**Functional consequences of dysfunctional splicing machinery and alternative splicing events across the cancer cell line encyclopedia**

Angelo Limeta1,✝, Leticia Castillon1,2,✝, Francesco Gatto1,3 and Jens Nielsen1,4,5

1 Department of Biology and Biological Engineering, Chalmers University of

Technology, Gothenburg, Sweden

2 Translational Cancer Medicine Research Program, Department of Surgery, Faculty of Medicine, Helsinki University Hospital, University of Helsinki, Helsinki, Finland

3 Elypta AB, SE114 28 Stockholm, Sweden

4 Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK2800 Kgs. Lyngby, Denmark

5 BioInnovation Institute, Ole Måløes Vej 3, DK2200 Copenhagen N, Denmark

✝ These authors contributed equally

Correspondence to; [nielsenj@chalmers.se](mailto:nielsenj@chalmers.se)

**Abstract**

The deregulation or disruption of the splicing process has been shown to play a role in the onset, development, and even response to treatment of some malignancies. Partly due to our incomplete understanding of the mechanism and regulation behind splicing and alternative splicing, the importance of aberrant splicing in oncogenesis is not yet understood. In this project, we set out to perform a systematic analysis of aberrant splicing events in the cancer cell lines from two perspectives: deregulated splicing because of dysfunctional splicing factors and the appearance of de novo splicing events because of splice-disruption mutations in the DNA sequence. Examining differences in the isoform expression landscape of cell lines with mutated vs wild-type spliceosomes revealed dysregulations in isoforms belonging to targets of the Myc transcription factor.

All the analyses have been performed using data publicly available in the DepMap portal. For the analysis of de novo splicing events caused by mutations in the DNA sequence, we used the deep-learning tool SpliceAI. To the extent of our knowledge, SpliceAI has not been previously used to analyze RNA-sequencing from cancerous samples - cell lines nor human tumour samples. In this project, we decided to test the deep learning algorithm in data from the Cancer Cell Line Encyclopedia collected by the DepMap consortium and evaluate its performance to detect splicing alterations that may affect oncogenesis. In order to try to evaluate the clinical relevance of ou findings, we used the MSK-IMPACT sequencing panel to narrow down the analysis to actionable genes - genes that can be targeted by drugs.

**Introduction**

Alternative splicing is a regulatory mechanism affecting approximately 95% of mammalian genes and the main contributor to transcriptome and proteome diversity in eukaryotes [1], [2]. On top of that, alternative splicing also affects regulatory processes such as chromatin modification and signal transduction [1], [2]. Thus, it is not surprising that alterations or errors in the splicing process may occur through disease-causing mutations - with an estimate of 15 to 50% of human disease causing mutations appearing to affect splice site selection [3], [1]. Furthermore, aberrant splicing is widespread in cancer, with splicing abnormalities conferring new capabilities to the cell that contribute to several hallmarks of cancer [4], [5], [6]. For example, splicing isoforms driven by SRSF1, one of the best studied splicing factors, participate in the maintenance of most hallmarks of cancer: sustaining of proliferative signaling, evasion of growth suppressors, escape from apoptosis, angiogenesis and invasion [6].

Such evidence has led to the development of anti-cancer compounds targeting the splicing process - splicing inhibitors - which may offer new therapeutic windows against defective splicing driven malignancies [6], [7]. In fact, H3B-8800, a recently reported modulator of the SF3b complex, has shown promise in clinical trials [7].

The process of splicing is performed by a dynamic ribonucleoprotein complex called the spliceosome [2], [8]–[11] [2], [8]–[11]. The spliceosome recognizes and assembles around DNA sequences known as splice sites. However, the spliceosome has different affinity for different splice sites, being able to recognize more often “strong” splice sites than other “weak” splice sites [2]. The efficiency by which the splicing factors are able to recognize the splice sites is completely context-dependent, with the position of functional splice sites and additional regulatory elements in the DNA sequence having a great influence in the recognition and binding to the DNA of spliceosome proteins [11], [12]. The competence between strong and weak splice sites is the main force behind alternative splicing. However, deregulation of the splicing process may be driven by somatic mutations, antisense RNAs, RNA binding proteins, transcriptional factors and epigenetic factors [13]. Additionally, such events may affect both the proteins that are part of the spliceosome complex, and the consensus DNA sequences that act as targets for the splicing factors.

Here, we propose a systemic data-driven approach to understand the effects of somatic mutations in the splicing process using the cell lines studied in the Cancer Cell Line Encyclopedia (CCLE) and the Broad Cancer Dependency Map (DepMap) projects [14]. We approached this problem from two different perspectives: on one hand, we compared the state of the transcriptome in cell lines presenting mutations in the splicing factors - the genes forming the spliceosome machinery. Previous research has shown that mutations in the splicing factors may affect their ability to interact with other members of the spliceosome machinery or to bind to the splicing sites in the DNA, resulting in usage of cryptic splicing sites [5]. On the other hand, we used the deep learning tool SpliceAI to understand how point mutations or other alterations in the genome may alter the splicing process of a gene, by the creation of new splice sites and/or the removal of existing ones [15]. The deep learning tool showed great potential to identify cryptic splice mutations in rare genetic diseases, in particular childhood neurodevelopmental disorders such as autism, where cryptic splice variants have been found to play a critical role [16]. To the best of our knowledge, SpliceAI has not been previously used to characterize the effect of somatic mutations in the context of cancer cell lines. **Results**

