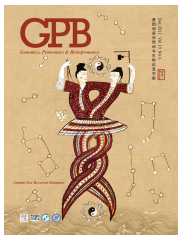
Genomics Proteomics Bioinformatics 19 (2021) 901–912

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



ORIGINAL RESEARCH

Differential Splicing of Skipped Exons Predicts Drug Response in Cancer Cell Lines

Edward Simpson[1,2,3], **Steven Chen**[1,3], **Jill L. Reiter**[1,3], **Yunlong Liu**[1,2,3,*]

[1]*Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[2]*Department of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA*
[3]*Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Received 20 May 2019; revised 19 August 2020; accepted 7 February 2021
Available online 2 March 2021

Handled by Yi Xing

Abstract Alternative splicing of pre-mRNA transcripts is an important regulatory mechanism that increases the diversity of gene products in eukaryotes. Various studies have linked specific transcript isoforms to altered drug response in cancer; however, few algorithms have incorporated splicing information into drug response prediction. In this study, we evaluated whether basal-level splicing information could be used to predict drug sensitivity by constructing doxorubicin-sensitivity classification models with splicing and expression data. We detailed splicing differences between sensitive and resistant cell lines by implementing quasi-binomial generalized linear modeling (QBGLM) and found altered inclusion of 277 skipped exons. We additionally conducted RNA-binding protein (RBP) binding motif enrichment and differential expression analysis to characterize *cis*- and *trans*-acting elements that potentially influence doxorubicin response-mediating splicing alterations. Our results showed that a classification model built with skipped exon data exhibited strong predictive power. We discovered an association between differentially spliced events and epithelial-mesenchymal transition (EMT) and observed motif enrichment, as well as differential expression of *RBFOX* and *ELAVL* RBP family members. Our work demonstrates the potential of incorporating splicing data into drug response algorithms and the utility of a QBGLM approach for fast, scalable identification of relevant splicing differences between large groups of samples.

KEYWORDS Alternative splicing; Transcript isoform; Splicing regulation; Drug sensitivity; Precision medicine

Introduction

The splicing of pre-mRNA transcripts is an important regulatory control mechanism that significantly increases the diversity of protein isoforms in a cell [1]. Alternative splicing plays a major role in the differentiation and maintenance of cellular identity, and as much as 95% of multiexon genes may be alternatively spliced [2,3]. Alternative splicing is also known to contribute to cancer development and progression, and has been linked to every major signature of cancer transformation [4]. Furthermore, splicing variants can help cells evade cancer therapies, and investigators have already started to explore splicing-focused therapeutic options [5–10]. Additionally, certain gene isoforms have been found to alter cancer drug response through altered kinase signaling [11,12]. Therefore, it is likely that alternatively-spliced isoforms play large roles in drug response, and that additional research in this area could

*Corresponding author.
E-mail: yunliu@iu.edu (Liu Y).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2021.12.002

have a major impact on the development of targeted therapeutics and drug response modeling.

Precision medicine, or tailoring treatment strategies to the patient, is dependent on clinical and molecular profiling [13]. Currently, precision medicine primarily relies on limited genetic screening of well-characterized high-impact genes, such as *HER2* and *KRAS* [14]. However, complex predictive models built with machine learning techniques are expected to revolutionize precision medicine in the years to come [15,16]. Nevertheless, the use of complex predictive algorithms has yet to be widely accepted in clinical settings [15]. While early models lacked sufficient study sizes or could not be validated, a major concern of current models is the failure to account for the complexity of tumor transcriptomes [17]. Many predictive models have been trained solely on gene expression data or a combination of expression data and limited sequence variant information, such as single nucleotide polymorphisms (SNPs), copy number variants (CNVs), and small nucleotide insertions or deletions (indels) [18]. Previous studies, however, have concluded that algorithms capable of integrating knowledge from various experimental techniques need to be developed in order for predictive modeling to progress [18–20]. As such, a variety of experimental data, including mRNA-splicing data, must be considered in order to build more realistic and comprehensive models.

Although long-read isoform sequencing technologies exist, they are often prohibitively expensive for large-scale studies. Consequently, short-read data are commonly used to infer isoform-specific information; the drawback being that the true identities of mRNA isoforms remain unknown. This uncertainty must be accounted for in quantitative techniques [21]. There are two main approaches to quantify isoform outcomes in short-read RNA-sequencing data: isoform- and exon-centric quantification [22]. Isoform-centric techniques measure the expression of whole isoforms by integrating read data across multiple exons, whereas exon-centric techniques measure the relative expression of individual exons. While both isoform- and exon-centric techniques are susceptible to short-read sequencing limitations, gene complexity and the heavy reliance on mathematical modeling to address combinatorial possibilities across exons often make isoform-centric approaches less attractive [23].

To date, few studies have incorporated splicing information into predictive modeling techniques. One such study produced the SURVIV pipeline, a system for discovering mRNA isoforms associated with patient survival [24]. These authors used exon-centric quantification and a binomial generalized linear model (GLM) with length normalization function on invasive ductal carcinoma data. They found that splicing information not only predicted patient survival but it also consistently outperformed expression-based models. Additionally, the authors found that combining clinical, expression, and splicing profiles produced the best performance. In another study, isoform-centric biomarker expression and drug response in cancer cell lines were investigated using a linear model to select an isoform for each response-mediating gene that showed the strongest correlation with drug sensitivity [25]. A small number of these biomarkers were validated in breast cancer cell lines and significantly associated with four anti-cancer therapeutics. Together, these two studies established a connection between mRNA splicing and drug response, demonstrating the potential utility of splicing data in tumor biology. However, a drug response classification model has not yet been established, and the relationship between individual exons and cancer drug response is still largely unexplored.

Therefore, to limit the noise introduced from short-read data and avoid the use of complicated probabilistic models, we proposed an exon-centric approach to investigate the relationship between alternative splicing data and cancer drug response. Our study was defined by three primary goals: 1) establish if splicing data predict drug response to a specific anti-cancer drug; 2) evaluate the pretreatment differences in splicing between cell lines that are sensitive or resistant to the drug; and 3) identify *cis*-acting elements that help explain the observed splicing differences. To address these challenges, we merged RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE) with drug response data from the Cancer Therapeutic Response Portal (CTRP) [26,27]. We first applied a machine learning-based approach to determine whether basal splicing profiles predicted doxorubicin sensitivity. Then, we systematically evaluated the pretreatment differences in splicing patterns using quasi-binomial generalized linear modeling (QBGLM), which allowed us to account for the uncertainty in splicing quantification and minimize the computational resources required to perform splicing analysis. Additionally, taking an exon-centric approach for quantification allowed the use of sequence information around the differentially spliced exons to identify enrichment of *cis*-acting motifs and their corresponding RNA-binding proteins (RBPs), thereby providing insight into the regulation of differentially spliced exons.

