A systematic analysis of the effects of splicing on the diversity of post-translational modifications in protein isoforms

Sam Crowl[1],Maeve Bella Coleman[1], Andrew Chaphiv[1], Kristen M. Naegle[1,*]

1 University of Virginia, Department of Biomedical Engineering and the Center for Public Health Genomics, Charlottesville, VA, 22903

*** kmn4mj@virginia.edu**

Abstract

Post-translational modifications (PTMs) and splicing are known to be important regulatory processes for controlling protein function and activity. However, there have been limitations in analyzing the interplay of alternative splicing and PTMs, both from the standpoint of PTM presence and in the possible diversification of the regulatory windows of PTMs, which define the connection to regulatory enzymes and possible binding partners. Limitations stem from the deep differences in genomic and proteomic databases, where PTMs are predominantly identified by mass spectrometry and subsequently assigned to the canonical isoform of the protein in databases. In this work, we bridge the protein- and genome-centric world views to map PTMs to genomic locations for subsequent projection of PTMs onto alternative isoforms. We then perform a systematic analysis of the diversification of PTMs within all defined protein isoforms, focusing on the PTM-specific profiles that may differ across the various major modifications found in humans, including exploration of how often alternative splicing leads to diversification of the regulatory sequences directly flanking a PTM. We found the interplay between splicing and PTMs is PTM-specific across a range of behaviors, such as PTM inclusion rates across isoforms and tissues. Additionally we found that $\approx 2\%$ of prospective PTM sites exhibited altered regulatory sequences surrounding the modification site, suggesting that regulatory or binding interactions might be diversified in these proteoforms. In addition to exploring isoforms as defined by Ensembl, we applied this PTM-to-isoform mapping approach to explore the impacts of disease-related splicing in prostate cancer, identifying possible new hypotheses as to the variable mechanisms of ESRP1 expression in different cancers.

Introduction

Post-translational modifications (PTMs) are tightly regulated chemical modifications that play important roles in protein activity, localization, interactions, and degradation. Across the human proteome, there have been 356,949 experimentally observed post-translational modifications across 18,360 different proteins, encompassing 130 unique modification types, indicating that the vast majority of proteins are modified post-translationally in some way [1]. Most post-translational modifications are regulated by specific classes of enzymes that can either add or remove the modification on a specific residue. For example, phosphorylation is added by kinases and removed by phosphatases. Additionally, many post-translational modifications can trigger protein-interactions through PTM-specific recognition domains: bromodomains bind to acetyl lysines [2], WW domains and 14-3-3 proteins bind phosphorylated serines and threonines [3], and ubiquitin-binding domains bind ubiquitinated lysines [4] (Supplementary Figure 1). The affinity of enzymes and PTM-specific binding domains for their targets is largely driven by the flanking residues directly surrounding the modification site, a phenomenon often referred to as domain-motif interactions [5] [6]. For example, not all 90 tyrosine kinases within the

human proteome can phosphorylate all known phosphotyrosines. Instead, kinases are 'wired' to phosphorylate only certain sequences (called the enzyme or domain 'motif'), which allows different kinases to elicit different physiological effects in a cell. In this way, post-translational modifications allow for both specific and controlled regulation of protein function and interactions, and loss of a PTM site or change to the flanking sequence can have large implications on protein function and regulation.

Nearly all multi-exonic genes are expected to undergo alternative splicing to produce multiple unique protein isoforms from the same gene, suggesting the potential for PTMs to be modulated by alternative splicing [7] [8]. In fact, there are several examples of isoform-specific PTMs throughout literature that have resulted in distinct functional differences between isoforms, such as a glycosylation site found in an isoform of neuroligin-1 important for regulating its interaction with neurexin, altering synapse formation [9]. Globally, PTM sites, particularly those found within disordered regions of a protein, have been suggested to be enriched within tissue-specific exons (differentially expressed across tissues) [10] [11], although there is some work that disputes this idea [12]. Further, alternatively spliced regions of proteins are enriched for linear motifs corresponding to several PTM-specific binding domains important for faciliatating protein interaction and signaling networks, including SH2 and WW domains [13]. However, most proteomic experimental pipelines cannot capture different protein isoforms from the information in small peptide fragments. Hence, the majority of published experiments, which then get wrapped into PTM-databases, assign PTMs to a single reference protein, typically the "canonical" isoform [14]. As a result, pre-mRNA splicing and post-translational modifications are often thought of as distinct mechanisms of altering protein function and behavior both within and between tissues [15]. When considered together, it is typically based on a small subset of data (tissue-specific exons) [10] [11] [12], reliant on sparse UniProt annotations [13] [16], and/or looks at PTMs as one entity, limiting the broad utility of these results.

To overcome the limitations of the proteomic annotations of PTMs and assess the relationship with splicing, we developed a reproducible computational pipeline to merge proteome-level annotations of post-translational modifications with genome-level annotations of alternative transcripts, identifying "prospective" PTM sites in alternative isoforms. To the best of our knowledge, this is the most extensive global analysis of the relationship between splicing and post-translational modifications to date, and the first to consider different types of modifications separately. We predominantly focused on two mechanisms by which alternative splicing can modulate PTMs – exclusion of the modification site from an isoform or alteration of the short linear motif surrounding the modification site. We illustrated that splicing plays a significant role in regulating the presence of PTM sites in alternative isoforms ($\approx 30\%$ of PTMs are excluded from at least one isoform), but that certain modification types are more likely to be maintained across different isoforms, such as glycosylation. Further, we found that a subset of prospective PTM sites ($\approx 2\%$, amounting to 10,460 prospective PTM sites) exhibit altered flanking sequences that may serve to change the regulatory enzymes responsible for the modification, as well as the binding domains that interact with that site.

Finally, having performed an exhaustive exploration of all possible human isoforms defined by Ensembl, we wished to understand how considering PTM regulation can be informative in the analysis of isoform expression differences within human tissues. As a case study, we applied a PTM-focused analysis of splicing patterns in prostate cancer patients with increased ESRP1-expression using data from The Cancer Genome Atlas (TCGA), as this splicing factor has been shown to be detrimental to prostate patient outcomes. We identified a possible signaling mechanism that explains ESRP1's effect on patient prognosis by altering the negative regulation of mTORC1. From this work, we have highlighted the importance of including post-translational modifications in the analysis of splicing and its impact on cell phenotype, an area that is not commonly discussed in splicing literature.

Results

In order to perform a systematic analysis of when and how alternative splicing diversifies PTMs across protein isoforms, we first needed to expand our knowledge of the post translational modifications present in alternative isoforms. The majority of post-translational modifications in modern-day PTM compendia

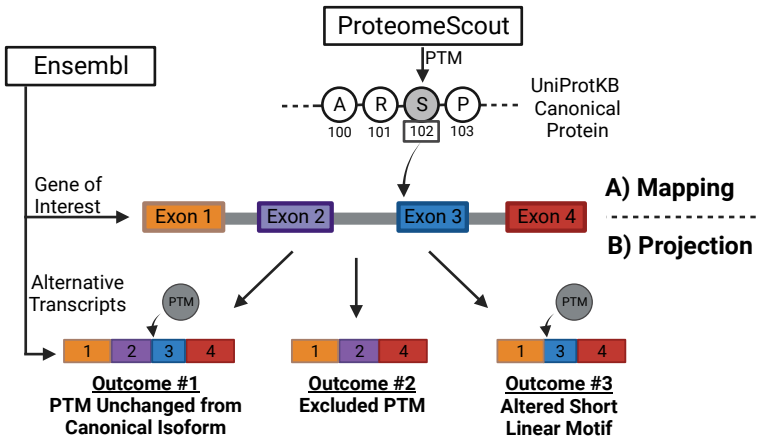


Figure 1. Mapping and Projecting Post-Translational Modification (PTMs) onto Alternative Transcripts. In order to expand the knowledge of PTMs in alternative isoforms, we mapped PTMs from a comprehensive database of post translational modifications (ProteomeScout) onto the genomic coordinates of Ensembl **(A)**, then project those genomic locations onto Ensembl-defined protein coding transcripts **(B)**. We refer to these projected PTMs as "prospective" PTMs in alternative isoforms. We defined three key outcomes of splicing on PTMs based on the predominant events observed in our analysis (PTM denoted as the gray circle on the third exon in the canonical transcript): **(1)** the PTM was successfully mapped onto the alternative isoform, with no changes to its surrounding region, **(2)** the PTM was excluded from the alternative isoform, or **(3)** the PTM was successfully projected onto the alternative isoform, but residues surrounding the PTM site were altered, which may influence the specificity of enzymes and binding partners.

were discovered by mass spectrometry-based approaches, which typically rely on trypsin-based proteolysis to yield MS-amenable peptide fragments. PTM annotations coming from tryptic-fragment identification are usually assigned to the canonical isoform, unless the tryptic fragment can be uniquely assigned to an alternative isoform (Supplementary Figure 2). In fact, 81% of alternative isoforms have no recorded PTMs, despite the fact that a peptide containing a modified residue could be mapped to more than one possible isoform. In order to overcome this limitation, we integrated three key resources using a Python-based framework. We used genomic information from Ensembl to provide the set of all human protein-coding transcripts supported by experimental evidence and the genomic location of their exons [17]. We used the protein database UniProt to establish the list of unique human protein identifiers, the canonical isoform of each human gene, and any reviewed alternative protein isoforms [18]. We used ProteomeScout as the resource for post-translational modifications, since it aggregates PTMs from five major compendia, creating a comprehensive and integrated resource of PTMs [1]. Using these resources, we ultimately sought to translate between the fields of genomics and proteomics for a comprehensive analysis of the control of PTMs by splicing, avoiding aggregating PTMs into a single class of information and focusing on modifications that have at least 150 unique observations in the human proteome. Importantly, we developed this framework for reusability and new analyses as our knowledge of both splicing and post-translational modifications continues to improve, with both the data and code associated with this work intended for use by the broader scientific community.

