources for AS are insufficient. Therefore, we developed the AS event signature for each pathway and evaluated it in multiple ways. Particularly, knowledge-based splicing pathway signatures provided a novel clue about the hidden regulation features.

During the evaluation of knowledge-based signatures, our database exhibited remarkable concordance with external studies including RNA-seq analysis and wet-lab experiment. However, knowledge-based signatures providing only gene-level evidence without exon or isoform-level profile is a limitation. We considered the isoform name or exon number recognition, which expressed using irregular terminology or substantial literature was absent in exonic information. For example, the mutually exclusive splicing event of *FGFR2* was described as 'exon IIIb' or '*FGFR2*-IIIb' isoform, and three splicing variants of CaMKII marked with β_M, β, and β_e' [46,47]. The terminology diversity hinders the determination of high-coverage regular expression for entity recognition. In-depth recognition drastically decreased the number of recognized entities and consequently declined the possibility to identify the associations between genes and pathways. To obtain statistical power, gene-level pathway associations are required to be collected. However, the obtained transcriptome-based signatures were limited to support precise splicing event-level signatures. It can compensate for the weak point of the knowledge-based dataset. Further, we expect to generate a higher-resolution splicing profile using text-mining with further research.

None of the ML model methods achieved outstanding performance in AUROC. Therefore, we compared our results with the obtained knowledge-based signature collection and external datasets. Particularly, the external datasets were obtained using DAS analysis from RNA-seq and confirmed in wet-lab experiments for biological function verification. Therefore, the external datasets were found to be the strongest biological evidence for evaluation. Although RF was not computationally outperformed, it exhibited favorable comparison results between external datasets and knowledge-based signatures. RF was chosen to learn the splicing profile, implying that the performance metric is a versatile measurement to demonstrate learning models. Our investigation elucidated that various cross-evaluations of biological aspects are useful in selecting the analysis model.

Tumor molecular subtypes indicate intra-tumor heterogeneity and elucidate biological processes. In the TCGA project, patients were clustered using integrative analysis encompassing a multi-omics profile [36]. The classification also included factors of clinical relevance like diagnosis markers, therapeutic targets, and prognosis. However, previous cancer splicing studies focused mostly on tumorigenesis to compare normal and tumor samples, or uncovered aberrant AS events induced from a single SF [11,48]. The approaches were insufficient to indicate the mechanism by which specific splicing events modulate tumor subtypes and biological functions. Especially, molecular subtypes of the TCGA project were derived by integrative analysis from multi-omics profiles, including methylation, mutation, transcriptome, and proteomics. The splicing candidates approximately manifest the various oncogenic pathways within each tumor type, and the splicing-SI scores successfully classified various molecular subtypes. Meanwhile, pathway-level cases presented the clinical relevance details of AS events in the EMT pathway and of the regulation by splicing factors in T-cell activation. In summary, the obtained splicing signatures provide implicated profile delineating biological processes for cancer transcriptome.

The obtained signature set was supported in the website with the previously developed database ASpedia [41]. The database includes the comprehensive multi-omics contents of the splicing regions, encompassing DNA, RNA, and protein. In the browser, the multi-omics splicing profile was interconnected with our splicing