**Mutational landscape of splicing factors**

We studied the distribution of mutations in core and accessory splicing factors across the entire CCLE, limiting our study to somatic mutations that would probably lead to a deleterious effect in the protein, namely mutations causing insertions or deletions, or incorporating a stop codon in the DNA sequence **(Fig. 1A).** Overall, the mutational load in spliceosome proteins varies greatly across cancer types and correlates well with the total mutational load of each cell line, an indication that no cancer type seems to present a significant enrichment in mutations in the spliceosome **(Fig. 1B-C)**. We observed differences in the frequency of deleterious mutations among individual core splicing factors across cell lines, with some of them, including SF3B1, TCERG1, PPW1 and SDE2 being often mutated across a variety of cancers **(Fig. 1C).** Seeing that mutations in splicing factors are ubiquitous across most of the cell lines in the Cancer Cell Line Encyclopedia, we decided to investigate further whether alterations in the spliceosome would lead to clear differences in the transcriptional program of the cell lines. For this, we performed PCA on the gene-level expression data (reported by CCLE in the form of TPM), labelling the cell lines according to whether mutations in genes encoding for core spliceosome proteins were present - resulting in two classes of cell lines: spliceosome mutated and wild-type spliceosome **(Fig. S1)**. The PCA showed a slight separation between the two classes of cell lines in some malignancies: lung cancer, bladder cancer or breast cancer. However, the observed differences were not striking enough to conclude that mutations in the spliceosome lead to major changes in the expression landscape of the cell. Likewise, extending this analysis to mutations in proteins that are part of the non-core splicing factors and in proteins that are involved in splicing regulation but not part of the spliceosome itself showed a similar trend **(Fig. S2)**.

Diagram

Description automatically generated with low confidence

**Figure 1. Distribution of mutations in splicing factors across the CCLE. A)** Number of core and accessory splicing factors included in this analysis. **B)** Average number of deleterious mutations in core splicing factors grouped by cancer type of each cell line. **C)** Binary clustering and heatmap of deleterious mutations in core splicing factors across all cell lines in the CCLE. Red cells indicate that a particular cell line contains a deleterious mutation in the corresponding splicing factor. Columns (cell lines) are color coded by cancer type according to the same color schema as in **B)**.

**Transcriptional differences related to spliceosome deficiencies in cancer**

Differential expression at the gene-level using a linear regression model to evaluate the relationship between expression levels of the isoforms and mutations in the spliceosome **(Materials & Methods)** revealed no significant transcriptional changes occurred at the gene-level. However, considering that mutations in the splicing factors are expected to mostly alter the splicing process, which affects the maturation of the pre-mRNA, it is not entirely surprising that these changes are not reflected in the expression of the gene, but might instead alter the relative expression of isoforms **(Fig. 2A)**. Examining the differences in diversity of expressed transcript isoforms in the CCLE for cell lines harboring wild-type spliceosomes versus cell lines containing deleterious mutations in the core-spliceosome highlighted overall increases in isoform diversity in the latter group **(Fig. 2B)**. For this reason, we decided to focus on changes occurring in the transcriptome by analyzing the relative expression of transcript isoforms. In order to perform differential expression and functional enrichment at the transcript isoform level, we developed custom workflow, as most tools for differential expression analysis are designed to work using gene-level p-values, regardless of the isoform type **(Fig. 1C)**. First, a linear regression model was used to obtain transcript-level p-values comparing mutated cell lines versus spliceosome-wild-type cell lines. To obtain biological relevant insights from the data, we aggregated the transcript-level p-values into gene-level p-values, while retaining the variability observed at the transcript level. To achieve this, we followed the framework suggested by Yi et al. (2018) [17] to translate transcript-level fold-changes and p-values into gene-level p-values using the Lancaster method [18]. Once gene-level statistics were re-generated with incorporated isoform information, we used them as an input for the gene-set analysis software PIANO [19]. We used PIANO to analyze cell lines with mutations in the core spliceosome, with mutations in non-core splicing factors, and mutations in proteins related to the splicing process but lacking classification as either core or non-core **(Methods, Table S1)**. There are several pathways that appear enriched in genes showing distinct isoform variation, many of which are also important for cancer progression, including the PI3K/Akt/mTOR pathway, oxdative phosphorylation, cell cycle, unfolded protein response and Myc targets **(Fig. 1D)**. The gene sets consisting of genes that are targets for the transcription factor Myc consistently appear as enriched in genes showing differential expression when having mutated spliceosome proteins - for core spliceosome proteins, non-core spliceosome proteins, and uncategorized splicing factors. The gene sets contain a subgroup of genes regulated by Myc, a transcription factor that regulates important cell processes such as cell growth, differentiation, metabolism and death. Myc has been intensively studied in the context of cancer because it appears deregulated in over 50% of human cancers [20] and it has been associated with poor prognosis and patient survival. Furthermore, although its function in many malignancies would make it an attractive therapeutic target, due to its structure Myc has been rendered “undruggable” thus far [21]. Examining the top 10 most significantly deregulated genes in the Myc targets gene sets for differences in expression for canonical and non-canonical transcript isoform between cell lines containing wild-type vs mutated core-spliceosomes displayed overall increases in isoform expression for both canonical and non-canonical transcripts **(Fig. 2E**,p < 0.001, two-way ANOVA test**)**. This suggests that targeting Myc and subsequently the expression of its target genes might be an attractive therapeutic avenue for spliceosome-deficient cancers.