Projecting Prospective Post-translational Modification Sites onto Alternative Isoforms

To expand the data available for analysis, we mapped post-translational modification sites in canonical protein isoforms to their location in their corresponding exon, transcript and gene recorded in Ensembl. With the genomic location of each PTM, we then project the PTM onto all alternative transcripts in Ensembl, allowing us to account for whether the final transcript includes some or all of the

PTM-carrying exon (Fig. 1, Supplementary Figure 3). We refer to the PTM sites projected onto alternative transcripts as "prospective", indicating that the genomic location is known to code for a modifiable residue in the canonical isoform, though there may not be direct annotated evidence of the modification existing in the alternative isoform. Using this pipeline, we successfully expanded the annotations of PTMs on alternative isoforms – 96.2% of prospective PTMs were not previously annotated. Of the post-translational modifications recorded in UniProt isoforms, we were able to capture 79.8% of the PTMs through this process. Of the ≈ 20% of PTMs not captured, some of these are cases in which the PTM has only been annotated in the alternative isoform (about two-thirds, or 3,322 PTMs), while the rest are unique to the alternative isoform (1,616 PTMs). However, these missed PTMs comprise less than 1% of the total PTM sites found in alternative isoforms, indicating that this approach greatly expands the amount of available PTM information while minimizing information loss.

Given the successful projection of PTMs onto alternative isoforms, we next assessed the effect of splicing on PTMs based on a set of key, possible effects (also described in Fig. 1). First, a PTM may be included in an isoform, maintaining any potential interactions that have been shown for the canonical isoform. Second, a PTM could be excluded from an isoform altogether, removing any potential functional changes or interactions induced by the modification. Lastly, if a prospective PTM is included in the alternative isoform, it may contain a unique flanking sequence directly surrounding the PTM distinct from the canonical isoform, potentially leading to different interactions with regulatory enzymes or PTM-specific binding domains. The effect of these changes was also annotated with the splicing event that resulted in the change relative to the canonical isoform, such as a skipped exon, alternative splice site, or mutually exclusive exons. The entirety of the datasets, prospective PTMs, and regulation through splicing are freely available to the research community (see data availability secton for details).

PTMs are differentially included through alternative splicing

We sought to evaluate how splicing might impact the presence of PTMs in alternative isoforms across the entire transcriptome and whether the relationship between splicing and PTMs is dependent on the specific modification type. To address this question, we first considered the number of PTMs within Ensembl-defined constitutive exons – exons found in every transcript associated with a gene. We found constitutive exons exhibited a lower density of PTM sites than non-constitutive exons, with the most densely populated exons tending to be non-constitutive (Supplementary Figure 4). This matches previous reports comparing tissue-specific exons to constitutive exons measured from RNA sequencing data and provides further evidence that splicing plays a role in the inclusion of PTM sites in protein isoforms [10] [11]. To identify which PTMs in particular are or are not being regulated by splicing, we identified "constitutive PTMs", or PTMs that are included in all protein-coding transcripts of a gene. Overall, 69.6% of PTM sites could be defined as constitutive across Ensembl transcripts and were unaffected by alternative splicing. However, not all types of PTMs are controlled by splicing at the same rate, with constitutive rates ranging from 49.4% (Palmitoylation, 168 unique observations) to 86.8% (sulfation, 152 unique observations) (Fig. 2A,B, Supplementary Figure 5). PTMs found within structural domains were slightly more likely to be included in isoforms, but this alone failed to explain the variation across modification types (Supplementary Figure 6). Further, when compared against a null model where splicing is entirely random and non-specific to modifiable residues, we found that much of the observed variability cannot be explained by randomness alone (Fig. 2B). Phosphorylation, methylation, sumolyation, hydroxylation and palmitoylation modifications all exhibited a lower constitutive rate than expected based on the null model. On the other hand, ubiquitination, glycosylation, and sulfation were much more likely to be found in alternative isoforms than in the null models (Fig. 2B).

Given the broader classes of PTMs demonstrated variability, we also assessed the subtypes within each modification type, and found further variability in splicing regulation (Supplementary Figure 7). For example, glycosylation as whole exhibits a higher constitutive rate than most other modifications (76.3%). This is largely due to N-linked glycosylation (75.3%), while O-linked glycosylation has a constitutive rate closer to the average across all modifications (69.9%). Similarly, phosphoserines exhibited slightly lower constitutive rates than either phosphothreonine or phosphotyrosine sites. This indicates that there is variability in splicing regulation even amongst similar modification types.

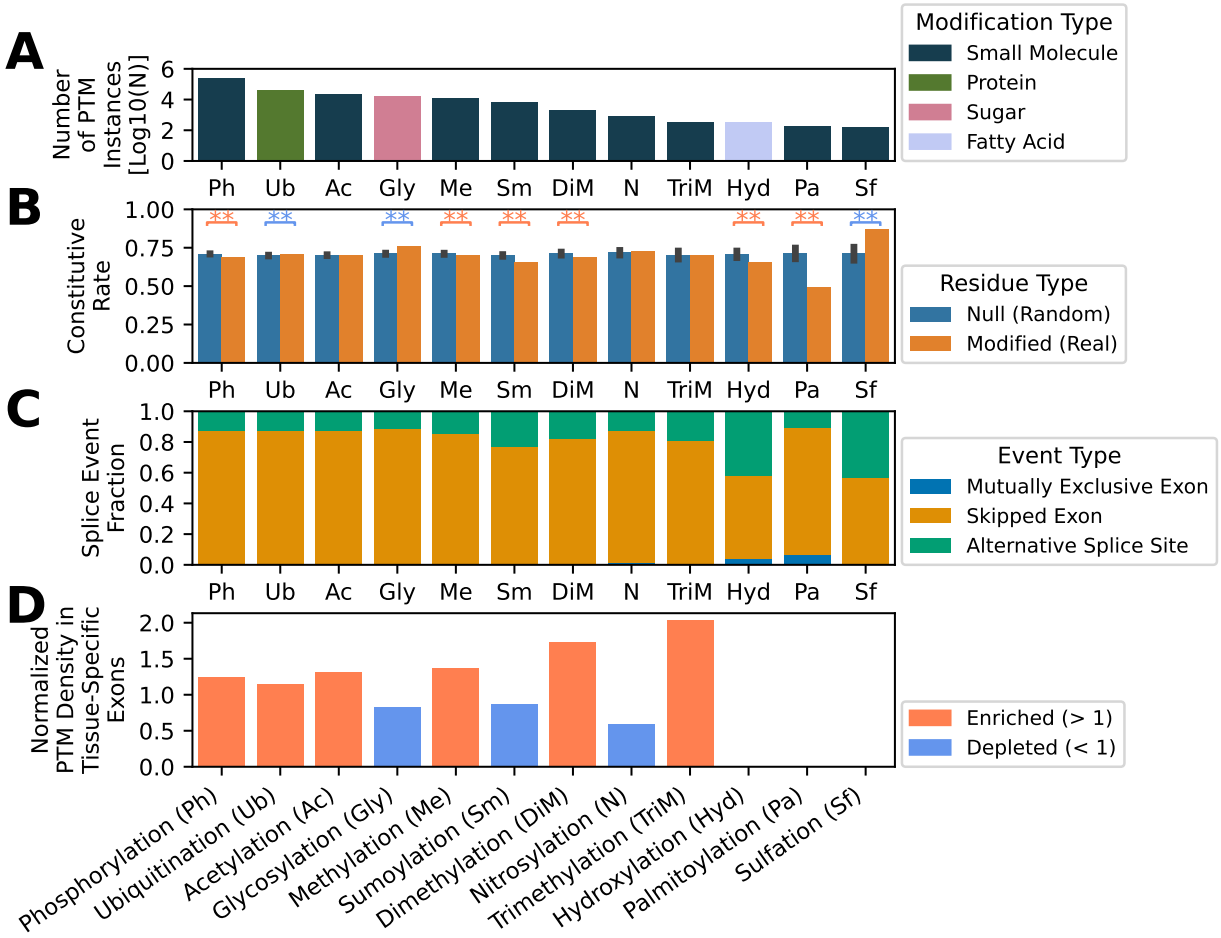


Figure 2. Splicing controls the presence of post-translational modifications in alternative isoforms. Modification-specific statistics of PTM inclusion across human alternative transcripts. Each panel shares the same x-axis, sorted from the largest number of identified modifications (phosphorylation) to the least (sulfation). Here, we only show modifications with at least 150 instances across the human proteome. For data of all modification types, see Supplementary Figure 5. **A)** The number of modifications documented in canonical UniProt proteins for each class of post-translational modification. Each bar is colored based on what type of molecule is being added: small molecules, proteins, sugars, or fatty acids. **B)** The fraction of modification sites that were found across all isoforms of a given gene, defined as a constitutive PTM rate. We compared the results for the real PTM data to a null model generated by randomly selecting residues across the proteome in matching proportions and treating them as real PTM sites. **C)** Breakdown of the types of alternative splicing events that result in the loss of a PTM site relative to the canonical isoform, considering skipped exon events, alternative splice sites (ASS), and mutually exclusive exons (MXE). Alternative splice sites are defined as described in Wang et al. and elsewhere [7]. Mutually exclusive exons were identified based on criteria from Pillmann et al. [19]. **D)** Density of each type of PTM in tissue-specific exons identified across three different publications [10] [20] [21]. PTM density is defined as the number of PTMs per 1000 amino acids and has been normalized by the density of that same PTM across the entire proteome (regardless of tissue-specific expression). A normalized density greater than 1 indicates more PTMs than expected within tissue-specific exons, a normalized density less than 1 indicates fewer PTMs than expected.

As many Ensembl transcripts have not yet been shown to produce functional proteins, it is possible that these results are influenced by transcripts that do not ultimately produce functional proteins. To account for this possibility, we repeated the same PTM-specific analysis using different definitions of functional transcripts, including functionality predictions from TRIFID [22]. As we implemented more restrictive criteria, we observed a decrease in the number of isoforms per protein and an increase in the constitutive PTM rate, suggesting that the true constitutive rate is likely somewhere between 70-90% (Supplementary Figure 8). Regardless of functional criteria used, the rates continued to retain similar variability across most modification types, suggesting that control of PTMs by splicing is PTM-specific and PTMs should always be evaluated in a detailed, PTM-specific manner, even within the same modification class – such as seen in O- vs. N-linked glycosylation and serine vs tyrosine phosphorylation.