Results

Dataset, drug, and model selection

We integrated RNA-seq data for 975 cell lines from CCLE with drug response data for 860 cell lines from CTRP [26,27]. After intersecting cell lines in CCLE and CTRP, we observed the number of cell lines with both data types differed by drug. Per-drug area under the concentration-response curve (AUC) values from the CTRP were plotted (Figure S1). A higher AUC value, a surrogate for cell growth under increasing concentrations of a designated

drug, corresponds to superior drug resistance. We chose doxorubicin to investigate further because it is a widely active chemotherapeutic used to treat a variety of malignancies, and it affects cells through multiple mechanisms, including DNA damage by intercalation and inhibition of topoisomerase II [28,29]. Additionally, we reasoned that doxorubicin would be a right drug for proof-of-principal testing because the alternatively spliced exons we identified would likely be relevant to a variety of cancer types. In contrast, spliced exons associated with targeted therapeutics might be relevant only to cancers containing specific genomic alterations. Furthermore, doxorubicin has been used in many drug modeling studies, and therefore, our results would be expected to have greater context and build upon an existing body of knowledge.

Following drug selection, we labeled cell lines according to their AUC values: cell lines at or below the 33rd percentile of the AUC distribution were considered doxorubicin sensitive and cell lines at or above the 66th percentile as resistant (**Figure 1**). This provided a total of 755 cell lines with intersected RNA-seq and doxorubicin response data; 253 were classified as sensitive and 258 as resistant.

Splicing and expression data individually predict drug sensitivity class

We postulated that alternative splicing profiles from untreatedcancer cell lines would hold predictive power for doxorubicin drug response. Hence, we built a machine learning model with elastic net logistic regression and exon-centric splicing data. Skipped exon event annotation, percent-spliced-in (PSI) calculation, and uncertainty estimation were done with the Mixture of Isoforms (MISO) software package [21]. For the splicing-based model, we required skipped exon events (model features) to be present in a minimum of 35% of cell lines and to exhibit PSI values with confidence intervals (CIs) between 0.01 and 0.2. We observed that PSI values with CIs outside of this range tended to be either calculated on low read counts or exhibited unrealistically precise distributions; these PSI values were filtered because small non-consequential changes in PSI would have been incorrectly considered highly significant. Skipped exon events were then limited to only those with the highest (top 5%) PSI standard deviation, thereby targeting events with higher variance and selecting for greater model impact. From a total of 40,178 pre-filtered skipped exon events, 805 remained. Cell line data were then randomly split into 7:3 (training set, n = 354; testing set, n = 157); each set consisted of approximately 50% sensitive and resistant cell lines. The predictive model was fit using elastic net logistic regression. The final splicing model contained a total of 42 non-zero weight events (Table S1). Model performance was assessed on the testing data, and performance metrics are provided in **Table 1**.

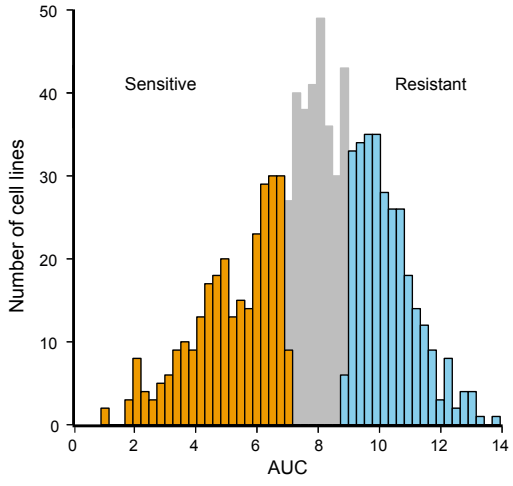


Figure 1 **CTRP cell line response to doxorubicin**
Distribution of the AUC values for doxorubicin in the CTRP cell lines. Lower and upper tertiles were labeled as sensitive (orange) or resistant (blue), respectively. AUC, area under the concentration-response curve; CTRP, Cancer Therapeutic Response Portal.

To assess whether splicing information would provide additional predictive power compared to an expression-based approach, we constructed an expression-only model. We first used featureCounts to quantify reads mapped to gene expression features [30]. To reduce the number of sparse genes, we filtered gene features with less than 10 reads in ≥ 35% of RNA-seq data. Using the same training set as the splicing-based model, we conducted differential expression analysis with edgeR to reduce the number of features [31]. We retained genes with Benjamini-Hochberg false discovery rate (FDR) < 0.05 and $\log_2$ fold change > 1.74 (top 5%) [32]. Read counts were then transformed to $\log_{10}$ counts per million. Out of 57,905 pre-filtered gene expression features, only 1103 remained. After running the elastic net, we obtained an expression-only model comprised of 67 non-zero weight features (Table S2). The performance of the expression-based approach was also strong (Table 1). In comparison with the splicing-based model, the sensitivity was lower (0.68 *vs*. 0.75), but the specificity

Table 1 **Performance metrics**

Model	Sensitivity	Specificity	Accuracy	Precision	AUROC	F1 score	*P* value
Splicing	**0.75**	0.88	0.82	0.85	0.85	0.80	5.3E−15
Expression	0.68	**0.96**	**0.83**	**0.95**	0.90	0.79	2.8E−16
Expression + splicing	0.71	0.95	**0.83**	0.93	**0.92**	**0.81**	**6.2E−17**

Note: The best value for each column is put in bold. AUROC, area under the receiver operating characteristic curve.

(0.96 *vs*. 0.88) and area under the receiver operating characteristic curve (AUROC; 0.90 *vs*. 0.85) were both higher. These metrics indicated that while splicing predicted more doxorubicin-sensitive cell lines correctly, it also predicted more false positives; on the other hand, expression-only modeling was more specific.

An integrated modeling approach outperforms standalone models

Based on our findings that splicing- and expression-based models showed strengths in sensitivity and specificity, respectively, we asked whether integrating the information from both models would lead to increased model performance. An integrated model was fit by merging the 805 events obtained after applying the splicing filter with the 1103 gene expression features remaining after applying the differential expression filter. From this combined feature set, elastic net selected 95 splicing and 216 gene expression features (Table S3). Receiver operating characteristic curve (ROC) plots for all three models are shown in **Figure 2**. The integrated model showed the highest accuracy and AUROC (Table 1). From this outcome, we concluded that splicing information enhanced the expression-based model and that splicing and expression data contributed improvements to sensitivity and specificity, respectively, to build a more balanced model. Bootstrapping the model building process revealed that although the combined model consistently showed a slight increase in specificity, the overall performance of the combined and expression-based models was largely the same (Figure S2).