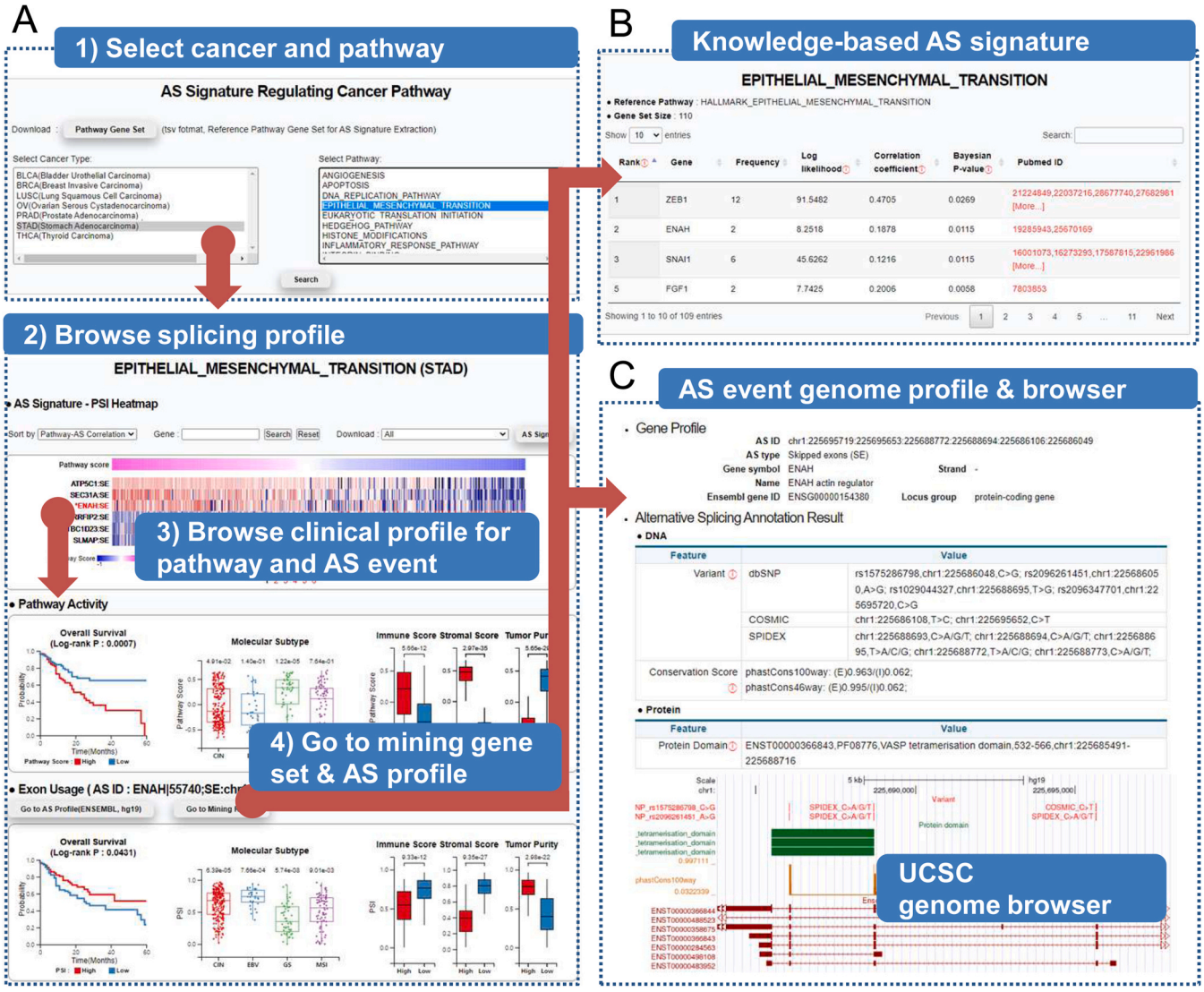

Fig. 5. AS signature search workflow. A) Transcriptome signature retrieval from cancer type and pathway selection of interest (Step1). Splicing browser presents an exon usage heatmap (Step 2) and clinical information for each selected AS event (Step 3). B) Knowledge-based signature was provided for the corresponding pathway. C) Multi-omics sequence profile of AS event can be browsed in table contents and UCSC genome browser (Step 4).

signatures. The integrative database can be useful in determining the functional evidence of splicing events.

In conclusion, both knowledge and transcriptome-based signatures exhibited plausible performance with multiple evaluations. Our splicing signature database is an informative resource to support clinical relevance and reliable biological functions. Moreover, it is applicable to an identified benchmark for biological and computational studies. We believe that our resource can become a solid reference to reveal the biological functions during spliceosome studies.

Funding

This work was supported by the National Research Foundation of Korea grant funded by the Korean government (NRF-2022R1A2C1005708; NRF-2022M3A9I2017587); National Cancer Center Grant (NCC-2210550).

CRediT authorship contribution statement

Kyubin Lee: Formal analysis, Data curation, Visualization, writing. **Daejin Hyung**: Formal analysis, Data curation, Visualization, Software development. **Soo Young Cho**: Software development, writing. **Sewha Hong**: Formal analysis, Investigation, Funding. **Jihyun Kim**: Formal analysis, Investigation. **Sunshin Kim**: Funding. **Ji-Youn Han:** Funding. **Charny Park:** Conceptualization, Project administration, Supervision, Funding, Writing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Korea Institute of Science and Technology Information (KISTI) SuperComputing Center for machine learning model optimization.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.02.052.

References

[1] Dominguez D, Tsai YH, Weatheritt R, Wang Y, Blencowe BJ, Wang Z. An extensive program of periodic alternative splicing linked to cell cycle progression. Elife 2016;5:e10288.

[2] Wang BD, Ceniccola K, Hwang S, Andrawis R, Horvath A, Freedman JA, et al. Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer. Nat Commun 2017;8:15921.

[3] Maslon MM, Heras SR, Bellora N, Eyras E, Cáceres JF. The translational landscape of the splicing factor SRSF1 and its role in mitosis. Elife 2014;3:2028.

[4] Yang Y, Park JW, Bebee TW, Warzecha CC, Guo Y, Shang X, et al. Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. Mol Cell Biol 2016;36:1704–19.

[5] Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. J Biomed Inf 2013;46:200–11.

[6] Kveler K, Starosvetsky E, Ziv-Kenet A, Kalugny Y, Gorelik Y, Shalev-Malul G, et al. Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. Nat Biotechnol 2018;36:651–9.

[7] Xie B, Ding Q, Han H, Wu D. MiRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics 2013;29:638–44.

[8] Shah PK, Bork P. LSAT: learning about alternative transcripts in MEDLINE. Bioinformatics 2006;22:857–65.