To determine the potential functional relevance of splicing-regulated PTMs, we obtained molecular functions and biological processes associated with PTMs from PhosphoSitePlus [23] and tested for enriched annotations within constitutive and non-constitutive PTMs. PTMs controlled by splicing (non-constitutive) were found to be significantly associated with many different functions that are important for regulating protein interactions and activity, including PTM sites important for intracellular localization and regulating molecular associations, indicating that alternative splicing may rewire protein-interaction networks through exclusion of the PTM sites driving these protein interactions (Supplementary Figure 9, Supplementary Table 1). Constitutive PTMs, on the other hand, were only found to be enriched for sites regulating receptor desensitization, and a larger than expected percentage of constitutive PTM sites did not have a known function. Similarly, constitutive PTM sites were not found to be enriched for any biological processes, while non-constitutive PTMs were enriched for a variety of processes including cell growth and gene transcription (Supplementary Figure 9). Combined, this points to the idea that splicing control of PTMs across isoforms will impact protein function and the regulatory networks associated with the spliced protein.

Next, we wished to understand if the way in which PTMs were excluded had a relationship with the types of splicing that can occur – such as by skipped exons, alternative splice sites, or mutually exclusive exons. While the rate of splicing control of PTMs was modification-specific, most modifications were impacted by different splice events in similar proportions, with most excluded PTMs being a result of a skipped exon event. There were a few small exceptions, as hydroxylation and sulfation sites were lost due to an alternative splice site more often than other modifications (Fig. 2C). In rare cases, mutually exclusive exon events may also impact the presence of PTMs. We identified 370 mutually exclusive exon events associated with PTM sites, using criteria defined previously by Pillman et al. [19]. These events encompass 142 different PTM sites and 14 modification types (Supplementary Figure 10). Based on global alignment of mutually exclusive exon pairs, we found that PTM sites were likely to be conserved in only 39.7% of cases, perhaps indicating that mutually exclusive exon events may be a mechanism by which PTM sites can be removed from alternative isoforms without significantly disrupting the overall structure of a protein.

While these results suggest a potential role of splicing in controlling PTM inclusion, we did not consider expression of transcripts within and across tissues, instead looking at all protein-coding alternative transcripts that have been recorded in Ensembl. To assess differential inclusion of PTM sites across tissues, we compiled a set of tissue-specific exons identified across three different publications [10] [21] [20]. Overall, PTMs tended to be found at a higher density in tissue specific exons than across the entire proteome, in line with prior results [10] [11]. As when looking at the entire transcriptome, PTM density in tissue-specific exons was largely dependent on the modification type. Glycosylation and nitrosylation exhibit lower density within tissue-specific exons than found across the entire proteome, matching their higher constitutive rate (Fig. 2D). Interestingly, sumoylation and palmitoylation, which were observed to have low constitutive rates, were not found to be abundant within these tissue-specific exons. Although there are some discrepancies between PTM density within tissue-specific exons and constitutive rates across the transcriptome for certain modification types, these results confirm that alternative splicing is an important mechanism for modulating the presence of certain PTMs, in particular those involving phosphorylation and methylation sites.

Alternative Splicing affects PTM-regulatory sequences

PTMs are important components of signaling pathways, cellular communication, and protein interactions, all of which are facilitated by direct interactions with a PTM by its corresponding modifying enzyme(s) and recognition domain(s). The specificity of these interactions are driven by the residues immediately upstream and downstream of the modification site, with the linear motif directly surrounding the PTM being a prerequisite for recognition of the PTM site [5] [6] [24]. Given this, we hypothesized that splicing may alter PTM function by changing the flanking sequence surrounding the PTM site. There are several mechanisms by which an alternative splice event could lead to changes in the flanking sequence of a PTM, including a skipped/retained exon adjacent to the PTM, an alternative splice site in a nearby exon or in the PTM-containing exon, or a mutually exclusive exon event (Fig. 3A). With this in mind, we sought to characterize how often alternative splicing leads to changes in the flanking sequence surrounding PTM sites, defining the flanking sequence as the five amino acids on either side of the PTM, unless otherwise stated. Although the specific residues important for PTM recognition are variable, depending on the modification and interacting protein, a regulatory window size of five residues likely captures the most common residues important for facilitating interactions across most modifications/proteins.

When comparing flanking sequences of a PTM site in the canonical and corresponding alternative isoforms, we found that 2.04% of prospective PTM sites observed changes to the flanking residues (Supplementary Table 2). Importantly, 25.1% of these sites have a matching tryptic fragment compared to the canonical tryptic fragment, indicating that 0.52% of prospective PTM sites have altered regulatory regions that would not be detected by trypsin-mediated MS (Fig. 3B). Across the possible altered regulatory regions, we observed that acetylation, trimethylation, and palmitoylation exhibited the highest rate of alteration (Fig. 3C). When compared against the null model, we found that the high alteration rate could not be explained by randomness alone (Supplementary Figure 11). In the majority of cases, we observed that altered regulatory windows are caused by changes to the adjacent exon, rather than the PTM-containing exon itself, usually via a skipped/retained exon (Fig. 3D). Both glycosylation and nitrosylation, which also had higher constitutive rates, have lower rates of altered flanking sequences, which were at rates that could be attributed to the distribution of amino acids across the transcriptome (Supplementary Figure 11).

We further explored the residue positions in the flanking sequences that were most affected by splicing and the potential consequence of these changes. Due to the nature of splicing-induced changes, most regulatory regions are altered predominantly only on one side of the modifications flanking sequence, and the farther you move from the PTM the more likely it is to see changes to the regulatory region (Supplementary Figure 11). Among the $\approx 2\%$ prospective PTMs with altered linear motifs, 17% of these PTMs (1,824 prospective PTMs) have a change in the +/-1 residue, immediately adjacent to the PTM site, which has the potential to significantly impact enzyme or binding domain interaction (Fig 3C, Supplementary Figure 11). For example, proline-directed kinases are a large subset of serine/threonine kinases with a strong preference for a proline in the +1 position, which includes cyclin dependent kinases (CDKs) and ERK [24]. Similarly, the CDY family of proteins contain a chromodomain that binds to methylated lysines with a strong preference for lysines that immediately follow an arginine [25]. In both cases, the remaining residues in the regulatory window serve to further drive specificity and the strength of the interaction, with more minor effects. We further confirmed that splicing-induced regulatory sequence changes typically preserve information in the motif, but create smaller changes. Regardless of the regulatory window size chosen (3, 5, 7, or 10 amino acids on either side of the PTM), regulatory sequences generally still maintained $> 60\%$ sequence identity (Fig. 3E). Altogether, these results indicate that splicing can alter regulatory sequences in ways that could impact PTM-specificity and recognition, but suggests that it is unlikely to fully disrupt interactions due to conservation of key residues nearby the PTM and that prospective PTMs with altered regulatory specificity windows may still be targets of modification in those alternative isoforms.

We next sought to understand how changes in the regulatory sequences might affect PTM regulation and protein interactions. Given that there is minimal data regarding isoform-specific binding or enzymatic activity, we compared the predicted affinity of various serine/threonine kinases for both canonical sequences and alternative regulatory sequences using a kinase motif enrichment tool, Kinase