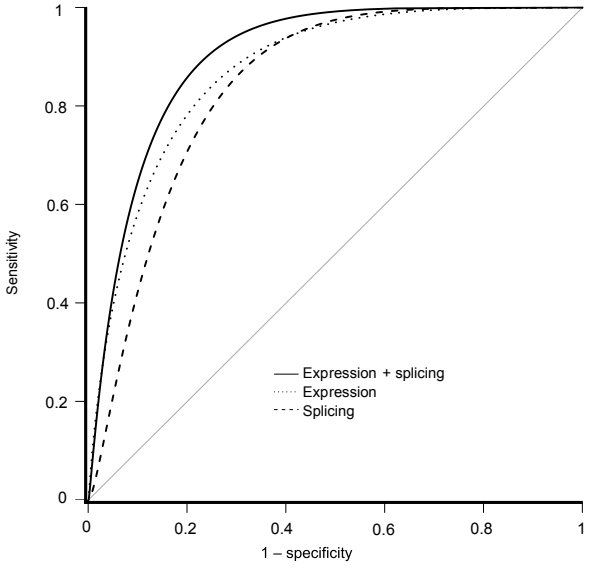


Figure 2 Comparison of model prediction of cell line response to doxorubicin
ROC plots for predicting cell line response to doxorubicin on the testing data set. Expression-only, splicing-only, and combined expression and splicing models are shown. ROC, receiver operating characteristic curve.

Finally, we asked whether splicing data contribute unique information to the final model or if the skipped exons selected by elastic net are also reflected by the gene expression features. We found that skipped exon features in the splicing-only model were not located in genes in the expression-only model. Similarly, no overlapping expression and skipped exon features were observed in the combined model. These findings indicate that the information contributed by splicing data to our models was unique.

QBGLM identifies differentially spliced events

Next, we evaluated the alterations in splicing between sensitive and resistant groups that contribute to drug response differences. For this part of our study, we performed differential splicing analysis. During modeling, feature selection was not considered because machine learning algorithms tend to choose a limited number of features with the strongest predictive value, somewhat arbitrarily, rather than capture the full set of biologically relevant features. Genome-wide annotation for skipped exons resulted in a total of 38,108 events that were then filtered to retain only those with reads supporting inclusion and exclusion for a minimum of 35% of cell lines in the sensitive and resistant groups. This filter significantly decreased the event space and left 18,409 events for analysis. QBGLM was then performed in R [33]. In the quasi-binomial distribution, the dispersion parameter provides for the fitting of increased variance; this property is especially useful for biological data, where the variability between samples is expected. Additionally, fitting variance by QBGLM helped account for the uncertainty introduced when using short-read data in splicing analysis and situations where a low number of reads inaccurately represent the probability of inclusion in some samples. Our procedure was also unique for splicing data normalization in that no consideration was made for exon or read length. As such, QBGLM modeled uncertainty without assuming an equal probability of reads aligning to every position in the event. Wald *P* values, corresponding to the weight on the class of the cell line, were FDR-adjusted using the Benjamini-Hochberg procedure and filtered for significance less than 0.01 [32]. Events were again filtered after QBGLM by requiring a difference in mean inclusion-to-total read counts of 0.1 between sensitive and resistant groups. This filter reduced false-positive identifications by selecting events that were more likely to exert meaningful biological consequences. In total, 277 significant alternatively spliced events were identified: 180 with higher (Table S4) and 97 with lower (Table S5) exon frequency in resistant cells. A volcano plot of the results and examples of raw data for two significant events are presented in **Figure 3**.

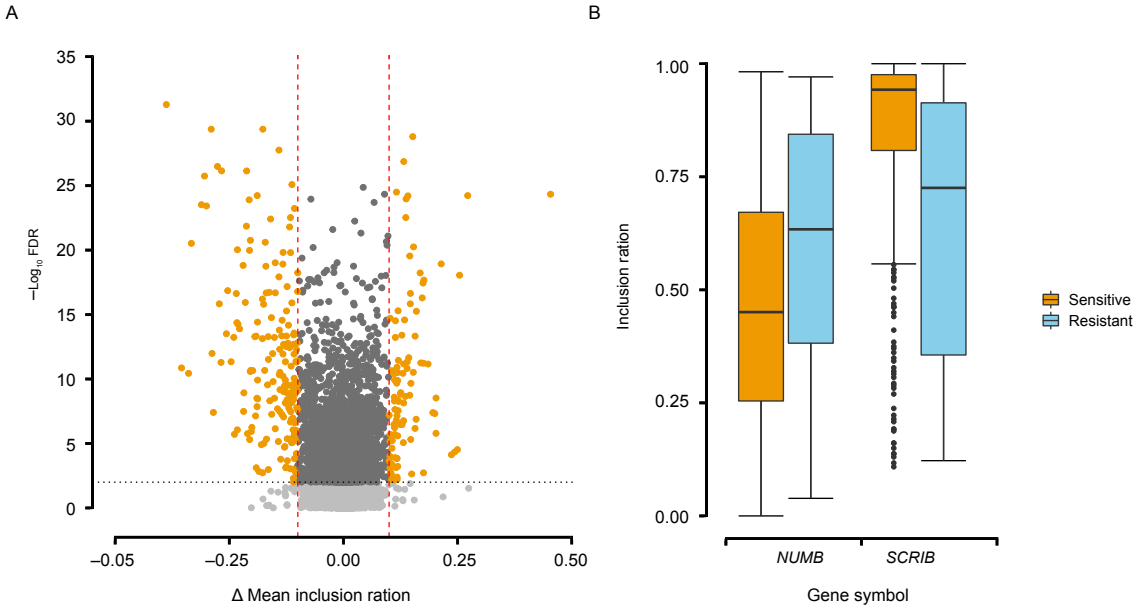


Figure 3 **Differentially spliced events analyzed by QBGLM**
A. Volcano plot of events analyzed by QBGLM. The horizontal dotted line marks the FDR cutoff of 0.01 for significance. Vertical dashed lines separately denote −0.1 and 0.1 difference (Δ) in mean inclusion-to-total read counts. **B.** Boxplot for inclusion ratios showing the overall change between sensitive and resistant groups for two genes with significant spliced events. The box denotes the first-to-third quartile, and the inner-line represents the mean. Whiskers extend to 1.5× the interquartile range, and outliers are marked as points. FDR, false positive discovery; QBGLM, quasi-binomial generalized linear modeling.

Over-representation analysis reveals enrichment for epithelial-mesenchymal transition

Gene ontology (GO) term enrichment was performed on gene symbols from significant alternatively spliced events to assess their relevance. Over-representation analysis with clusterProfiler revealed significant enrichment (FDR ≤ 0.05) for several biological processes including cell junction organization, regulation of cytoskeleton organization, and positive regulation of GTPase activity (**Figure 4**) [34,35]. Alterations in these processes have been previously implicated in uncontrolled cellular proliferation, epithelial-mesenchymal transition (EMT), and drug resistance [36–40]. Noteworthy, genes affected by splicing alterations included *SCRIB*, *ADAM15*, *MACF1*, *NUMB*, *VEGFA*, and *FOXM1*. While the majority of splicing consequences were in-frame alternatively included or excluded exons with unknown significance, an exon identified in *NUMB* (exon 11, chr14:73,745,989−73,746,132) contained an alternative translational start site, and another in *SCRIB* (exon 16, chr8:144,889,722−144,889,784) included a portion of a PKC phosphorylation site. *NUMB* is a key protein in cell fate determination, and increased expression has been found to inhibit the propagation of chronic myelogenous leukemia cells [41,42]. Additionally, *NUMB* mRNA processing is regulated by a variety of splicing factors, including RBM6, and alternative *NUMB* isoforms are consistently found in cancer [43,44]. *SCRIB* exon 16 has been reported to be associated with misregulation of EMT in specific cell types [45].