[9] Tagore S, Gorohovski A, Jensen LJ, Frenkel-Morgenstern M. rotFus: a comprehensive method characterizing protein-protein interactions of fusion proteins. PLoS Comput Biol 2019:15.

[10] Balamurali D, Gorohovski A, Detroja R, Palande V, Raviv-Shay D, Frenkel-Morgenstern M. ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. Nucleic Acids Res 2020;48:D825–34.

[11] Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res 2017;27:1759–68.

[12] Zhang Y, Qian J, Gu C, Yang Y. Alternative splicing and cancer: a systematic review. Signal Transduct Target Ther 2021:6.

[13] Seiler M, Peng S, Agrawal AA, Palacino J, Teng T, Zhu P, et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep 2018;23(282–296):e4.

[14] Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. 2016;26:732–44.

[15] Carazo F, Romero JP, Rubio A. Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors. Brief Bioinform 2019;20:1358–75.

[16] Lee K, Yu D, Hyung D, Young Cho S, Park C. ASpediaFI: functional interaction analysis of alternative splicing events. Genom Proteom Bioinforma 2022;22:00006–7.

[17] Warzecha CC, Shen S, Xing Y, Carstens RP, Warzecha CC, Shen S, et al. The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. RNA Biol 2009;6:546–62.

[18] Kang HG, Hwangbo H, Kim MJ, Kim S, Lee EJ, Park MJ, et al. Aberrant transcript usage is associated with homologous recombination deficiency and predicts therapeutic response. Cancer Res 2022;82:142–54.

[19] Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The gene ontology resource: enriching a gold mine. Nucleic Acids Res 2021;49:D325–34.

[20] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503.

[21] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database Hallmark gene set collection. Cell Syst 2015;1:417–25.

[22] Wu C, Schwartz JM, Nenadic G. PathNER: a tool for systematic identification of biological pathway mentions in the literature. BMC Syst Biol 2013;7:S2.

[23] Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, p. 55–60.

[24] Leaman R., Graciela Gonzalez. BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing 2008:652–63.

[25] Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res 2015;43:D36–42.

[26] Schimek MG, Budinská E, Kugler KG, Švendová V, Ding J, Lin S. TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. Stat Appl Genet Mol Biol 2015;14:311–6.

[27] Kahles A, Lehmann K, van, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell 2018;34(211–224):e6.

[28] Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol 2018;19:40.

[29] Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinforma 2013;14:7.

[30] Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics 2015;31:2595–7.

[31] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013:4.

[32] Ajith S, Gazzara MR, Cole BS, Shankarling G, Martinez NM, Mallory MJ, et al. Position-dependent activity of CELF2 in the regulation of splicing and implications for signal-responsive regulation in T cells. RNA Biol 2016;13:569–81.

[33] Jiang YZ, Ma D, Suo C, Shi J, Xue M, Hu X, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell 2019;35(428–440):e5.

[34] Sohn BH, Hwang JE, Jang HJ, Lee HS, Oh SC, Shim JJ, et al. Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project. Clin Cancer Res 2017;23:4441–9.

[35] Walter V, Yin X, Wilkerson MD, Cabanski CR, Zhao N, Du Y, et al. Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. PLoS One 2013:8.

[36] Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 2018;173:291.

[37] Cancer T, Atlas G. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9.

[38] Xu Y, Gao XD, Lee JH, Huang H, Tan H, Ahn J, et al. Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. Genes Dev 2014;28:1191.

[39] Li J, Choi PS, Chaffer CL, Labella K, Hwang JH, Giacomelli AO, et al. An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer. Elife 2018:7.

[40] Selvanathan SP, Graham GT, Grego AR, Baker TM, Hogg JR, Simpson M, et al. EWS–FLI1 modulated alternative splicing of ARID1A reveals novel oncogenic function through the BAF complex. Nucleic Acids Res 2019;47:9619.

[41] Hyung D, Kim J, Cho SY, Park C. ASpedia: a comprehensive encyclopedia of human alternative splicing. Nucleic Acids Res 2018;46:58–63.

[42] Gertler FB, Niebuhr K, Reinhard M, Wehland J, Soriano P. Mena, a relative of VASP and drosophila enabled, is implicated in the control of microfilament dynamics. Cell 1996;87:227–39.

[43] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013;4:2612.

[44] Anczuków O, Krainer AR. Splicing-factor alterations in cancers. RNA 2016;22:1285.

[45] Maghsoudi Z, Nguyen H, Tavakkoli A, Nguyen T. A comprehensive survey of the approaches for pathway analysis using multi-omics data integration. Brief Bioinform 2022.

[46] Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Carstens RP. ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. Mol Cell 2009;33:591–601.

[47] Bayer KU, Koninck P, de, Schulman H. Alternative splicing modulates the frequency-dependent response of CaMKII to Ca2+ oscillations. EMBO J 2002;21:3590.

[48] Jha A, Quesnel-Vallières M, Wang D, Thomas-Tikhonenko A, Lynch KW, Barash Y. Identifying common transcriptome signatures of cancer by interpreting deep learning models. Genome Biol 2022:23.