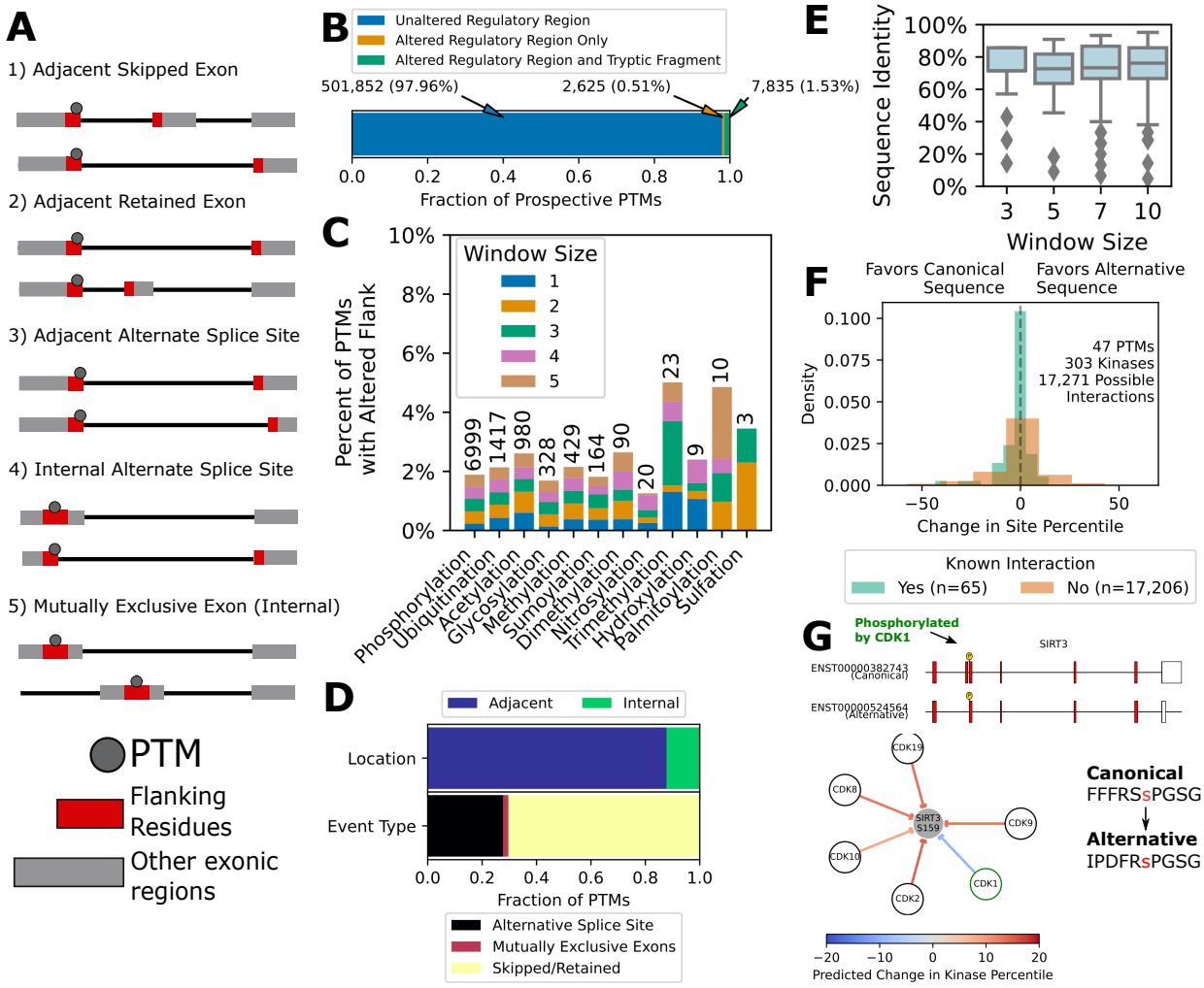


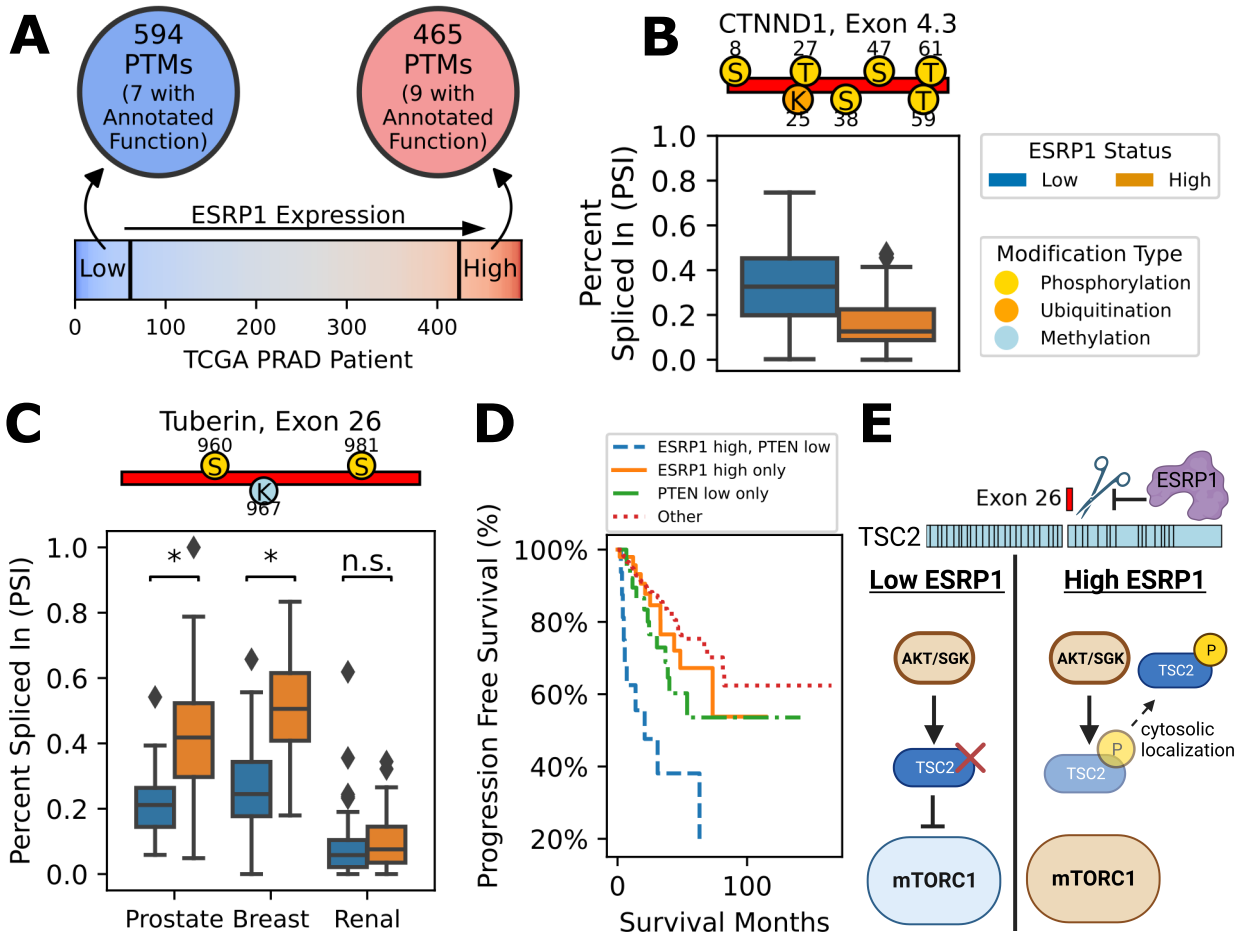
Figure 3. Alternative splicing rewires PTM-driven interaction networks by altering flanking sequences. A) Illustrative examples of the modes of alternative splicing that result in changes to the flanking sequence around a PTM, comparing two transcripts with differing flanking sequences (highlighted in red) surrounding the same PTM. **B)** Number of prospective PTMs found in alternative isoforms with an unchanged flanking sequence (+/- 5 amino acids), an altered flanking sequence but unchanged tryptic fragment, or both an altered flanking sequence and tryptic fragment. **C)** Fraction of prospective PTMs with altered flanking sequences, broken down by modification type, with the total number of events indicated above each bar. The proximity of the change to the PTM is denoted by the different bar colors, ranging from immediately next to (1) or up to 5 amino acids away from the PTM on either side. **D)** Breakdown of which events described in panel A most commonly lead to altered regulatory regions. **E)** Sequence identity between differing flanking sequences in the canonical and alternative isoforms, using different window sizes of amino acids on either side of the PTM. **F)** Change in the likelihood of kinase interaction from the canonical to the alternative flanking sequence, based on predicted site percentiles derived from Kinase Library software [24]. We only considered PTM sites with an altered flanking sequence and conserved tryptic fragment (orange bar in panel B) that also had at least one known kinase interaction annotated in PhosphoSitePlus [23] **G)** An example of an altered flanking sequence surrounding the phosphorylation site at S159 in SIRT3. The change in Kinase Library site percentile between the canonical and alternative isoform of SIRT3 for various cyclin-dependent kinases are shown, with positive changes (red arrows) indicating a preference for the alternative isoform, while negative changes (blue arrows) indicating a preference for the canonical isoform.

Library [24]. We focused on serine and threonine phosphorylation sites with a known kinase interaction, an altered flanking sequence, and a conserved tryptic fragment, as these are the modifications that are most likely to be missed by trypsin-mediated MS and be misannotated on the canonical isoform of the protein. In total, we looked at 57 different events. We used the Kinase Library tool to predict the likelihood of the annotated kinase interaction (kinase annotated as being known to phosphorylate the target) with either the canonical or alternative regulatory sequence, and then compared changes in the Kinase Library percentile for all kinases in the serine/threonine kinome. Interestingly, in many cases, flanking sequence changes had a wide range of possible impacts on both the annotated kinase interaction and on kinases not annotated as a regulatory kinase: decreased likelihood in some cases, increased it in others, and in an equal number of cases it had no effect on the predicted likelihood (Fig. 3F). Kinases that are not known to interact with the particular phosphorylation site exhibited more extreme cases than the annotated, known kinase interaction(s). In some cases, the predicted ranking for unannotated kinases differed by more than 100 ranks between the canonical and alternative isoform, while predicted changes in the annotated interaction tended to be more subtle. However, there are also cases where the annotated kinase's predicted interaction significantly differs when looking at the alternative isoform. For example, SIRT3, a deacetylase that regulates energy metabolism, contains an activating phosphorylation site at S159 that has been shown to be phosphorylated by CDK1. In an alternative isoform of SIRT3, a skipped exon event results in a unique regulatory sequence preceding S159 (Fig. 3G). While the proline in the +1 position remains present, a key feature of proline-directed kinases like CDKs, the surrounding changes result in a decreased likelihood of phosphorylation by CDK1 while increasing affinity for other CDK family members like CDK2, suggesting a potential difference in the regulation of phosphorylation at S159, moving away from CDK1 in the alternative isoform (Fig. 3G). It appears that splicing-induced changes to regulatory sequences causes subtle changes in PTM-specificity and, in some cases, leads to altered regulatory networks. Although limited by constraints on known enzyme interactions for extending confidently to more phosphorylation sites or interaction changes with other differentially spliced PTMs, this analysis indicates that altered flanking sequences of PTMs introduced by splicing has the capacity to diversify regulatory networks of a protein.

Overexpression of splicing factor, ESRP1, leads to tumor-specific post-translational modifications that drive cell growth and DNA damage response in prostate cancer

Having assessed the global relationship between all PTMs and Ensembl-annotated human transcripts and their isoforms, we wished to understand if PTM projection could be useful to understanding alternative PTM networks in disease. Aberrant splicing is involved in several diseases, including cystic fibrosis, muscular dystrophy, and cancer [26] [27] [28] [29] [30]. In our global analysis, we found that non-constitutive PTM sites were enriched for sites important to carcinogenesis, suggesting that changes to splicing patterns may influence the presence of functionally important PTM sites for the development and progression of cancer (Supplementary Figure 9). Further, many splicing factors important for regulating splicing of pre-mRNA have been implicated in altering cancer progression, such as epithelial splicing regulatory proteins (ESRPs), splicing factors responsible for the splicing of epithelial-specific isoforms [31]. In prostate cancer, it has been reported that elevated levels of ESRP1 leads to worsened prognosis [32] [33] [30]. For this reason, we selected prostate cancer as a system to study the impact of differential inclusion of PTMs as a mechanism for disease progression.