RBP binding motif enrichment and regulatory splicing factors

To elucidate a regulatory mechanism for the splicing differences between sensitive and resistant groups, we searched for RBP binding motifs corresponding to potential splicing factors. Motif analysis was conducted on seven sequence regions for each skipped exon. These regions consisted of the entire skipped exon sequence, the 300-bp sequences from the 5′ and 3′ ends of both flanking introns, the 150-bp sequence from the 3′ end of the upstream exon, and the 150-bp sequence from the 5′ ends of the downstream exon (**Figure 5**A). Sequences from these regions were extracted from the hg19 reference genome and scanned for motifs using FIMO [46]. All annotated skipped exons across the genome were scanned, and null distributions of counts for each motif were made from bootstrapped events to determine enrichment for identified motifs. RBP binding motifs for seven RBPs (SNRPA, PPRC1, RBM6, PCBP3, RBFOX1, EIF2S1, and ELAVL1) were identified. Locations and enrichment *P*values of the identified RBP binding motifs are shown in Figure 5A and **Table 2**. Fisher's exact test was used to determine association with higher or lower exon frequency. Splicing outcome, Fisher's *P* values, and descriptions of the identified RBPs are shown in Table 2.

Finally, we asked whether any of these enriched RBPs are differentially expressed between sensitive and resistant cell lines. Differential expression analysis was conducted

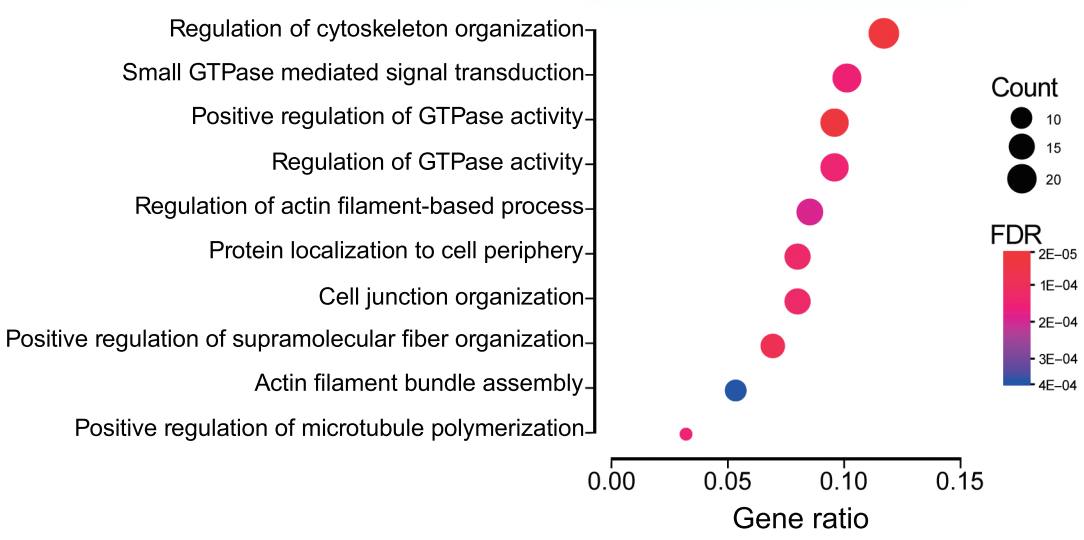


Figure 4 Enrichment of biological processes identified in differentially spliced events
Biological processes identified with over-representation analysis were sorted by gene count ratio from top to bottom, with the highest ratio of found genes for a specific process on top. Point diameters are scaled by the total number of genes in that process, and warmer colors indicate significance.

using edgeR on featureCount data from the two groups (Table S6) [31,53]. Significantly differential expression patterns were only observed for RBFOX and ELAVL family proteins (Figure 5B). This finding was particularly interesting as RBFOX and ELAVL family members have previously been linked to EMT and other cancer-related processes (Table 2) [45,54,55]. Notably, Blencowe et al. [56] previously found that PPRC1 increased splicing of *NUMB* exon 5 in CGR8 mouse embryonic stem cells compared to differentiated N2A neuroblastoma cells [56]. In our work, *NUMB* exon 10 was differentially spliced. However, we did not see differential expression of *PPRC1* as it was filtered before edgeR analysis due to low read count. In contrast, RBFOX, ELAVL, and PPRC1 were not selected as predictive features in the expression-based predictive model. Based on these findings, we conclude that additional biological information gained from splicing analysis could not be found using expression-based analysis alone.

Discussion

The major conclusions of the work herein are that skipped exon splicing data independently predict drug response and, when integrated with gene expression data, can increase the power of predictive drug response algorithms. These conclusions are supported by the following experimental evidence. First, we demonstrated the strong performance of the splicing-only elastic net GLM and determined that the most balanced model was obtained by combining splicing and expression data. Second, we showed that splicing-only and expression-only models had no genes in common, which indicates that each data type contributed unique information. Additionally, we demonstrated that an exon-centric approach positively impacts downstream analysis by identifying *cis*-acting RBP regulatory motifs, allowing researchers to find associations between regulatory elements of differentially spliced exons with essential biological processes. When employing RBP binding motif enrichment, we identified several candidate splicing factors, including RBFOX and ELAVL family members, which were differentially expressed between drug response groups. Moreover, we identified signatures of EMT, which affect cellular plasticity and stemness in tumor subpopulations and are thought to contribute to mechanisms behind cancer drug resistance [57–61]. Collectively, our results indicate that incorporating splicing information into predictive models improves performance and provides new biological insights.

Following our analysis, we assessed the cell line origin of the classified dataset to investigate if differences in the proportions of cell lineages could help explain the enriched biological processes we observed (Figure S3). The distribution of cell lineages, specifically the proportion of hematopoietic and lymphoid cells, differed greatly across sensitive and resistant groups: hematopoietic and lymphoid cells made up 44% of sensitive cell lines compared to only 1% of resistant cell lines. Hematopoietic cell types exhibit enhanced cytotoxicity to doxorubicin treatment, a consequence of treating highly proliferative cells with a topoisomerase inhibitor [62]. These cells, being more stem-like in nature compared to solid tumor tissue, are also expected to display signatures of EMT as stemness and EMT are related [63]. Our over-representation analysis of dif-