Diagram

Description automatically generated

**Figure 2. Transcript Isoform level differential expression analysis comparing cell lines with wild-type versus mutated core-spliceosomes. A)** Example visualization of why isoform level differences may not be captured by standard differential expression workflows. Although the isoform expression levels for a particular gene may vary greatly between conditions, gene level expression levels may not. **B)** Violin plot of the Shannon entropy distribution in cell lines with wild-type vs mutated core-spliceosomes. **C)** Workflow for isoform differential expression, Lancaster aggregation and functional analysis. **D)** Significant Hallmark gene sets following the differential isoform expression analysis **E)** Boxplots of the canonical and non-canonical isoform expression levels for top 10 most significant genes in the Myc targets gene sets across cell lines with wild-type vs mutated core-spliceosomes. log TPM: log-normalized transcripts-per-million. \* = FDR < 0.05, \*\* = FDR <0.01, \*\*\* = FDR <0.001, \*\*\*\* = FDR <0.0001, ns = non-significant.

**Predicting the effects of genomic variants on the splicing process**

Shifting our focus from mutations in the splicing machinery, we examined the effects of how individual genomic mutations may result in the generation new splice variants at nearby loci. We used a deep-learning based algorithm called SpliceAI for this analysis as it is the current state-of-the-art for predicting the effects on splicing for individual mutations. The SpliceAI algorithm classifies mutational events in cell lines according to whether a particular mutation in a genomic locus results in the alteration of the splicing process for neighboring genes and therefore results in the generation of new transcript variants for those particular genes **(Fig. 3A)**. For each mutation, it generates as a ground truth, we used the splice altering mutation as classified by DepMap, which tags known splice consensus sites as “Splice\_site” on the somatic mutation dataset. Therefore, this can be considered as a binary classification problem and the performance can be evaluated using receiver-operating-characteristic (ROC) and precision-recall curve analysis (PR). To evaluate whether SpliceAI has higher accuracy for specific cancer types, we visualized the distribution of the probabilities calculated by the algorithm per cancer type, labelling them according to whether they were categorized as a “Splice\_site” mutation by DepMap. Most of the mutations that had not been annotated as occurring in a splice site received a low probability from SpliceAI, while the majority of mutations that scored above the 0.5 threshold has been annotated as a splice site mutation by Depmap. In general, the algorithm performs consistently across cancer types; poorer performances correspond to malignancies with fewer cell lines available in the database - for example teratoma, a disease for which only one cell line was analyzed.

Diagram

Description automatically generated

**Figure 3. Splice variant predictions resulting from primary mutations in the genome. A)** Visualization of the SpliceAI algorithm. Individual variants and surrounding 10kb context sequence is used to directly predict alternative splicing using a deep neural network. **B)** Performance of SpliceAI on predicting annotated mutations resulting in splice variants in the CCLE. Left: violin plot for the distribution of probabilities of a mutation resulting in a splice variant, grouped by the variant type annotation in the CCLE. Middle & right: receiver-operating-characteristic and precision-recall curves for the prediction of splice variants for SpliceAI. AUC = area under the curve. **C)** Distribution of distances between mutations and splice site boundaries for all potential splice altering variants (Probability of being splice altering > 0.5). Red-shaded areas indicate +/- 5bp from the mutated base.

**Evaluation of the predicted variants**

Over 1 million somatic mutations were fed to SpliceAI from the somatic mutations database generated by DepMap. The neural network predicted that slightly over 2% of these were splice altering mutations.

Next, we were interested in knowing the distance from the mutation to the actual splice-altering event in the sequence. The figure shows that most of the predicted splicing events occur relatively close to the somatic mutation causing the alteration, yet the distribution is highly skewed towards the tails, which indicates that a minority of events have an effect elsewhere in the genome. This is further strengthened by the fact that almost 6% of the predicted splice-altering mutations have an effect in a different gene than the one carrying the somatic mutation - or alter the splicing for both the carrier gene and a second one. We defined these events as off-target splicing. A complete list of the predicted off-target splicing events can be found in …

Next, we wanted to know the distribution of the predicted splice-altering events across exons and introns. To achieve this, we used the transcription data provided by DepMap and selected all the transcripts that are generated from a gene carrying a somatic mutation. Most of the splice variants that are predicted by SpliceAI fall within intronic regions. ...

**Deleterious splice altering mutations are enriched in tumor suppressor proteins and may affect the therapeutic response of drugs**

**MSKCC**

We were interested in assessing the impact of the altered splicing events in the function of the gene and the sensitivity of the cell line to drugs. As explained before (*Refer to Methods section*), we focused on the genes included in the MSK-IMPACT panel. To obtain robust statistics, we decided to show only those genes for which a predicted variant was present in more than 10 cell lines. Interestingly, all the genes that meet this criterion are well-studied tumor suppressors [22], [23], [24], [25], [39]. The Figure … shows the effect of cell line growth of knocking out a certain gene. A negative score means that depleting the gene has a negative impact on the growth, whereas a positive score means that deletion of the gene improves the growth of the cell line. We can observe that for those cell lines carrying splice-altered variants, the scores tend to be closer to 0, implying that knocking-out the altered gene has no (or little) effect in growth. Conversely, for all genes (except CDK2A) lacking splice-altering variants, the effect of we see a significant difference in the . Similarly, we wanted to see whether splice-altering mutations could cause differences in the sensitivity to certain drugs. For this analysis, we used the sensitivity data generated by the PRISM repurposing project [26] … should we add the sensitivities faceted per cancer?