We sought to identify whether ESRP1 expression resulted in differential inclusion of specific PTM sites, which may contribute to the difference in phenotype and worsened prognosis in prostate cancer. We used The Cancer Genome Atlas (TCGA) SpliceSeq database to look at exon utilization in ESRP1-low versus ESRP1-high prostate cancer patient samples. This dataset allows us to look at exon utilization across samples, defined by the percent spliced in (PSI) rates of each exon, and by projection, the PTMs that might exist on those exons. Given that this data is exon-level information, we focused on exclusion of PTMs rather than changes to their regulatory sequences as that data is lacking from this resource. While most exons and projected PTM sites were unaffected by changes in ESRP1 expression,



ESRP1 drives phenotypic changes in cancer by altering the presence of specific PTM sites. A) To assess the role of ESRP1 in prostate cancer progression, we explored the relationship between ESRP1 expression and the inclusion of PTM sites in prostate tumor samples from The Cancer Genome Atlas (TCGA), based on RNA expression levels of ESRP1 and percent spliced in (PSI) data from SpliceSeq [34] [35]. Patients were binned into two groups: ESRP1-high ($z \geq 1$) and ESRP1-low ($z \leq -1$). In total, 1059 PTMs were found to be differentially regulated (indicated by large circles), of which only 16 were associated with a specific biological process in PhosphoSitePlus [23]. **B)** An example of a known splicing target of ESRP1 captured by this process, an exon of CTNND1 containing several phosphorylation sites and one ubiquitination site. **C)** Inclusion of exon 26 of tuberin (TSC2) across multiple cancer types, which contains a phosphorylation site at S981 important for localizing TSC2 to the cytosol and away from mTORC1. ESRP1 plays in an oncogenic role in prostate and breast cancer, but has been suggested to play a protective role in renal cancers. **D)** Progression-free survival of prostate cancer patients with high expression of ESRP1, low expression of PTEN, both, or neither. **E)** Potential consequence of splicing of TSC2 Exon 26 and its relationship to the PI3K/mTOR pathway. Created with Biorender.com.

we identified a total of 1059 PTM sites that were differentially included based on ESRP1 expression, with 594 and 465 PTM sites specific to either the ESRP1-low and ESRP1-high groups, respectively (Fig. 4A, Supplementary Table 3). Phosphorylation makes up the the majority of differentially included modifications, but we also identified several acetylation, ubiquitination, and methylation sites, among others (Supplementary Figure 13). Genes with regulated PTM sites were enriched for those found at the cytoskeleton and involved in cell adhesion, matching ESRP1's known association with the EMT transition and changes to cell motility (Supplementary Figure 12). In addition, it includes genes and exons that have previously been identified as ESRP1-regulated, such as an exon in CTNND1 containing several phosphorylation sites and an ubiquitination site (Fig. 4B) [36]. All together, this suggests that this approach successfully captures at least some of ESRP1-mediated splicing in patient samples.

In order to better understand the role of ESRP1-mediated splicing on protein function, we delved deeper into the function of the differentially included PTM sites, based on annotations in PhosphoSitePlus [23]. Of the 1059 ESRP1-regulated PTMs, we were only able to identify 16 that were known to be relevant to a specific biological process (Fig 4A). Among these, PTMs relevant to carcinogenesis were among the most enriched biological processes (Supplementary Figure 13). However, somewhat surprisingly, of the six PTMs identified as relevant to carcinogenesis, inclusion values of five of these sites would suggest increased carcinogenesis potential in ESRP1-low patients, contrary to the worsened prognosis of patients with ESRP1 amplification. When studying ESRP1's role in cancer progression, the majority of available literature focuses on its role in EMT, with low expression of ESRP1 leading to a more mesenchymal and invasive state [37] [38] [31] [39]. However, based on PTM-inclusion, there is not a clear upregulation of PTMs known to drive increased motility in ESRP1-low patients (Supplementary Figure 14). It actually has been shown in breast cancer, in which ESRP1 is also oncogenic, that ESRP1 downregulation does not induce a more mesenchymal state despite changes in expression to EMT-related genes, supporting the idea that ESRP1-driven EMT may be tissue-specific [40]. In addition to EMT, studies have linked ESRP1 expression to oncogenic autophagy, primarily as a result of splicing of FGFR2 [41]. However, we observed that ESRP1 was correlated with inclusion of Tuberin S981 (inhibits autophagy when phosphorylated) and anticorrelated with inclusion of ATG16L1 S278 (induces autophagy when phosphorylated), suggesting that ESRP1-high tumors might be less likely to undergo autophagic processes as a result of differential inclusion of PTMs (Supplementary Figure 14). Given these results, we propose that ESRP1's oncogenic capabilities in prostate cancer may be due to mechanisms not previously identified.

Previous literature has identified a strong link between increased ESRP1 expression, high copy numbers, and chromosomal deletions [32] [30]. To validate this in the TCGA cohort, we compared ESRP1 RNA expression with the fraction of genome altered and found a strong correlation, independent of TP53 alterations (data not shown). While the mechanism by which ESRP1 leads to chromosomal aberrations is still unclear, it has been hypothesized this is a result of dedifferentiation and genomic instability, rather than a direct role in DNA repair [32]. In support of this idea, several transcription and splice factors were found to be regulated by ESRP1, with phosphorylation sites at S199 in heterogenous nuclear ribonucleoprotein A1 (HRNPA1) and S213/S272 in Poly(Rc)-binding protein 2 (PCBP2) being important for regulating differentiation in various cell types. Inclusion of these sites in ESRP1-high tumors suggest that differentiation processes are more likely to be inhibited and exist in a dedifferentiated state. This combined with splicing of genes important for cell growth and cell cycle regulation, including CD44 and CDK1, supports the idea that increase in genomic alterations is due to genomic instability. We also found several p53 target genes that are differentially regulated by ESRP1, including CDK1, CASP8, DDB2, and TSC2. Tuberin (TSC2) is a particularly interesting case, as TSC2 is a negative regulator of mTOR signaling and loss of TSC2 has been shown to lead to aberrant S-phase progression upon DNA damage and increased sensitivity to stress, in large part due to sustained mTOR signaling [42] [43] [44]. In addition to loss of expression, TSC2 inhibition of mTORC1 signaling can be abrogated by phosphorylation of S981 by AKT1 or SGK1, which leads to interaction with 14-3-3 proteins and localization away from mTORC1 [45]. Exon 26 of TSC2, which contains S981, is found in a higher percentage of TSC2 transcripts from ESRP1-high patients (Fig 4C). This matches previous reports that have shown that TSC2 is alternatively spliced in response to ESRP1/2 knockdown [36].

Strikingly, we found similar ESRP1-mediated regulation of TSC2 S981 in breast cancer, where ESRP1 is oncogenic, but not in renal cancer, where ESRP1 has actually been suggested to be protective.

Given that TSC2 S981 appears to be included at higher rates in ESRP1-high tumors and its phosphorylation is known to be mediated by the PI3K/PTEN pathway, we were curious if PI3K/PTEN pathway activity was related to the oncogenic effects of ESRP1. In prostate cancer, PTEN deletions are a common alteration, found in 17.3% of the TCGA cohort. When comparing progression-free survival in patients that express high levels of ESRP1 but low levels of PTEN, we found that these patients exhibited significantly worse outcomes than other patients, even for those who had high ESRP1 or low PTEN alone (Fig. 4D). This suggests a synergistic relationship between ESRP1 and the PI3K/PTEN pathway, most likely through phosphorylation of TSC2 and subsequent activation of mTORC1. While the PI3K/PTEN/MTOR pathway is most commonly associated with AKT1, we actually observed an unexpected enrichment in the available substrates of SGK1 by ESRP1-mediated splicing, but not for AKT1 substrates, indicating that ESRP1 may be facilitating mTOR activity via SGK/PI3K/PTEN signaling axis (Supplementary Figure 16). Thus, patients with low expression of ESRP1 may contain a constitutively activated isoform of TSC2, leading to suppressed mTORC1 activity and less aggressive prostate cancers, while high expression of ESRP1 may lead to increased mTORC1 activity and proliferation in response to PI3K/PTEN pathway activity (Fig. 4E).

Discussion

This comprehensive analysis of the effect of splicing on PTMs and their linear motifs demonstrated a few key considerations for future studies, such as that inclusion/exclusion of PTMs by splicing is extensive, that regulatory or interaction connections might be diversified through alterations in motifs through splicing changes near PTM junctions or by exon swapping (mutually exclusive exons that preserve the predominant component of a PTM's motif), and the importance of considering PTMs within their specific groups (given the different behaviors across PTM types regarding constitutive rates, tissue specificity, flanking sequence regulation, and even the type of splicing events that alter PTMs). In addition to the systematic analysis of the global interplay between splicing and PTMs, exploring protein functional diversity within specific biological contexts driven by isoform-specific PTMs yields fruitful and new hypotheses. Here, we explored ESRP1-related prostate cancer differences, generating new hypotheses regarding the complexity that governs the disparate behaviors of ESRP1 across cancer types – that ESRP1 in some cancers relates to better outcomes (such as renal cancer), yet in other cancers is correlated with significantly worse prognosis (prostate and breast).