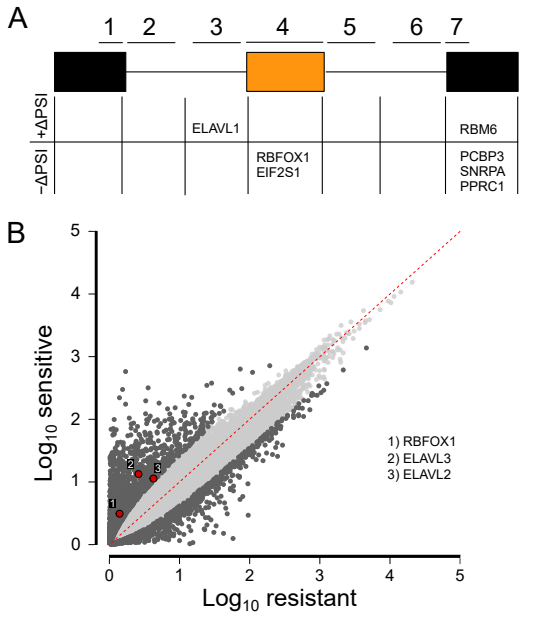


Figure 5 RBP binding motifs identified in differentially spliced events

A. Significantly enriched RBP binding motifs by skipped exon event region with respect to resistant cell lines. The schematic shows two constitutive exons (black boxes), one skipped exon (orange box), and two introns (connecting lines), as observed in skipped exon splicing. Regions of interest are shown as horizontal lines numbered 1 to 7. These regions consisted of: 1) up to 150 bp of the upstream exon; 2) 5′ 300 bp from the 5′ end of the upstream intron; 3) 300 bp from the 3′ end of the upstream intron; 4) the entire length of the skippedexon; 5) 300 bp from the 5′ end of the downstream intron; 6) 300 bp from the 3′ end of the downstream intron; and 7) up to 150 bp of the downstream exon. +ΔPSI and −ΔPSI indicate a higher and lower exon frequency in resistant cells, respectively. RBP binding motifs were identified using FIMO, analyzed for enrichment against motifs found in randomly drawn events, and deemed significant for association with included or excluded exons by Fisher's exact test. **B.** Mean expression levels of genes in sensitive and resistant groups using $\log_{10}$ read counts. Three differentially expressed RBPs belonging to the RBFOX or ELAVL families are numbered and shown in red. RBP, RNA binding protein; PSI, percent-spliced-in.

ferentially spliced events identified a number of biological processes, including EMT, proliferation, and drug resistance. Among other biological processes, we identified an exon in *SCRIB* previously described by Shapiro et al. [45] to be alternatively spliced and associated with an EMT signature. It is possible that the machine learning models picked up signatures related to proliferative and stem-like differences in the underlying cell types. However, we applied our modeling approach to the other 500 drugs in CTRP and observed strong performance for the vast majority (Figure S4; Table S7); many of these other drugs did not exhibit a large difference in cell type proportion (Figure S5). Further, the accuracy of our splicing-only model (82%) and accuracy of our combined model (83%) exceeded the proportion of hematopoietic and lymphoid cells in the testing set. This indicated that cell lines from other lineages are also discriminated by our classifiers according to their splicing profiles.

To determine the potential influence that cell type distribution might have had on the QBGLM results, we performed QBGLM in a tissue-specific manner on hematopoietic and lymphoid tissue and lung tissue. We then overlapped the differentially spliced events from all tissues as well as hematopoietic and lymphoid tissue and lung tissue, and found that hematopoietic and lymphoid tissue and lung tissue had many events in common (Figure S6). The vast majority of events found to overlap between the specific tissues and all tissues were present in both hematopoietic and lymphoid tissue and lung tissue. Additionally, we did not observe an imbalance in the number of events from the overlaps between all tissues and specific tissues. These findings support the conclusion that QBGLM also identified events from other tissue types besides hematopoietic and lymphoid and that many events found in hematopoietic and lymphoid cells are recapitulated by other cell types. Lastly, we did not consider predictive features from modeling during follow-up analysis because machine learning techniques arbitrarily select the most predictive features without regard for biological significance. Instead, we used only those events identified by QBGLM to explore relevant genes, pathways, and RBP binding motifs. Analyzing events from QBGLM allowed us to capture a more comprehensive list of differentially spliced events that are more likely to be relevant to the other cell types in the dataset.

To accurately assess the contribution of differential splicing to predictive drug modeling, we sought to identify a comprehensive and well-characterized dataset with drug response measurements and paired RNA-seq data. Although the widespread availability of high-throughput datasets offered a number of options for computational modeling, the majority of large-scale studies were done before RNA-seq became the predominant expression quantification method, and most of the pharmacological profiling experiments were paired with array-based expression data, making splicing analysis impossible. We searched for datasets with large numbers of samples to increase the power of our machine learning-based approach and to avoid overfitting. We also targeted diverse datasets to investigate predictive features with broad applicability, corresponding to multiple drugs and cell types. These criteria led us to integrate two large independent datasets rather than use a single resource with limited transcriptomic or pharmacological data.

While some investigators have challenged the integration of drug response datasets, the integration of these resources by others has shown reasonable consistency [64,65]. Additionally, other investigators have argued that isolated testing of individual cancer cell lines is an incomplete representation of tumors and that databases containing large collections of cells better represent the heterogeneity and

Table 2 **Enriched RBPs identified by motifs in significant events from QBGLM**

RBP	Position	Enrichment *P* value	Exon inclusion	Inclusion *P* value	Description	Refs.
RBFOX1	Skipped exon	8.7E−08	−	0.015	RBFOX1 and its family members (RBFOX2 and RBFOX 3) bind to (U)GCAUG stretches; they are generally found to enhance splicing when bound downstream and suppress splicing when bound upstream	[47,48]
EIF2S1	Skipped exon	1.6 E−07	−	0.011	EIF2S1 (or EIF2alpha) is one of three key members of the EIF2 complex and is responsible for delivering Met-tRNA for initiation of translation	[47]
RBM6	3′ exon	1.2 E−04	+	0.021	RBM6 is an RBP first identified by cloning a tumor suppressor locus and has been linked to lung as well as other cancers	[49]
PPRC1	3′ exon	1.6 E−04	−	0.021	PPRC1 (or PGC-1) is a coactivational transcription factor commonly associated with metabolic stress and little is known about its potential role in splicing; however, an important paralog of this protein (PGC-alpha) has been connected to altered splicing of VEGF	[47,50]
ELAVL1	5′ intron	1.9 E−03	+	0.005	ELAVL family members traditionally bind to AU-rich elements in 3′ UTR of mRNA	[51,47]
PCBP3	3′ exon	0.015	−	0.002	PCBP3 is a member of the poly(rC)-binding protein family and is paralogous to PCBP1/2/4; members of this family have strong motif homology and share a wide variety of functions, but PCBP3 lacks the nuclear localization signals that other members have	[47,52]
SNRPA	3′ exon	0.026	−	0.015	SNRPA is an essential component of the U1 splicing complex and is required for recognition of the pre-mRNA 5′ end; the U1 complex binds to the 5′ splicing site of an exon-intron boundary	[51]

Note: The enrichment *P* value is the FDR-adjusted *P* value against randomly bootstrapped events from the genome. Exon inclusion is with respect to resistant cells, where "+" represents higher exon frequency in resistant cell lines and "−" represents lower exon frequency in resistant cell lines. The inclusion *P* value was calculated using Fisher's exact test. RBP, RNA-binding protein; QBGLM, quasi-binomial generalized linear modeling; RBFOX1, RNA-binding protein Fox-1 homolog 1; EIF2S1, eukaryotic translation initiation factor 2 subunit alpha; RBM6, RNA binding motif protein 6; PPRC1, peroxisome proliferator-activated receptor gamma coactivator-related protein 1; ELAVL1, ELAV-like RNA-binding protein 1; PCBP3, poly(rC)-binding protein 3; SNRPA, small nuclear ribonucleoprotein polypeptide A.

tissue-level characteristics of cancer [66]. Because our goal was to specifically target global splicing patterns, we sought to use large datasets to reduce the impact of individual differences across databases. Therefore, we feel that our approach accurately reflected the transcriptomic and drug response measures of various cancer types, that the number and composition of cell lines in it reduced the possible influence of lineage inconsistency, and that our dataset is a reliable source of information for investigating global trends in transcript splicing or expression.