**Diagram

Description automatically generated**

**Figure 4. Consequences of perturbing genes harboring putative splice variants in the MSK-IMPACT panel A)** CRISPR knock-down effects for each corresponding gene in the MSK-IMPACT panel. Only genes with at least 10 cell lines harboring splice-variants were included in the analysis. Significance was assessed using the two-sided Wilcoxon rank-sum-test. **B)** Sensitivity of compounds targeting each corresponding gene in the MSK-IMPACT panel (the target gene is written in parenthesis in each facet label). Only genes with at least 3 cell lines harboring splice-variants were included in the analysis. Significance was assessed using the two-sided Wilcoxon rank-sum-test.

**Discussion**

Here, we show that gene-level differential expression fails to capture the effects of mutations in the spliceosome that do have an impact on transcript-level expression. Most pre-processing pipelines upstream of differential expression analysis focus on gene-level quantification of reads at the gene-level because the accuracy is much better - transcripts contain less unique sequence than genes, which means that counting methods relying on alignment to a reference genome are less accurate when processing transcripts. Nowadays, there exist alignment-independent tools that are able to accurately estimate counts per transcript, but the reads still need to be aggregated to gene-level counts for most standard DE analysis tools such as DESeq2/edgeR. Similarly, most gene set analysis algorithms require gene-level p-values to work. Therefore, capturing the differences at the transcript level is not straightforward, especially when working with the processed data that is made available by many of the databases nowadays, namely DepMap or TCGA. Yet, here we show that aggregating transcript-level p-values into gene-level p-values using the Lancaster method makes it possible to go from expression data to a gene-level statistic that retains the differences present in the transcriptome while being compatible with many downstream gene expression analysis tools. To show that in fact there is a difference in the transcriptome composition for cell lines with intact spliceosome and for cell lines with mutated splicing factors, we calculated the Shannon Entropy (SE) for each transcript for both groups (spliceosome mutated and spliceosome WT). In accordance with previous research showing that alterations in the spliceosome correlate with higher entropy levels, we saw that mutated cell lines consistently return a higher SE value. Both analyses indicate that there are significant differences in the transcriptome composition due to mutations in the spliceosome, yet the changes caused by mutations affecting the spliceosome and splicing factors seem to be subtle and span across the whole transcriptome network rather than large and localized. Furthermore, our results show that mutations in the spliceosome translate to an altered composition of the expressed transcripts, especially increasing the expression of alternative transcripts compared with canonical ones. The gene set analysis using the aggregated Lancaster p-values revealed that several pathways with an important role in oncogenesis present significant alterations in transcript composition caused by mutations in the spliceosome. Among these, we decided to focus on the study of the differences in the genes part of the MYC targets gene set, since the gene set appeared enriched for transcripts that had significant differences in expression when calculating the GSA for core-spiceosomal proteins, non-core spliceosomal proteins and other splicing factors. Furthermore, MYC has been shown to be widely deregulated in cancer [21], [27], [28] yet it is hard to target therapeutically [29], [30]. When studying the most significantly differentially expressed genes in the gene set, we observed that some of them, such as RNPS1 and PABPC1, are part of the non-core splicing machinery but also have a role in the process of nonsense-mediated mRNA decay (NMD). NMD plays an extensive role in the regulation of gene expression, and its disruption has been shown to lead to many pathologies, including cancer [31], [32]. PABPC1 acts by binding to the poly(A) tail of mRNAs and, when this binding occurs at specific positions, it can repress NMD. Therefore, hyperactive or promiscuous isoforms of PABPC1 could hinder NMD and allow for aberrant transcripts to be expressed [31]. On the other hand, an up-regulated expression of factors involved in NMD may actually indicate that this process is more active than usual, which could be a side-effect of aberrant spliceosome machinery producing high amounts of aberrant transcripts. In such a case, inhibition of NMD may appear as a therapeutic strategy by increasing the load of aberrant transcripts in the cell and allowing for the expression of neoantigens in cancerous cells, which could enhance the efficiency of immunotherapy [32]. The role of NMD in cancer is context dependent [31], [33] and therefore a more detailed analysis is needed to understand in which cases targeting the NMD process could be beneficial for spliceosome-mutated cell lines. … Previous research has already shown that the spliceosome is a possible therapeutic vulnerability for certain cancers [34], [35] and some molecules, such as H3B-8800, have shown promise in clinical trials thus far [36].

The second part of our work focused on the prediction of new splicing variants using the deep learning tool SpliceAI. The performance of the algorithm is remarkable, especially considering that it has not been previously trained on cancer data. Furthermore, our estimation of the performance of the algorithm is limited by the available data from DepMap, which only reports mutations in the splice site when they occur at the consensus sequence and does not consider mutations elsewhere that may have an impact in the splicing process. Therefore, we are most likely underestimating the actual performance of SpliceAI.

Could we ask someone for raw RNA data and process it ourselves for a couple of cell lines?