Of course, both these global and cancer-specific studies highlighted a major limitation stemming from the gaping lack of PTM annotation, including functional roles and regulatory connections (less than 5% of PTMs are annotated with additional functional context in PhosphoSitePlus [23]), which limits the current understanding of the full scope of the impact of isoform-specific PTMs on protein functional diversification. Yet, despite the limitation in individual PTM annotation, larger patterns between splicing and PTM-driven regulation still emerge. It has previously been suggested that transcription and alternative splicing are functionally coupled [46], and the activity of many splicing factors are regulated by post-translational modifications to induce splicing changes [47]. Examples of potential signaling-transcription-splicing cross-talk exist throughout literature, such as HRAS-mutant dependent splicing via SRSF proteins in lung cancer [48] and combined control of NF-κB signaling by splicing and PTMs [49]. In this work, we found a predicted coordination amongst possible SGK1/2/3 substrates differentially spliced in ESRP1-expressing prostate cancers, suggesting that ESRP1-induced splicing may facilitate SGK signaling in tumors where SGK is present and that there might be larger patterns in how signaling and protein interaction networks are controlled by splicing. Although we did not find a strong correlation between ESRP1 and SGK expression in the TCGA cohort, SGK1 has previously been shown to be upregulated in response to androgen receptor (AR) activity [50], a transcription factor that is one of the most common targets for prostate cancer treatments and shares many of the same transcriptional targets as ESRP1/2's splicing targets [51]. This suggests a potential dual mechanism of regulation where AR and ESRP1 activity stimulates the expression of both SGKs and SGK's substrates (via splicing),

altering SGK-related pathway activity including mTOR signaling. This relationship between SGKs and ESRP1 may not be limited to prostate cancer, as in breast cancer where ESRP1 expression also worsens patient prognosis, SGKs are amplified in some TCGA breast cancer patients and have been linked to estrogen receptor activity [52] [53]. Both ESRP1 and SGKs have been linked to resistance to therapies targeting estrogen or the estrogen receptor in breast cancer, which may be a result of both functioning in the same pathway [40] [54]. This suggests there may be a consistent mechanism of regulation that explains the tissue-dependent role of ESRP1. Further study is needed to confirm the link between ESRP1 and SGK signaling in vitro, but these initial findings provide an excellent picture as to how splicing, post-translational modifications, and their regulatory networks may work in tandem to produce phenotypic changes in cells.

Finally, although we overcame a significant barrier in integrating proteomic-level and genomic-level data to assess how PTMs might be regulated by splicing, and the process clearly has the potential to produce novel and interesting biological hypotheses, experimental connection will remain a challenge. Unfortunately, the vast majority of differential spliced PTMs occur in such a way that they are effectively indistinguishable amongst isoforms by traditional mass spectrometry workflows. For the same reason that databases of PTMs predominantly annotate canonical isoforms, the current experimental approaches will be unable to attribute the peptide-wise modification to the full protein details of an isoform. However, this work also highlights the interesting flip side of this conundrum – that the relative quantification of PTM measurements by mass spectrometry could potentially be a reflection of changes in overall modifiability of that target peptide across isoforms. For example, increases in a phosphorylation site, often attributed to an increase in a phosphorylation site on a single protein isoform, might also be a reflection of increased exon inclusion of that modifiable residue, rather than an overall increase in the kinase responsible for that site. Together, this study suggests key avenues for advancement include improving annotations of modifications (their function and interaction partners) and developing the experimental and computational approaches to connect modifications to specific isoforms. Additionally, we know that the field is rapidly improving and expanding in identification of new modification sites and transcript isoform identification. Given that growth, we have provided this pipeline for the easy ability to perform this analysis or approach to new datasets for mapping PTMs and exploring the interplay with splicing.

Methods

Data acquisition and processing

Exon and coding sequences for protein-coding genes belonging to version GRCh38.p14 of the GENCODE basic set were downloaded using the Biomart web interface. Additional exon, transcript, and gene meta information was downloaded from Biomart using the pybiomart package. Transcript sequences were derived from exon sequences and exon rank information, and the coding sequence was matched to its location in the transcript. Transcripts producing a protein isoform with fewer than twenty amino acids were removed from analysis. Post-translational modifications and other protein-level information associated with UniProt proteins were downloaded from ProteomeScout, a comprehensive database of post-translational modifications documented across five major compendia [1]. Protein sequences from Ensembl and ProteomeScout were compared and any discrepancies were removed from analysis. Transcripts that coded for the same protein sequence were collapsed into a single isoform entry. TRIFID functional scores associated with each transcript were downloaded from the APPRIS database [55].

Mapping PTMs to the Genome and Projecting them onto Alternative Isoforms

In order to map post-translational modifications to their location in the genome, coding sequences were first translated into the corresponding amino acid sequences using Biopython. Coding sequences that did not start with a start codon or had incomplete codons were removed from analysis. Based on the

position of the PTM site in the protein and the location of the coding sequence in the transcript, the transcript location of the codon responsible for the modifiable residue was obtained and the corresponding exon containing that codon identified. Using the genomic coordinates of the exon of interest, the PTM could then be mapped to its location in the genome. Next, to project PTM sites onto alternative transcripts and protein isoforms, we identified all exons in the transcriptome whose genomic coordinates contained the mapped PTM location. For each of these exons, the position of the PTM in the alternative protein was determined using location of the exon and start of the coding sequence in the alternative transcript. With PTMs projected, we validated that no frame shifts were introduced that would lead to a change in the residue associated with that position. In the rare case where a PTM site exists at the splice junction and is coded for by two exons, we checked for conservation of at least one of the two contributing exons in the alternative transcript, then validated that the residue remained unchanged. We call the PTMs that were successfully projected on an alternative transcript a "prospective" PTM site, as the majority of these sites have not been experimentally validated. See Supplementary Figure 3 for a toy example of this process.

Identifying Splice Events

When thinking about the different types of splice events that may impact PTMs, we focused predominantly on skipped exon events, alternative splice sites, and mutually exclusive exon events, as these are the events that can lead to loss of PTM sites and are a direct result of splicing. Given that the PTM-level information used in this work is documented on canonical UniProt isoforms, splice events were defined relative to the transcripts associated with the canonical UniProt isoform. Genes for which there was not a clear canonical Uniprot isoform were excluded from this analysis. For each exon associated with the canonical isoform, we first checked to see if the Ensembl exon ID matched any exons in the alternative transcripts. In cases where there were no matching IDs, we compared the genomic location of the exon in the canonical isoform to the genomic locations of exons in the alternative transcript. From this, we identified cases where 1) an alternative exon had the same genomic coordinates ("Conserved"), 2) no alternative exons have any overlapping genomic coordinates ("Skipped"), 3) or an alternative exon has partially overlapping genomic coordinates ("Alternative Splice Site"). For "Conserved" cases, we checked to see if the protein sequences coded for by the exon within both the canonical and alternative isoform were the same, or if they had been altered due to a frameshift or alternative promoter. These cases were removed from analysis, so that focus was specifically on splicing mechanisms. For exons that were identified as "Skipped" in an alternative transcript, we assessed whether there was potential for a mutually exclusive exon event, using criteria defined by Pillman et al. [19]. In brief, we identified exons in the alternative transcript as candidate mutually exclusive exons if they were adjacent to the skipped exon location and not found in the canonical isoform. We then defined true mutually exclusive exons as those that had a length difference of no more than 20 amino acids (60 nucleotides) and a sequence similarity of 15% or more. Sequence similarity was determined through global pairwise alignment between the canonical exon and MXE candidate, using gap penalties of -10 when introducing gaps and -2 for extending gaps. Alignment was performed using Biopython [56].

Calculating PTM and flanking sequence conservation across isoforms

For each known PTM, we calculated the fraction of alternative transcripts associated with the canonical isoform for which the prospective PTM site was found. PTM sites that were found in all isoforms of a given gene were defined as "constitutive". Given that multiple transcripts can code for the same protein sequences, we collapsed transcripts with matching protein sequences into a single isoform entry prior to calculating conservation. We could then calculate a "constitutive rate" for different modification types, which is the the fraction of PTMs of that modification type that could be defined as constitutive. For cases in which transcripts were filtered out based on different functional criteria (Supplementary Figure 8), potentially dysfunctional transcripts were removed from analysis prior to calculating constitutive rates.

In order to compare flanking sequences between canonical and alternative isoforms, we obtained the flanking sequences based on the projected location of the PTM in the alternative isoform, usually considering the five amino acids on either side of the PTM unless mentioned otherwise. To obtain tryptic fragments, we identified the nearest lysine and/or arginine residues to the PTM that are not proceeded by a proline, as this is the preferential cut sites for the enzyme trypsin [57]. We then compared the flanking sequences between the canonical and alternative isoform and identified cases in which the flanking sequence did not match (at least one amino acid was different). When calculating the overall rate of alteration for these flanking sequences, we only considered PTM sites that were successfully projected onto the alternative isoform of interest (i.e. it does not incorporate PTMs that are excluded from the isoform entirely).

Generating the null splicing model for comparison with patterns of modifications

In order to determine if the observed splicing control of PTMs was unique to modifiable residues and was not a general property of all residues, we developed a null model in which "modifiable" residues are randomly distributed across the transcriptome. For each modification type, we randomly sampled an identical number of residues as there are the modification of interest, with the same distribution of the specific amino acids that can be modified. For example, for the phosphorylation null model, we randomly sampled a total of 223,659 residues (131,236 serines, 54,358 threonines, 38,025 tyrosines), the same number of phosphorylation sites that have been observed in the proteome. With the random sample of residues, we then repeated the same analysis as done for the true modifiable residues (constitutive rate, flanking sequence identity, etc.). We repeated this process 100 times for each modification type to create a null model distribution of constitutive rates and altered flanking sequence rates. Finally, we considered the rate statistically significant if the observed value was found to have a higher or lower value than all but five or fewer of the random null models ($p \leq 0.05$).

Identifying tissue-specific PTM sites

Tissue-specificity data was downloaded from three different publications which used different approaches to identifying for tissue-specific exons and/or transcripts [21] [10] [20]. In cases in which only tissue-specific transcripts were reported, we compared the two transcript isoforms to identify tissue-specific exons, excluding alternative splice site events. Across all tissue-specific exons, density of PTMs, defined as the number of PTM sites per 1000 residues, were calculated by counting the number of PTM sites whose genomic coordinates fall within the tissue-specific exons and dividing that number by the amino acid length of each exon. Amino acid length excluded residues which were found at splice junctions and were coded for by two exons. PTM density was then normalized to the density of PTMs found across the entire proteome, defined as the $\frac{total\ number\ of\ PTM\ sites\ in\ canonical\ isoforms}{total\ number\ of\ residues\ in\ canonical\ isoforms}$. Thus, a normalized density greater than one indicates a higher density of PTM sites then expected relative to the entire proteome.