When performing machine learning, we elected to build a classification model rather than a continuous model as the CCLE and Genomics of Drug Sensitivity in Cancer (GDSC) consortia recommended dividing cell lines into sensitive and resistant groups when analyzing drug response data across datasets [67]. This recommendation was based on the observation that using all cell lines in a database tended to introduce noise due to increased drug response variance from cell lines that did not have influential genetic differences [67]. We analyzed performance consistency by bootstrapping the model building procedure and found that combined and expression-based models were almost equivalent (Figure S2). While we did not find splicing-based model to outperform expression-based model as previous researchers have [24,25], our approach differed from these earlier models as it was designed to determine the importance of alternative splicing in doxorubicin drug response using a minimalistic procedure rather than generating the best possible classifier. Nevertheless, while our work provides evidence that adding splicing information to expression-based models in a more controlled manner produces a better classifier, there remains room for improvement in the model building process.

Finally, we noted that differential splicing analysis with QBGLM could be achieved in minutes. Even for groups containing hundreds of samples, the analysis time is negligible if inclusion and exclusion reads are counted beforehand as part of a standard pipeline. While our analysis works well for large groups of samples, it struggles with smaller sets; however, we expect the model's ability to handle large groups of samples to be a key strength, as the volume of sequencing data and the number of samples included in studies continue to rise.

The experimental evidence from this study strongly supports the overall hypothesis that alternative splicing data can be used to predict doxorubicin drug response, and that splicing data can contribute to valuable insights into drug response mechanisms. Resistance to doxorubicin has long been a major challenge and facilitates the resurgence of disease as well as increased patient mortality [28,68]. Nevertheless, doxorubicin remains a widely prescribed antineoplastic agent and is extremely important in breast cancer treatment [68]. Recently, discoveries in targeted drug delivery have expanded the variety of cancers that can be treated with doxorubicin, and there is now a greater focus on combating doxorubicin resistance [69–71]. Our findings suggest that splicing information could uncover new avenues for improving the effectiveness of doxorubicin treatment. Additionally, our findings indicate that there is much more to learn about the influence of splicing on cancer drug

response. Ultimately, splicing information may have a major impact on how, or under what circumstances, doxorubicin and other cancer therapeutics are used.

Materials and methods

Datasets

975 RNA-seq files corresponding to pretreatment cancer cell lines were downloaded from the CCLE and matched to post QC AUC values for 860 cancer cell lines from the CTRP v2 using the cell line name [26,27]. While integrating data from two separate sources is not ideal, this approach was chosen because it provided the largest available overlap between RNA-seq and drug profiling data. Intersecting these data sets for cell lines profiled with doxorubicin yielded 755 cell lines with drug response and RNA-seq data. Cell lines were split into three groups using the tertiles of the AUC distribution. The low AUC group was labeled "sensitive" (n = 253), the high group was "resistant" (n = 258), and the middle group was omitted from the analysis.

MISO splicing analysis

Splicing analysis for predictive modeling was done with MISO [21]. RNA-seq files belonging to sensitive and resistant groups were analyzed using exon-centric version 2 annotations for hg19 and the standard pipeline from the MISO documentation website, http://miso.readthedocs.io/en/fastmiso/. Data corresponding to 40,178 skipped exon events was obtained.

Gene expression quantification and differential expression analysis

Read counts for predictive modeling with expression data and for differential expression analysis were calculated with featureCounts [30]. A genomic feature was defined as any record with a valid gene_id and was counted at the meta-feature level. RNA-seq files were processed for 57,095 genomic features that were annotated in the GRCH37(v87) GTF file downloaded from ftp.ensembl.org, using a minimum read length overlap of 2 bp. Differential expression analysis was performed on featureCount data from the training dataset using edgeR [31]. Only features with more than 10 reads in more than 35% of training cell lines were evaluated, leaving 22,201 features before differential expression and downstream filtering. Log_{10} counts per million were used as feature values. The same annotation set of quantified genomic features as those used for predictive modeling (57,095) were again used for assessing differential expression of genes coding for RBPs. In this case, filtering to include features with at least 10 reads in more than 20% of cell lines reduced the number to 28,110 before differential expression analysis. In edgeR, a negative binomial generalized log-linear model with quasi-likelihood F-test (glmQLFit) was used. Differentially expressed features with an FDR ≤ 0.05 and a $\log_2$ fold change ≥ 1.5 were considered significant, producing a final number of 2943 differentially expressed gene features.

Predictive modeling

Using the glmnet and caret packages in the R language, elastic net logistic regression was used to fit all predictive models [72,73]. Following splicing and expression analyses, feature selection was performed to restrict the parameters of the models. Splicing features were defined as skipped exon events identified by MISO and were required to have PSI values with CIs between 0.01 and 0.2 for a minimum of 35% of cell lines. This requirement reduced the number of potential splicing features from 40,178 to 15,007. We also filtered events having a PSI standard deviation less than 0.14, based on the top 5% of the remaining skipped exon events, which reduced the number of splicing features to 805. Any missing values were then imputed randomly from all samples with data for a particular event. Gene expression features were filtered by requiring a minimum of 10 reads in more than 35% cell lines. This lowered the number of potential features from 57,905 to 22,201. Cell lines were divided into training (70%) and testing (30%) sets. This produced 354 training cell lines (177 sensitive and 177 resistant) and 157 testing cell lines (76 sensitive and 81 resistant). Individual and combined models were trained on the same training cell lines. Expression features were further restricted after training and testing set separation by conducting differential expression analysis on the training set and applying the cutoffs: FDR < 0.05 and $\log_2$ fold change > 1.74 (top 5%). Expression- and splicing-only models were then trained using their respective filtered feature sets, while the combined model was trained by merging the two filtered feature sets and allowing the elastic net to choose freely between the whole.

A 10-fold cross-validation approach with grid search (to scan for the highest performing alpha and lambda values) was used to train the models. The models were then assessed with the testing cell line data. Sensitivity, specificity, accuracy, and precision were calculated. The AUROC, F1 score, and P value (corresponding to accuracy against the no-information rate) were also produced. Lastly, when building models to assess the generalizability with the remaining 500 drugs in CTRP, all event types including skipped exon, mutually exclusive exon, retained intron, alternative 5′ splice site, and alternative 3′ splice site were used for modeling.