More interestingly, the tool has shown to be able to predict variants outside of the genes where the mutation is located, proving its ability to use the context of the mutation to enhance its prediction power and revealing putative variants that were not detected through classical analysis of the genome such as whole genome sequencing (WGS) or similar techniques. …

To further assess the capability of SpliceAI to predict splice-altering variants, we studied the functional effect of the putative splice-altering mutations on cell growth - using the data from the Achilles project - and on drug sensitivity - using the PRISM repurposing dataset. The analysis of cell growth revealed that those genes carrying a mutation that alters splicing are rendered nonfunctional significantly more often than genes carrying other kinds of mutations, as evidenced by the lower variance of gene effect and the values closer to zero. This analysis also indicates that SpliceAI is able to predict splice-altering mutations, since it seems clear that at least some of the predicted mutations have a functional consequence and that their effect seems to be larger than other mutations annotated by DepMap. Furthermore, the genes that more often carry a predicted splice-altering variant (genes presenting a splice-altering mutation in more than 10 cell lines) are important tumor suppressor genes. This is not surprising, since the function of such genes is to prevent cells from becoming cancerous and hence it is common for malignant cells to present aberrant isoforms that allow for the maintenance of aberrant cellular rhythms [37]. Furthermore, hereditary cancer genes, such as NF1 and RB1, have been previously reported to accumulate splice-altering mutations [38]. Hence, the functional analysis indicates that SpliceAI is able to predict actual splice-altering mutations and that at least some of these have a functional effect in oncogenesis. ...

**Shannon Entropy (SE)**

The SE was calculated for each transcript using the transcription expression data downloaded from DepMap, as described above *(refer to SE section in Methods).* The results consistently show a significantly higher entropy for transcript expression in cell lines with mutations in the spliceosome. Higher entropy in cell lines with a mutated spliceosome may indicate a wider variety of transcripts being expressed in those cells than compared to “wild-type” cell lines. This is consistent with the results observed from the differential expression analysis at transcript level, which shows significant differences in the transcript expression between mutated and wild-type cell lines, and strengthens the idea that mutations in the spliceosome may cause subtle but network-wide changes that are not detectable by using classical approaches to measure differential expression.

**Materials & methods**

**Datasets**

All omics data and perturbation screens for the CCLE were freely accessible and downloaded from the DepMap portal [14]. For the somatic mutations, expression and Achilles CRISPR knock-down screening datasets, we used the 21Q1 release, made available in February 2021*.* The most recent drug sensitivity dataset available was the release 19Q4 (PRISM Repurposing Primary Screen), which was uploaded in December 2019.

**List of spliceosome & splicing factor genes**

The list of genes coding for elements of the spliceosome machinery and accssory splicing factors, a total of 404 genes, were extracted from previous work by Seiler and colleagues [36]. For the purpose of this study, we categorized each gene into “core” spliceosome, “non-core” spliceosome, or “other” in accordance with the labels produced in a prior study by Hegele and colleagues [39]**.** According to this study, genes labelled as “core” spliceosome genes were present in high abundance and have a well-defined function, or were known for being associated with some element of the spliceosome. Genes classified as “non-core”, on the other hand, include elements such as mRNA binding proteins and regulatory proteins.Some genes in the list do not belong to either category and have been labelled as “other” for the purpose of this analysis. In total, 141 genes are part of the core spliceosome, 103 are part of the non-core spliceosome, and 225 did not belong to any of these categories.

**Software & tools**

All local analysis, unless otherwise stated, was carried out in the R software package (version R 3.6.2; The R Project for Statistical Computing, [http://www.R-project.org](http://www.r-project.org)). To analyze the impact a mutation could have in the splicing process by modifying or creating splice donor or acceptor sites, we used the deep learning tool SpliceAI [15] (https://github.com/Illumina/SpliceAI). SpliceAI computations were performed in the HPC-cluster Hebbe, a high-performance computing environment allowing for computational job submissions using the slurm scheduler. The access to the cluster was enabled by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE).

**Code availability**

All scripts used to run the analysis can be accessed at the following Github repository: https://github.com/angelolimeta/SplicingInCancer

**Differential expression in the transcriptome**

In order to analyse the impact of a defective spliceosome in the transcriptional programs of the cell lines, we split the data in two groups: on one hand, cell lines harboring a mutation in **any** splicing factor, i.e. spliceosome deficient cell lines; and on the other hand, cell lines that retain an intact spliceosome, i.e. wild-type spliceosome cell lines (meaning that the genes encoding for splicing factors do not carry mutations for these cell lines).

We used the dataset *CCLE\_RNAseq\_transcripts.csv* downloaded from the CCLE/DepMap web portal. This file contains RNAseq generated expression levels, reported in transcript-per-million (TPM) data for each transcript at the individual isoform level. The transcript quantification was performed by the DepMap team using RSEM. The reporter expression levels were log2 transformed, using a pseudo-count of 1. Since the expression levels were reported using TPM, a format not suitable for further analysis using standard differential expression (DE) tools such as edgeR, we decided to test the relationship between the expression levels of the isoforms (log(TPM) data) and the presence of mutations in the splicing factors. In order to do this, we selected cancer cells from different cancer type that had at least 3 spliceosome deficient cell lines and 3 cell lines with a wild-type spliceosome. Then, we used a linear regression model to evaluate the relationship between expression levels of the isoforms and mutations in the spliceosome. To correct for multiple testing, we used the Benjamini-Hochberg (BH) procedure to adjust the p-values. Adjusted p-values less than 0.05 were considered significant. As a result, we obtained p-values for each transcript or isoform - i.e. “transcript-level p-values”.

**Lancaster method for the aggregation of transcript-level p-values**

In order to aggregate the transcript-level p-values into gene-level p-values that can be used to perform gene set analysis, while retaining the variations observed in the transcriptome when using a linear model, we used the Lancaster method. The Lancaster method was first suggested by Yi et al. (2018) [17] as a framework to translate transcript p-values to gene p-values. The method proposes a weighted aggregation of transcript-level p-values, weighting the p-value according to the expression level of the transcript. In order to weigh the transcript p-values generated for the linear model, we calculated the mean expression for each transcript for both cell lines with defects in the spliceosome and for wild-type cell lines. Then, using the Lancaster method, we generated the gene-level statistics.