Predicting changes in kinase interactions using Kinase Library

To assess how changes to a PTM's flanking sequence may impact could impact which proteins interact with that site, we turned to a kinase motif enrichment tool, Kinase Library [24]. To get scores from Kinase Library, we first restricted our analysis to prospective PTMs in alternative isoforms with altered flanking sequences (based on five amino acids), a matching a tryptic fragment, and at least one established kinase interaction. We also restricted it to phosphoserine and phosphothreonine sites, as Kinase Library is currently only able to score serine/threonine kinases. In total, we scored a total of 57 different events affecting 47 different PTMs. For each, we used the Kinase Library web interface to score the flanking sequence in both the canonical and alternative isoforms, and then extracted the 'site percentile', which indicates where the phosphorylation site ranks relative to all other phosphorylation

sites scored for that kinase. We then calculated the change in site percentile from the canonical to alternative sequence.

Identifying PTM sites with variable inclusion in ESRP1-high and -low prostate cancer patients

In order to identify PTM sites specific to individual prostate cancer patients from The Cancer Genome Atlas (TCGA), "percent spliced in" values (PSI) specific to each patient were downloaded from the TCGASpliceSeq database, which indicate the fraction of transcripts within a given patient for which the exon of interest is included [34]. PSI values were calculated using SpliceSeq, an algorithm designed to predict alternative splicing patterns from RNA sequencing data, like the ones generated by TCGA [34]. While the majority of this work relies on Ensembl hg38 coordinates, TCGASpliceSeq reports exons in terms of hg19 coordinates, so genomic coordinates of PTM sites were converted into the hg19 coordinates system using the pyliftover python package. From these genomic locations, we projected PTMs onto SpliceSeq exons using the same procedure as with Ensembl transcripts. Next, we downloaded patient-specific expression of ESRP1 mRNA across the from cBioPortal, normalized across all samples [35]. We also downloaded survival data and alteration data for ESRP1 and PTEN from cBioPortal. ESRP1-high and low patients were defined as patients with ESRP1 mRNA expression greater or less than one standard deviation from the mean expression across the TCGA cohort, respectively. Finally, we compared inclusion of SpliceSeq exons across the ESRP1-low and -high groups and identified statistically significant differences using a Mann Whitney U test and Benjamini-Hochberg FDR correction ($p \leq 0.05$, $r \geq 0.25$). PTM sites projected onto these differentially included exons were then also defined as differentially included.

PTM- and gene-level enrichment analysis

PTM-site specific enrichment analysis was performed utilizing annotations downloaded from PhosphoSitePlus which include known molecular function or biological process of a given PTM site, if any is known [23]. Enrichment for a specific molecular function or biological process was assessed using a Fisher's Exact test, where the background population either included all PTM sites in canonical isoforms (as in Fig. 2) or all PTMs identified in the TCGA prostate cancer SpliceSeq dataset (as in Fig4). P-values were corrected using Benjamini-Hochberg false positive correction [58]. Gene-specific enrichment analysis was performed using the Enrichr wrapper from the gseapy python package [59] [60] [61]. Gene sets from Gene Ontology and KEGG were used [62] [63]. The background population was defined as all genes identified in the TCGA prostate cancer SpliceSeq dataset (as in Fig. 4).

To identify kinases with enriched substrates regulated by ESRP1 (Supplementary Figure 16), we utilized kinase-substrate networks of known connections from PhosphoSitePlus [23] or predicted connections from KSTAR [64]. In both cases, a one-tailed hypergeometric test was used to assess statistically significant enrichment of a kinase's substrates across ESRP1-regulated PTMs, with all phosphorylation sites identified in the TCGA SpliceSeq data used as the background population. For the KSTAR networks, the median p-value enrichment was extracted across the ensemble of 50 predicted kinase-substrate networks to obtain a final enrichment value.

Data and Code Availability

All data generated for this work has been deposited in Figshare at `https://doi.org/10.6084/m9.figshare.24969993`. The computational pipeline used to project PTMs onto transcripts is freely available at the following github repository: `https://github.com/NaegleLab/ExonPTMapper`. All additional code and analysis done for this work can be found at a separate github repository here: `https://github.com/NaegleLab/PTM_Splicing_Analysis`.

Acknowledgements

We would like to thank Noah Perry and Mete Civelek for their guidance and input throughout the development of this work. Additionally, we would like to thank Ben Jordan for early mentorship of undergraduates working on this project. We would also like to thank the Systems Biology Community at the University of Virginia for their invaluable feedback on the data analysis and the manuscript figures. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM138127 and by the National Science Foundation Graduate Research Fellowship under Grant No. 1842490. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

References

1. Matlock, M. K., Holehouse, A. S. & Naegle, K. M. Proteomescout: A repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Research* (2015).

2. Mujtaba, S., Zeng, L. & Zhou, M. M. Structure and acetyl-lysine recognition of the bromodomain. *Oncogene 2007 26:37* **26**, 5521–5527 (2007). URL `https://www.nature.com/articles/1210618`.

3. Reinhardt, H. C. & Yaffe, M. B. Phospho-ser/thr-binding domains: navigating the cell cycle and dna damage response. *Nature Reviews Molecular Cell Biology 2013 14:9* **14**, 563–580 (2013). URL `https://www.nature.com/articles/nrm3640`.

4. Dikic, I., Wakatsuki, S. & Walters, K. J. Ubiquitin-binding domains — from structures to functions. *Nature Reviews Molecular Cell Biology 2009 10:10* **10**, 659–671 (2009). URL `https://www.nature.com/articles/nrm2767`.

5. Akiva, E., Friedlander, G., Itzhaki, Z. & Margalit, H. A dynamic view of domain-motif interactions. *PLOS Computational Biology* **8**, e1002341 (2012). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002341`.

6. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist (2014). URL `http://dx.doi.org/10.1016/j.molcel.2014.05.032`.

7. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008). URL `http://www.nature.com/articles/nature07509`.

8. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics 2008 40:12* **40**, 1413–1415 (2008). URL `https://www.nature.com/articles/ng.259`.

9. Comoletti, D. *et al.* Characterization of the interaction of a recombinant soluble neuroligin-1 with neurexin-1β. *Journal of Biological Chemistry* **278**, 50497–50505 (2003). URL `http://www.jbc.org/article/S0021925820755705/fulltexthttp://www.jbc.org/article/S0021925820755705/abstracthttps://www.jbc.org/article/S0021-9258(20)75570-5/abstract`.

10. Buljan, M. *et al.* Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular Cell* **46**, 871–883 (2012).

11. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012). URL `https://www.science.org/doi/10.1126/science.1228186`.

12. Colak, R. *et al.* Distinct types of disorder in the human proteome: Functional implications for alternative splicing. *PLOS Computational Biology* **9**, e1003030 (2013). URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003030`.

13. Weatheritt, R. J., Davey, N. E. & Gibson, T. J. Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Research* **40**, 7123–7131 (2012). URL `https://dx.doi.org/10.1093/nar/gks442`.

14. Wang, X. *et al.* Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Molecular & Cellular Proteomics : MCP* **17**, 422 (2018). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836368/https://onedrive.live.com/view.aspx?resid=D7F68FF2A2AD0CF4%2125440&id=documents&wd=target%28Papers.one%7C2571AE6F-B743-44CC-925E-F28A5791D905%2FDiversityLimitedbyTrypsinCleavage%7CD4FF7E86-F077-4AF`.

15. Goldtzvik, Y., Sen, N., Lam, S. D. & Orengo, C. Protein diversification through post-translational modifications, alternative splicing, and gene duplication. *Current Opinion in Structural Biology* **81**, 102640 (2023). URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X23001148.

16. Zhou, J., Zhao, S. & Dunker, A. K. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *Journal of Molecular Biology* **430**, 2342–2359 (2018).

17. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988–D995 (2022). URL https://dx.doi.org/10.1093/nar/gkab1049.

18. Bateman, A. *et al.* Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021). URL https://academic.oup.com/nar/article/49/D1/D480/6006196.

19. Pillmann, H., Hatje, K., Odronitz, F., Hammesfahr, B. & Kollmar, M. Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics* **12**, 1–16 (2011). URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-270. Constraints on MXEs¡br/¿- MXE's must have the same length (difference less than 20aa)¡br/¿- Must have conserved splice site patterns (GT-AG, GC-AG, GG-AG, AT-AC) OR ¡br/¿- Reading frame must be conserved¡br/¿- They must show sequence similarity (¿15% similarity)¡br/¿¡br/¿.

20. Rodriguez, J. M., Pozo, F., Domenico, T. D., Vazquez, J. & Tress, M. L. An analysis of tissue-specific alternative splicing at the protein level. *PLOS Computational Biology* **16**, e1008287 (2020). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008287.

21. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* **14**, 1–11 (2013). URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-7-r70.

22. Pozo, F. *et al.* Assessing the functional relevance of splice isoforms. *NAR Genomics and Bioinformatics* **3**, 1–16 (2021). URL https://academic.oup.com/nargab/article/3/2/lqab044/6281449.

23. Hornbeck, P. V. *et al.* Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research* **40**, D261–70 (2012). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245126&tool=pmcentrez&rendertype=abstract.

24. Johnson, J. L. *et al.* An atlas of substrate specificities for the human serine/threonine kinome. *Nature 2023 613:7945* **613**, 759–766 (2023). URL https://www.nature.com/articles/s41586-022-05575-3.

25. Fischle, W., Franz, H., Jacobs, S. A., Allis, C. D. & Khorasanizadeh, S. Specificity of the chromodomain y chromosome family of chromodomains for lysine-methylated ark(s/t) motifs. *Journal of Biological Chemistry* **283**, 19626–19635 (2008). URL http://www.jbc.org/article/S0021925820814161/fulltexthttp://www.jbc.org/article/S0021925820814161/abstracthttps://www.jbc.org/article/S0021-9258(20)81416-1/abstract.