Differential splicing analysis by QBGLM

Exon-centric splicing is usually characterized by an inclusion ratio or PSI value [74]. This PSI value ranges from 0 to 1, describes the inferred percentage of transcripts containing the exon, and is heavily dependent on sequence information spanning exon–exon boundaries. Reads supporting inclusion are those reads overlapping the upstream and skipped exon junction, skipped and downstream exon junction, or all three exons including both junction boundaries. Reads supporting exclusion are those reads overlapping the upstream and downstream exon junction boundaries but not the skipped exon. The PSI value effectively compresses information from a distribution of mapped reads, where reads may correspond to multiple isoforms, into a single number. PSI has been treated as a point estimate with a margin of error rather than a definitive ratio [21]. Techniques for calculating this metric can vary; however, the more popular methods rely on length-normalized read density and may include counts for non-junction reads, as well as iterative procedures for establishing a CI.

Splicing analysis by the QBGLM was done using raw read counts. A total of 38,108 skipped exon events were extracted from isoforms annotated in the GRCH37(v87) GTF file downloaded from ftp.ensembl.org. Uniquely mapped and properly paired junction reads with a minimum exon overlap of 1 bp supporting the inclusion or exclusion of skipped exons were counted for each skipped exon event. After counting, events were filtered, retaining only those with at least 1 inclusion and 1 exclusion read in 35% of classified cell lines. A total of 18,409 events passed the filter. A QBGLM was fit using the glm package in R [33]. The inclusion read percentage for a given event was modeled as the probability of success. The cell line label (sensitive or resistant) was set as the dependent variable.

Events were filtered for significance by requiring a Benjamini-Hochberg adjusted P value ≤ 0.01 on the group weight ($Beta_1$). A total of 4309 events passed the filter. Events were further separated for relevance using the difference (Δ) in mean inclusion-to-total read counts (inclusion/inclusion + exclusion) in each group. A minimum of 0.1 difference in mean inclusion-to-total read counts between sensitive and resistant groups was required to maximize biological relevance; only 277 events met this threshold.

GO over-representation analysis

Significant skipped exon events identified from QBGLM were annotated for gene symbols by the genomic positions of the skipped exons using the biomaRt package [75,76]. Gene symbols for the entire set of significant events were then analyzed with the clusterProfiler packa-ge [35]. Results from biological process enrichment based on the GO database were then exported and assessed for relevance.

Motif enrichment

Significant skipped exon events were analyzed for enrichment of RBP binding motifs in three stages: 1) motif matches were counted for significant events in a region of interest; 2) the total count for the set of significant events was compared to a background of randomly drawn events; 3) significantly enriched motifs found were then filtered and sorted based on their associated splicing outcome. Seven regions surrounding each exon of interest (Figure 5A) were extracted from hg19 (GRCh37). These regions were: 150 bp maximum or the full length of the 5′ upstream exon, 300 bp of its 3′ flanking intron, 300 bp in the 5′ upstream intron flanking the skipped exon, the entire length of the skipped exon, 300 bp in the 3′ downstream flanking intron, 300 bp in the 5′ intron flanking the 3′ downstream exon, and 150 bp maximum or the full length of the 3′ downstream exon. Sequences for each region were scanned using FIMO and the CISBP-RNAv0.6 RNA-binding motif database [46,77]. Using a P value threshold for motif matches of 6.7E–4, as compared to the default 1E–4, it was necessary to find small splicing factor motifs in short extracted sequence lengths. Counts across significant events for a given motif were then compared to the genomic background in context by bootstrapping the same number of skipped exon events (without replacement) from all annotated events in the genome, repeating the procedure 10,000 times. P values for significant event motif counts were then calculated using this random distribution and adjusted using the Benjamini-Hochberg method. This is referred to as the enrichment P value in Table 2. Fisher's exact test was then used on enriched motifs to identify those associated with preferential increased or decreased exon inclusions. P values from Fisher's exact test are referred to as the inclusion P value in Table 2. In both enrichment and preferential inclusion analyses, a minimum P value of 0.05 was required.

Data availability

All data used in this study are publicly available through the CCLE (https://sites.broadinstitute.org/ccle) and CTRP (https://portals.broadinstitute.org/ctrp.v2). The code used for processing and analyzing the data is available on request.

CRediT author statement

Edward Simpson: Methodology, Software, Data curation, Formal analysis, Validation, Visualization, Writing - original draft. **Steven Chen:** Methodology. **Jill L. Reiter:** Writing -

review & editing. **Yunlong Liu:** Conceptualization, Supervision. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Institutes of Health, USA (Grant No. R01CA213466) awarded to YL. This work was also supports by the Precision Health Initiative at Indiana University.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.12.002.

ORCID

0000-0001-9015-9864 (Edward Simpson)
0000-0003-3463-5824 (Steven Chen)
0000-0001-5460-2355 (Jill L. Reiter)
0000-0002-2699-626X (Yunlong Liu)

References

[1] Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet 2001;17:100–7.
[2] Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol 2017;18:437–51.
[3] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5.
[4] David CJ, Chen M, Assanah M, Canoll P, Manley JL. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature 2010;463:364–8.
[5] Dehm SM. mRNA splicing variants: exploiting modularity to outwit cancer therapy. Cancer Res 2013;73:5309–14.
[6] Zammarchi F, de Stanchina E, Bournazou E, Supakorndej T, Martires K, Riedel E, et al. Antitumorigenic potential of STAT3 alternative splicing modulation. Proc Natl Acad Sci U S A 2011;108:17779–84.
[7] Hernandez-Lopez HR, Graham SV. Alternative splicing in human tumour viruses: a therapeutic target? Biochem J 2012;445:145–56.
[8] Pawellek A, McElroy S, Samatov T, Mitchell L, Woodland A, Ryder U, et al. Identification of small molecule inhibitors of pre-mRNA splicing. J Biol Chem 2014;289:34683–98.
[9] Bauman JA, Kole R. Modulation of RNA splicing as a potential treatment for cancer. Bioengineered Bugs 2011;2:125–8.
[10] Niedermeier M, Hennessy BT, Knight ZA, Henneberg M, Hu J, Kurtova AV, et al. Isoform-selective phosphoinositide 3′-kinase inhibitors inhibit CXCR4 signaling and overcome stromal cell–mediated drug resistance in chronic lymphocytic leukemia: a novel therapeutic approach. Blood 2009;113:5549–57.
[11] Cesi G, Philippidou D, Kozar I, Kim YJ, Bernardin F, Van Niel G, et al. A new ALK isoform transported by extracellular vesicles confers drug resistance to melanoma cells. Mol Cancer 2018;17:145.
[12] Peng H, Peng T, Wen J, Engler DA, Matsunami RK, Su J, et al. Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. Bioinformatics 2014;30:1899–907.
[13] Ashley EA. Towards precision medicine. Nat Rev Genet 2016;17:507–22.
[14] Ogilvie LA, Wierling C, Kessler T, Lehrach H, Lange BMH. Predictive modeling of drug treatment in the area of personalized medicine. Cancer Inform 2015;14:95–103.
[15] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[16] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;2:59–77.
[17] Vickers AJ. Prediction models in cancer care. CA Cancer J Clin 2011;61:315.
[18] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform 2016;18:bbw065.
[19] Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 2014;32:1202–12.
[20] Amin SB, Yip WK, Minvielle S, Broyl A, Li Y, Hanlon B, et al. Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. Leukemia 2014;28:2229–34.
[21] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 2010;7:1009–15.
[22] Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, et al. Prediction and quantification of splice events from RNA-seq data. PLoS One 2016;11:e0156132.
[23] Zhang C, Zhang B, Lin LL, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 2017;18:583.
[24] Shen S, Wang Y, Wang C, Wu YN, Xing Y. SURVIV for survival analysis of mRNA isoform variation. Nat Commun 2016;7:11548.
[25] Safikhani Z, Smirnov P, Thu KL, Silvester J, El-Hachem N, Quevedo R, et al. Gene isoforms as expression-based biomarkers predictive of drug response *in vitro*. Nat Commun 2017;8:1126.
[26] Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 2013;154:1151–61.
[27] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
[28] Cox J, Weinman S. Mechanisms of doxorubicin resistance in hepatocellular carcinoma. Hepatic Oncol 2016;3:57–9.
[29] Eliaz RE, Nir S, Marty C, Szoka Jr. FC. Determination and modeling of kinetics of cancer cell killing by doxorubicin and doxorubicin encapsulated in targeted liposomes. Cancer Res 2004;64:711–8.
[30] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30.
[31] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.
[32] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57:289–300.
[33] Core R Team. R: a language and environment for statistical computing. Austria: R Foundation for Statistical Computing; 2017.
[34] Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool en-