**Gene-set analysis (GSA)**

To perform the GSA, we used the R package PIANO [19]. As input, we used the gene-level statistics derived from the linear model and the Lancaster aggregation: the unmodified (not corrected) aggregated p-values and the log fold-change, not including information about directionality. The GSA algorithm includes a step to perform the FDR correction, therefore it is not necessary to correct the p-values before feeding them to the algorithm. To define the gene sets to analyze, we downloaded several curated gene sets from KEGG [40] and the hallmark gene sets from MSigDB [41]. To perform the GSA, we used two statistical methods: Fisher’s test and Reporter features (for detailed information about the statistical methods, see Väremo et al (2018) [19]). To overcome the lack of directionality information that results from the Lancaster aggregation, we extracted the genes contained in some of the gene sets of interest and aggregated the transcripts according to whether they are considered canonical or alternative. The canonical transcript was extracted from the ENSEMBL database for each gene, using the suite BioMart [42]. The rest of the transcripts are therefore considered alternative transcripts.

**Shannon Entropy (SE) and Coefficient of Variation (CV)**

The Shannon Entropy and Coefficient of Variation are metrics that measure the variability of numerical data [43]. We used the R package EntropyExplorer [44] to calculate the differential SE and differential CV. We categorized the cell lines in cell lines with mutations in the spliceosome - mutated cell lines and cell lines with an intact spliceosome - wild-type cell lines. For this calculation, we considered mutations in any splicing factor or spliceosomal gene regardless of their categorization into “core”, “non-core” or “other”.

**Running & evaluating the performance of SpliceAI**

SpliceAI calculates the probability that a particular mutation will affect the splicing process. We used the somatic mutations dataset (*CCLE\_mutations.csv)* as input for the deep learning algorithm. In order to get a suitable format for SpliceAI, we parsed the dataset into a VCF file. The SpliceAI algorithm was run using default parameters. To estimate the performance of the algorithm, we measured its ability to classify the mutation in two classes: “splice-altering” or “not splice-altering”. SpliceAI outputs four probability scores for a mutation, evaluating the likelihood of the mutation event to alter either the donor or the acceptor site by deleting existing sites or creating new ones. Since we are only interested in knowing whether the algorithm considers a mutation to be splice-altering or not, only the highest of the four calculated scores is considered. According to Jaganathan et al. (2019) [15], a mutation can be confidently considered to alter splicing when the estimated score is above 0.5. Therefore, this is the score we used as a cutoff; considering putative splice-altering events all mutations for which SpliceAI gave a score of 0.5 or higher. Then, we used a Receiver Operating Characteristic (ROC) curve to evaluate the performance of SpliceAI, using the annotations made by DepMap in their somatic mutations file as true labels. In order to account for the imbalance of the dataset, we additionally evaluated the performance using precision-recall (PR) curves. For both ROC and PR curves we calculated the Area Under the Curve (AUC) to obtain a direct metric of performance. To calculate the ROC/PR curves and the AUC we used the R packages pROC [45] and PRROC [46].

**Characterizing the predicted variants**

For the distance to event analysis, i.e. the …

To know whether a predicted event would fall within an exon or intron, we used the transcript list from DepMap and filtered for all the transcript coming from genes carrying a somatic mutation. Then, we extracted the genomic positions of exon boundaries using BioMart databases [42]. Next, we used the package GenomicRanges [47] to create intervals for the exons and for the mutations predicted by SpliceAI (adding +1 to the position where splicing is predicted to be altered). The function findOverlaps detects when the alteration predicted by SpliceAI falls within one of the exons.

**Evaluating the functional impact of splice-altering variants using the MSK-IMPACT panel**

To better understand the effect of the splicing alteration in the phenotype of the cell line, we used the information contained in the CRISPR knock-out screening dataset and the PRISM drug sensitivity screening dataset. To narrow down the analysis, we used the list of clinically relevant somatic mutations generated by the MSK-IMPACT project [48] to only consider clinically relevant genes. The list of targeted genes was retrieved from the cBio Cancer Genomics Portal [49], [50], where it can be accessed under the name MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017). Since the goal of this analysis is to compare the effect of simple SNPs or other less disruptive mutations vs. SNPs that cause an alteration in splicing events, we decided to exclude insertions and deletions from the analysis, because such mutations are usually so disruptive that they often result in loss of function of the gene.

**Funding**

This work was funded by the Knut and Alice Wallenberg Foundation and Cancerfonden.

**Acknowledgements**

The authors would like to thank X & Y for critical review and input to the manuscript, as well as X, Y and Z for insightful discussions and feedback throughout this project.

**Author contributions**

All authors helped conceive and design the study. AL & LC analyzed the data. All authors wrote the paper and were involved in interpretation, editing, and discussion.

**References**

[1] Y. Barash *et al.*, “Deciphering the splicing code,” *Nature*, vol. 465, no. 7294, pp. 53–59, May 2010, doi: 10.1038/nature09000.

[2] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz, *Alternative splicing: A pivotal step between eukaryotic transcription and translation*, vol. 14. Nature Publishing Group, 2013. doi: 10.1038/nrm3525.

[3] G. S. Wang and T. A. Cooper, *Splicing in disease: Disruption of the splicing code and the decoding machinery*, vol. 8. Nature Publishing Group, 2007. doi: 10.1038/nrg2164.