26. Scotti, M. M. & Swanson, M. S. Rna mis-splicing in disease (2016).

27. Wang, G. S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics 2007 8:10* **8**, 749–761 (2007). URL https://www.nature.com/articles/nrg2164.

28. Fletcher, S. *et al.* Antisense suppression of donor splice site mutations in the dystrophin gene transcript. *Molecular Genetics and Genomic Medicine* **1**, 162–173 (2013).

29. Voena, C. *et al.* Oncogenic alk regulates emt in non-small cell lung carcinoma through repression of the epithelial splicing regulatory protein 1. *Oncotarget* **7**, 33316 (2016). URL `/pmc/articles/PMC5078097//pmc/articles/PMC5078097/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5078097/`.

30. Lee, H. H. *et al.* Epithelial splicing regulatory protein (espr1) expression in an unfavorable prognostic factor in prostate cancer patients. *Frontiers in Oncology* **10**, 2274 (2020).

31. Liu, Y. *et al.* Underlying mechanisms of epithelial splicing regulatory proteins in cancer progression. *Journal of Molecular Medicine 2022 100:11* **100**, 1539–1556 (2022). URL `https://link.springer.com/article/10.1007/s00109-022-02257-5`.

32. Freytag, M. *et al.* Epithelial splicing regulatory protein 1 and 2 (esrp1 and esrp2) upregulation predicts poor prognosis in prostate cancer. *BMC Cancer* **20**, 1–12 (2020). URL `https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07682-8`.

33. Gerhauser, C. *et al.* Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996–1011.e8 (2018). - ESRP1 duplication correlated with ESRP1 expression (¿1.5 fold change)¡br/¿- Some patients exhibited both ESRP1 and Myc amplifications, but only ESRP1 showed changes at mRNA level¡br/¿- ESRP1 duplications exhibited higher Gleason scores (indicative of worse tumor stages)¡br/¿- Increased ESRP1 protein correlated with higher proliferation right by Ki67 labeling¡br/¿- Showed that it is an independent biomarker¡br/¿- Showed strong relationship between PEPCI (Purity-Adjusted Epigenetic Prostate CAncer Index) and ESRP1 amplification.

34. Ryan, M. *et al.* Tcgaspliceseq a compendium of alternative mrna splicing in cancer. *Nucleic Acids Research* **44**, D1018–D1022 (2016). URL `https://academic.oup.com/nar/article/44/D1/D1018/2503095`.

35. Cerami, E. *et al.* The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401–404 (2012).

36. Yang, Y. *et al.* Determination of a comprehensive alternative splicing regulatory network and combinatorial regulation by key factors during the epithelial-to-mesenchymal transition. *Molecular and Cellular Biology* **36**, 1704–1719 (2016). URL `https://www.tandfonline.com/doi/abs/10.1128/MCB.00019-16`.

37. Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B. & Carstens, R. P. Esrp1 and esrp2 are epithelial cell-type-specific regulators of fgfr2 splicing. *Molecular Cell* **33**, 591–601 (2009).

38. Brown, R. L. *et al.* Cd44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *The Journal of Clinical Investigation* **121**, 1064–1074 (2011). URL `http://www.jci.org`.

39. Ishii, H. *et al.* Epithelial splicing regulatory proteins 1 (esrp1) and 2 (esrp2) suppress cancer cell motility via different mechanisms. *Journal of Biological Chemistry* **289**, 27386–27399 (2014). URL `http://www.jbc.org/article/S0021925820483784/fulltexthttp://www.jbc.org/article/S0021925820483784/abstracthttps://www.jbc.org/article/S0021-9258(20)48378-4/abstract`.

40. Gökmen-Polar, Y. *et al.* Splicing factor esrp1 controls er-positive breast cancer by altering metabolic pathways. *EMBO reports* **20**, e46078 (2019). URL `https://onlinelibrary.wiley.com/doi/full/10.15252/embr.201846078https://onlinelibrary.wiley.com/doi/abs/10.15252/embr.201846078https://www.embopress.org/doi/10.15252/embr.201846078`.

41. Nanni, M., Ranieri, D., Persechino, F., Torrisi, M. R. & Belleudi, F. The aberrant expression of the mesenchymal variant of fgfr2 in the epithelial context inhibits autophagy. *Cells 2019, Vol. 8, Page 653* **8**, 653 (2019). URL https://www.mdpi.com/2073-4409/8/7/653/htmhttps://www.mdpi.com/2073-4409/8/7/653.

42. Soucek, T., Pusch, O., Wienecke, R., DeClue, J. E. & Hengstschläger, M. Role of the tuberous sclerosis gene-2 product in cell cycle control. *Journal of Biological Chemistry* **272**, 29301–29308 (1997). URL http://www.jbc.org/article/S0021925818508863/fulltexthttp://www.jbc.org/article/S0021925818508863/abstracthttps://www.jbc.org/article/S0021-9258(18)50886-3/abstract.

43. Zhang, H. *et al.* Loss of tsc1/tsc2 activates mtor and disrupts pi3k-akt signaling through downregulation of pdgfr. *The Journal of Clinical Investigation* **112**, 1223–1233 (2003).

44. Pai, G. M. *et al.* Tsc loss distorts dna replication programme and sensitises cells to genotoxic stress. *Oncotarget* **7**, 85365 (2016). URL /pmc/articles/PMC5356742//pmc/articles/PMC5356742/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5356742/.

45. Cai, S. L. *et al.* Activity of tsc2 is inhibited by akt-mediated phosphorylation and membrane partitioning. *The Journal of Cell Biology* **173**, 279 (2006). URL /pmc/articles/PMC2063818//pmc/articles/PMC2063818/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2063818/.

46. Kornblihtt, A. R., Mata, M. D. L., Fededa, J. P., Muñoz, M. J. & Nogués, G. Multiple links between transcription and splicing. *RNA* **10**, 1489–1498 (2004). URL http://rnajournal.cshlp.org/content/10/10/1489.fullhttp://rnajournal.cshlp.org/content/10/10/1489http://rnajournal.cshlp.org/content/10/10/1489.abstract.

47. Kretova, M., Selicky, T., Cipakova, I. & Cipak, L. Regulation of pre-mrna splicing: Indispensable role of post-translational modifications of splicing factors. *Life 2023, Vol. 13, Page 604* **13**, 604 (2023). URL https://www.mdpi.com/2075-1729/13/3/604/htmhttps://www.mdpi.com/2075-1729/13/3/604.

48. Lo, A., McSharry, M. & Berger, A. H. Oncogenic kras alters splicing factor phosphorylation and alternative splicing in lung cancer. *BMC Cancer* **22**, 1–13 (2022). URL https://bmccancer.biomedcentral.com/articles/10.1186/s12885-022-10311-1http://creativecommons.org/publicdomain/zero/1.0/.

49. Leeman, J. R. & Gilmore, T. D. Alternative splicing in the nf-κb signaling pathway. *Gene* **423**, 97–107 (2008).

50. Shanmugam, I. *et al.* Serum/glucocorticoid-induced protein kinase-1 facilitates androgen receptor-dependent cell survival. *Cell Death & Differentiation 2007 14:12* **14**, 2085–2094 (2007). URL https://www.nature.com/articles/4402227.

51. Shah, K. *et al.* Androgen receptor signaling regulates the transcriptome of prostate cancer cells by modulating global alternative splicing. *Oncogene 2020 39:39* **39**, 6172–6189 (2020). URL https://www.nature.com/articles/s41388-020-01429-2.

52. Toska, E., Castel, P., Leslie, C. S., Scaltriti, M. & Baselga, J. Pi3k inhibition activates sgk1 via a feedback loop to promote chromatin-based regulation of er-dependent gene expression. *CellReports* **27**, 294–306.e5 (2019). URL https://doi.org/10.1016/j.celrep.2019.02.111.

53. Wang, Y. *et al.* Sgk3 is an estrogen-inducible kinase promoting estrogen-mediated survival of breast cancer cells. *Molecular Endocrinology* **25**, 72–82 (2011). URL https://dx.doi.org/10.1210/me.2010-0294.

54. Wang, Y. *et al.* Sgk3 sustains erα signaling and drives acquired aromatase inhibitor resistance through maintaining endoplasmic reticulum homeostasis. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E1500–E1508 (2017). URL https://www.pnas.org/doi/abs/10.1073/pnas.1612991114.

55. Rodriguez, J. M. *et al.* Appris: annotation of principal and alternative splice isoforms. *Nucleic acids research* **41** (2013). URL https://pubmed.ncbi.nlm.nih.gov/23161672/.

56. Cock, P. J. *et al.* Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

57. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular and Cellular Proteomics* **3**, 608–614 (2004). URL http://www.mcponline.org/article/S1535947620338779/fulltexthttp://www.mcponline.org/article/S1535947620338779/abstracthttps://www.mcponline.org/article/S1535-9476(20)33877-9/abstract.

58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995). URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x.

59. Fang, Z., Liu, X. & Peltz, G. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics* **39** (2023). URL https://dx.doi.org/10.1093/bioinformatics/btac757.

60. Chen, E. Y. *et al.* Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics* **14** (2013). URL https://pubmed.ncbi.nlm.nih.gov/23586463/.

61. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016). URL https://pubmed.ncbi.nlm.nih.gov/27141961/.

62. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics 2000 25:1* **25**, 25–29 (2000). URL https://www.nature.com/articles/ng0500_25http://fruitfly.bdgp.berkeley.eduhttp://genome-www.stanford.eduhttp://www.informatics.jax.org.

63. Kanehisa, M. & Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30 (2000). URL https://dx.doi.org/10.1093/nar/28.1.27.

64. Crowl, S., Jordan, B. T., Ahmed, H., Ma, C. X. & Naegle, K. M. Kstar: An algorithm to predict patient-specific kinase activities from phosphoproteomic data. *Nature Communications 2022 13:1* **13**, 1–16 (2022). URL https://www.nature.com/articles/s41467-022-32017-5.

Supporting Information