hancements. Nucleic Acids Res 2017;45:D183–9.
[35] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol 2012;16:284–7.
[36] González-Mariscal L, Lechuga S, Garay E. Role of tight junctions in cell proliferation and cancer. Prog Histochem Cytochem 2007;42:1–57.
[37] Martin TA, Jiang WG. Loss of tight junction barrier function and its role in cancer metastasis. Biochim Biophys Acta 2009;1788:872–91.
[38] Tsukita S, Yamazaki Y, Katsuno T, Tamura A, Tsukita S. Tight junction-based epithelial microenvironment and cell proliferation. Oncogene 2008;27:6930–8.
[39] Provenzano PP, Keely PJ. Mechanical signaling through the cytoskeleton regulates cell proliferation by coordinated focal adhesion and Rho GTPase signaling. J Cell Sci 2011;124:1195–205.
[40] Parri M, Chiarugi P. Rac and Rho GTPases in cancer cell motility control. Cell Commun Signal 2010;8:23.
[41] Rhyu MS, Jan LY, Jan YN. Asymmetric distribution of numb protein during division of the sensory organ precursor cell confers distinct fates to daughter cells. Cell 1994;76:477–91.
[42] Zhao C, Chen A, Jamieson CH, Fereshteh M, Abrahamsson A, Blum J, et al. Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukaemia. Nature 2009;458:776–9.
[43] Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 2016;16:413–30.
[44] Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell 2013;52:720–33.
[45] Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. PLoS Genet 2011;7:e1002218.
[46] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8.
[47] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45.
[48] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456:470–6.
[49] Wang K, Ubriaco G, Sutherland LC. RBM6-RBM5 transcription-induced chimeras are differentially expressed in tumours. BMC Genomics 2007;8:348.
[50] Saint-Geniez M, Jiang A, Abend S, Liu L, Sweigard H, Connor KM, et al. PGC-1α regulates normal and pathological angiogenesis in the retina. Am J Pathol 2013;182:255–65.
[51] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49:D480–9.
[52] Makeyev AV, Liebhaber SA. The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. RNA 2002;8:265–78.
[53] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40:4288–97.
[54] Venables JP, Brosseau JP, Gadea G, Klinck R, Prinos P, Beaulieu JF, et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. Mol Cell Biol 2013;33:396–405.
[55] Wen J, Toomer KH, Chen Z, Cai X. Genome-wide analysis of alternative transcripts in human breast cancer. Breast Cancer Res Treat 2015;151:295–307.
[56] Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KCH, et al. Multilayered control of alternative splicing regulatory networks by transcription factors. Mol Cell 2017;65:539–53.e7.
[57] Pradella D, Naro C, Sette C, Ghigna C. EMT and stemness: flexible processes tuned by alternative splicing in development and cancer progression. Mol Cancer 2017;16:8.
[58] Seguin L, Kato S, Franovic A, Camargo MF, Lesperance J, Elliott KC, et al. An integrin β3–KRAS–RalB complex drives tumour stemness and resistance to EGFR inhibition. Nat Cell Biol 2014;16:457–68.
[59] Ma JL, Zeng S, Zhang Y, Deng GL, Shen H. Epithelial–mesenchymal transition plays a critical role in drug resistance of hepatocellular carcinoma cells to oxaliplatin. Tumor Biol 2016;37:6177–84.
[60] Shang Y, Cai X, Fan D. Roles of epithelial-mesenchymal transition in cancer drug resistance. Curr Cancer Drug Targets 2013;13:915–29.
[61] Salt MB, Bandyopadhyay S, McCormick F. Epithelial-to-mesenchymal transition rewires the molecular path to PI3K-dependent proliferation. Cancer Discov 2014;4:186–99.
[62] Minderman H, Linssen PC, Wessels JM, Haanen C. Doxorubicin toxicity in relation to the proliferative state of human hematopoietic cells. Exp Hematol 1991;19:110–4.
[63] Ye X, Weinberg RA. Epithelial–mesenchymal plasticity: a central regulator of cancer progression. Trends Cell Biol 2015;25:675–86.
[64] Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell 2016;166:740–54.
[65] Pozdeyev N, Yoo M, Mackie R, Schweppe RE, Tan AC, Haugen BR. Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. Oncotarget 2016;7:51619–25.
[66] Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. J Natl Cancer Inst 2013;105:452–8.
[67] The Cancer Cell Line Encyclopedia Consortium, The Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature 2015;528:84–7.
[68] AbuHammad S, Zihlif M. Gene expression alterations in doxorubicin resistant MCF7 breast cancer cell line. Genomics 2013;101:213–20.
[69] Arora HC, Jensen MP, Yuan Y, Wu A, Vogt S, Paunesku T, et al. Nanocarriers enhance doxorubicin uptake in drug-resistant ovarian cancer cells. Cancer Res 2012;72:769–78.
[70] Chamberlain GR, Tulumello DV, Kelley SO. Targeted delivery of doxorubicin to mitochondria. ACS Chem Biol 2013;8:1389–95.
[71] Zhao Y, Tang S, Guo J, Alahdal M, Cao S, Yang Z, et al. Targeted delivery of doxorubicin by nano-loaded mesenchymal stem cells for lung melanoma metastases therapy. Sci Rep 2017;7:44758.
[72] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
[73] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.
[74] Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh CS, et al. Identification of alternative splicing markers for breast cancer. Cancer Res 2008;68:9525–31.
[75] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21:3439–40.
[76] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009;4:1184–91.
[77] Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 2013;499:172–7.