[4] J. P. Venables, “Aberrant and Alternative Splicing in Cancer,” *Cancer Res.*, vol. 64, no. 21, pp. 7647–7654, Nov. 2004, doi: 10.1158/0008-5472.CAN-04-1910.

[5] S. C. Bonnal, I. López-Oreja, and J. Valcárcel, “Roles and mechanisms of alternative splicing in cancer — implications for care,” *Nat. Rev. Clin. Oncol.*, vol. 17, no. 8, pp. 457–474, Aug. 2020, doi: 10.1038/s41571-020-0350-x.

[6] S. Oltean and D. O. Bates, “Hallmarks of alternative splicing in cancer,” *Oncogene*, vol. 33, no. 46, pp. 5311–5318, Nov. 2014, doi: 10.1038/onc.2013.533.

[7] M. Seiler *et al.*, “H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers,” *Nat. Med.*, vol. 24, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/nm.4493.

[8] E. El Marabti and I. Younis, *The cancer spliceome: Reprograming of alternative splicing in cancer*, vol. 5. Frontiers Media S.A., 2018. doi: 10.3389/fmolb.2018.00080.

[9] J. P. Staley and J. L. Woolford, *Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines*, vol. 21. Elsevier Current Trends, 2009. doi: 10.1016/j.ceb.2009.01.003.

[10] M. C. Wahl, C. L. Will, and R. Lührmann, *The Spliceosome: Design Principles of a Dynamic RNP Machine*, vol. 136. Cell Press, 2009. doi: 10.1016/j.cell.2009.02.009.

[11] X.-D. Fu and M. Ares, “Context-dependent control of alternative splicing by RNA-binding proteins,” *Nat. Rev. Genet.*, vol. 15, no. 10, pp. 689–701, Oct. 2014, doi: 10.1038/nrg3778.

[12] S.-Y. Chen, C. Li, X. Jia, and S.-J. Lai, “Sequence and Evolutionary Features for the Alternatively Spliced Exons of Eukaryotic Genes,” *Int. J. Mol. Sci.*, vol. 20, no. 15, Art. no. 15, Jan. 2019, doi: 10.3390/ijms20153834.

[13] J. Li *et al.*, “Alternative splicing perturbation landscape identifies RNA binding proteins as potential therapeutic targets in cancer,” *Mol. Ther. - Nucleic Acids*, vol. 24, pp. 792–806, Jun. 2021, doi: 10.1016/j.omtn.2021.04.005.

[14] M. Ghandi *et al.*, “Next-generation characterization of the Cancer Cell Line Encyclopedia,” *Nature*, vol. 569, no. 7757, pp. 503–508, 2019, doi: 10.1038/s41586-019-1186-3.

[15] K. Jaganathan *et al.*, “Predicting Splicing from Primary Sequence with Deep Learning,” *Cell*, vol. 176, no. 3, pp. 535-548.e24, Jan. 2019, doi: 10.1016/j.cell.2018.12.015.

[16] T. A. Cooper, L. Wan, and G. Dreyfuss, “RNA and Disease,” *Cell*, vol. 136, no. 4, pp. 777–793, Feb. 2009, doi: 10.1016/j.cell.2009.02.011.

[17] L. Yi, H. Pimentel, N. L. Bray, and L. Pachter, “Gene-level differential analysis at transcript-level resolution,” *Genome Biol.*, vol. 19, no. 1, p. 53, Apr. 2018, doi: 10.1186/s13059-018-1419-z.

[18] H. O. Lancaster, “THE COMBINATION OF PROBABILITIES: AN APPLICATION OF ORTHONORMAL FUNCTIONS,” *Aust. J. Stat.*, vol. 3, no. 1, pp. 20–33, Apr. 1961, doi: 10.1111/j.1467-842X.1961.tb00058.x.

[19] L. Väremo, J. Nielsen, and I. Nookaew, “Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods,” *Nucleic Acids Res.*, vol. 41, no. 8, pp. 4378–4391, Apr. 2013, doi: 10.1093/nar/gkt111.

[20] C. M. Koh *et al.*, “MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis,” *Nature*, vol. 523, no. 7558, pp. 96–100, Jul. 2015, doi: 10.1038/nature14351.

[21] M. Gabay, Y. Li, and D. W. Felsher, “MYC Activation Is a Hallmark of Cancer Initiation and Maintenance,” *Cold Spring Harb. Perspect. Med.*, vol. 4, no. 6, Jun. 2014, doi: 10.1101/cshperspect.a014241.

[22] M. A. Karajannis and R. E. Ferner, “NEUROFIBROMATOSIS-RELATED TUMORS: EMERGING BIOLOGY AND THERAPIES,” *Curr. Opin. Pediatr.*, vol. 27, no. 1, pp. 26–33, Feb. 2015, doi: 10.1097/MOP.0000000000000169.

[23] C. Bazzichetto *et al.*, “PTEN as a Prognostic/Predictive Biomarker in Cancer: An Unfulfilled Promise?,” *Cancers*, vol. 11, no. 4, Mar. 2019, doi: 10.3390/cancers11040435.

[24] R. Di Fiore, A. D’Anneo, G. Tesoriere, and R. Vento, “RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis,” *J. Cell. Physiol.*, vol. 228, no. 8, pp. 1676–1687, Aug. 2013, doi: 10.1002/jcp.24329.

[25] R. Zhao, B. Y. Choi, M.-H. Lee, A. M. Bode, and Z. Dong, “Implications of Genetic and Epigenetic Alterations of CDKN2A (p16INK4a) in Cancer,” *EBioMedicine*, vol. 8, pp. 30–39, May 2016, doi: 10.1016/j.ebiom.2016.04.017.

[26] C. Yu *et al.*, “High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines,” *Nat. Biotechnol.*, vol. 34, no. 4, Art. no. 4, Apr. 2016, doi: 10.1038/nbt.3460.

[27] S. K. Madden, A. D. de Araujo, M. Gerhardt, D. P. Fairlie, and J. M. Mason, “Taking the Myc out of cancer: toward therapeutic strategies to directly inhibit c-Myc,” *Mol. Cancer*, vol. 20, no. 1, p. 3, Jan. 2021, doi: 10.1186/s12943-020-01291-6.

[28] C. Wang *et al.*, “Alternative approaches to target Myc for cancer treatment,” *Signal Transduct. Target. Ther.*, vol. 6, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41392-021-00500-y.

[29] H. Chen, H. Liu, and G. Qing, “Targeting oncogenic Myc as a strategy for cancer treatment,” *Signal Transduct. Target. Ther.*, vol. 3, no. 1, Art. no. 1, Feb. 2018, doi: 10.1038/s41392-018-0008-7.

[30] C. V. Dang, E. P. Reddy, K. M. Shokat, and L. Soucek, “Drugging the ‘undruggable’ cancer targets,” *Nat. Rev. Cancer*, vol. 17, no. 8, pp. 502–508, Aug. 2017, doi: 10.1038/nrc.2017.36.

[31] G. Nogueira, R. Fernandes, J. F. García-Moreno, and L. Romão, “Nonsense-mediated RNA decay and its bipolar function in cancer,” *Mol. Cancer*, vol. 20, no. 1, p. 72, Apr. 2021, doi: 10.1186/s12943-021-01364-0.

[32] F. Supek, B. Lehner, and R. G. H. Lindeboom, “To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases,” *Trends Genet.*, vol. 0, no. 0, Dec. 2020, doi: 10.1016/j.tig.2020.11.002.

[33] M. W. Popp and L. E. Maquat, “Nonsense-mediated mRNA Decay and Cancer,” *Curr. Opin. Genet. Dev.*, vol. 48, pp. 44–50, Feb. 2018, doi: 10.1016/j.gde.2017.10.007.

[34] T. Y.-T. Hsu *et al.*, “The spliceosome is a therapeutic vulnerability in MYC-driven cancer,” *Nature*, vol. 525, no. 7569, Art. no. 7569, Sep. 2015, doi: 10.1038/nature14985.

[35] B. Eymin, “Targeting the spliceosome machinery: A new therapeutic axis in cancer?,” *Biochem. Pharmacol.*, p. 114039, May 2020, doi: 10.1016/j.bcp.2020.114039.

[36] M. Seiler *et al.*, “Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types,” *Cell Rep.*, vol. 23, no. 1, pp. 282-296.e4, Apr. 2018, doi: 10.1016/j.celrep.2018.01.088.

[37] Y. Zhang, J. Qian, C. Gu, and Y. Yang, “Alternative splicing and cancer: a systematic review,” *Signal Transduct. Target. Ther.*, vol. 6, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41392-021-00486-7.

[38] C. L. Rhine *et al.*, “Hereditary cancer genes are highly susceptible to splicing mutations,” *PLOS Genet.*, vol. 14, no. 3, p. e1007231, Mar. 2018, doi: 10.1371/journal.pgen.1007231.

[39] A. Hegele *et al.*, “Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome,” *Mol. Cell*, vol. 45, no. 4, pp. 567–580, Feb. 2012, doi: 10.1016/j.molcel.2011.12.034.

[40] M. Kanehisa and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, vol. 28. Oxford University Press, 2000. doi: 10.1093/nar/28.1.27.

[41] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database Hallmark Gene Set Collection,” *Cell Syst.*, vol. 1, no. 6, pp. 417–425, Dec. 2015, doi: 10.1016/j.cels.2015.12.004.

[42] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt,” *Nat. Protoc.*, vol. 4, no. 8, pp. 1184–1191, 2009, doi: 10.1038/nprot.2009.97.

[43] K. Wang, C. A. Phillips, G. L. Rogers, F. Barrenas, M. Benson, and M. A. Langston, “Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes,” 3, 2014.

[44] K. Wang, C. A. Phillips, A. M. Saxton, and M. A. Langston, “EntropyExplorer: An R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression,” *BMC Res. Notes*, vol. 8, no. 1, p. 832, Dec. 2015, doi: 10.1186/s13104-015-1786-4.

[45] X. Robin *et al.*, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, Mar. 2011, doi: 10.1186/1471-2105-12-77.

[46] J. Keilwagen, I. Grosse, and J. Grau, “Area under Precision-Recall Curves for Weighted and Unweighted Data,” *PLOS ONE*, vol. 9, no. 3, p. e92209, Mar. 2014, doi: 10.1371/journal.pone.0092209.

[47] M. Lawrence *et al.*, “Software for Computing and Annotating Genomic Ranges,” *PLOS Comput. Biol.*, vol. 9, no. 8, p. e1003118, Aug. 2013, doi: 10.1371/journal.pcbi.1003118.

[48] A. Zehir *et al.*, “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients,” *Nat. Med.*, vol. 23, no. 6, Art. no. 6, Jun. 2017, doi: 10.1038/nm.4333.

[49] E. Cerami *et al.*, “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data,” *Cancer Discov.*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.

[50] J. Gao *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.