Deciphering sequence determinants of alternative splicing and polyadenylation in health and disease with massively parallel reporter assays

Samantha Elyse Koplik

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Georg Seelig, Chair

Andre Berndt

Barry Lutz

Program Authorized to Offer Degree:

Bioengineering

University of Washington

Abstract

Deciphering sequence determinants of alternative splicing and polyadenylation in health and disease with massively parallel reporter assays

Samantha Elyse Koplik

Chair of the Supervisory Committee:

Georg Seelig

Electrical & Computer Engineering

Paul G. Allen School for Computer Science & Engineering

Splicing and polyadenylation are major co- and post-transcriptional processes that regulate gene expression. Alternative splicing (AS) and alternative polyadenylation (APA) are frequent drivers of human disease, yet systematic maps of how genetic variants affect these processes in different cell types remain limited. To address this gap, this thesis presents high-throughput perturbations of splicing and polyadenylation using massively parallel reporter assays (MPRAs) to measure the impact of genetic variation on both AS and APA in human cell lines of diverse tissue origin. First, I introduce Cell-type Oriented Massively Parallel Assay of Splicing Signatures (COMPASS), an MPRA that quantifies splicing outcomes for 87,546 variants across five human cell lines. COMPASS targets disease-relevant genes, including

ACMG actionable and autism-associated genes, providing a resource to systematically dissect splicing impacts in health and disease. Benchmarking COMPASS data against predictive models highlights both strengths and weaknesses of current approaches. Biological relevance is further supported by prime editing experiments that validate selected variants in their native genomic context. Analyses of COMPASS data also reveal RNA-binding protein motifs whose disruption drives splicing changes and identify subsets of sequences that mediate cell type-specific splicing programs. Next, I applied a similar approach to dissect the *cis*-regulatory determinants of APA. APARENT2, a deep residual network for predicting APA, had previously identified variants enriched for gain-of-function for polyadenylation in autism GWAS cohorts. To validate these predictions, I developed an MPRA in multiple cell types, which confirmed these gain-of-function variants and uncovered additional cell type-specific effects. I further expanded this work to a larger APA MPRA that cataloged the effects of over 5,000 disease-associated variants, revealing both conserved and cell type-specific regulation. Together, these studies provide the most comprehensive cell type-resolved compendia of AS and APA to date. COMPASS delivers the largest atlas of splice-disrupting variants to date, and two APA MPRAs provide a complementary resource cataloging polyadenylation-disrupting variants. The resources developed in this thesis serve to support variant reclassification in clinical genomics, guide therapeutic target discovery, and aid in the refinement of predictive models.

Acknowledgments

I want to thank my advisor, Georg Seelig, for the encouragement, advice, and mentorship he has provided over my 5-plus years at UW. I would also like to thank the members of my doctoral committee, Andre Berndt, Barry Lutz, and my GSR, Jennifer Nemhauser.

I would like to express my deepest gratitude to Andrew, my fiancé and partner of eleven and a half years, whose unwavering love, patience, and encouragement have carried me through every stage of my PhD journey. From the early days of coursework and failed experiments to the stress of making presentations and the challenges of writing this thesis, you have been my constant source of strength and support. Your belief in me has made all the difference, and I am endlessly thankful to have shared this chapter of life with you. I also want to thank our dog, Luna, who has been with me since my first year of grad school. She has been a source of comfort and emotional support, has become a friend to my lab mates, and has kept me lots of company through writing this thesis.

I would like to thank all members of the Seelig Lab, past and present, for their support, camaraderie, and collaboration throughout my PhD. I am so grateful to my friends and lab mates in the Seelig Lab for making it such a fun and entertaining place to do science. From trivia nights, growing sea monkeys, hangouts at Shultzy's, lab coat competitions, bully matrices, and, of course, many bags of Nerds Gummy Clusters from MolES and Johnson. You have all made the lab a memorable place.

I am deeply thankful to my family for their love and encouragement throughout this journey. To my mom and dad, my step-parents, and my siblings. Your support has meant everything to me, and I could not have reached this milestone without you.

Acknowledgment of Contributions

The research described in this thesis was carried out in collaboration with past and present members of the Seelig Lab, under the guidance of my advisor, Georg Seelig, who has provided invaluable advice on all aspects of the research. Accordingly, I use "we" throughout the text to reflect both my own contributions and those of my collaborators and colleagues, though this thesis highlights the work that primarily represents my own contributions. Specific contributions are noted where relevant: **Angela M Yu** contributed to much of the splicing MPRA work, **Madelyn Shelby** contributed to the cell type-specific low-throughput experimental validations. All prime editing experiments and prime editing data processing were performed by **Madelyn Shelby. Johannes Linder** conceived and developed APARENT2 and was the main contributor to the computational aspects of this research, while I carried out all experimental validation. Together, Johannes and I also designed and developed a larger APA MPRA. Chapters 2 and 3 (AS MPRA) are based on collaborative work that is part of a manuscript that has been submitted for review at *Cell*, and I gratefully acknowledge the contributions of all co-authors on this project. In addition, much of Chapter 4 (APARENT2) has already been published in *Genome Biology* (Linder et al., 2022). My graduate work was supported by the National Science Foundation Graduate Research Fellowship Program (NSF GRFP) under Grant No. DGE-2140004.

Dedication

I dedicate this thesis to my grandpa, Bernard Koplik (1934-2021), who was a professor of mechanical engineering and who dedicated his life to learning and teaching. While he is no longer with us, I miss him deeply and often think back to the days he helped me with my engineering homework in college. He was an amazing grandfather, and his passion for teaching and his love for science continue to inspire me every day.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

Before beginning my PhD, I worked in experimental synthetic biology, developing genetic circuits for the treatment of human disease. This experience revealed how limited our understanding of biological systems remains and convinced me that a deeper grasp of the relationship between sequence and function is essential for engineering effective therapies. I became fascinated by the potential of machine learning (ML) to elucidate this relationship, especially in the context of disease-associated variants. While progress has been made in identifying sequence determinants of co- and post-transcriptional regulation, our understanding of how the *cis*-regulatory code shapes processes such as alternative splicing (AS) and alternative polyadenylation (APA) is still incomplete.

Massively parallel reporter assays (MPRAs) have proven especially powerful for this purpose, enabling high-throughput measurement of molecular phenotypes and producing the large-scale datasets needed to evaluate how well existing predictive models generalize. These measurements underscore the limitations of models trained largely on randomized sequence libraries or on population cohorts, which often fail to capture the effects of rare or clinically relevant variants absent from the population. Many such variants may exert regulatory effects through mechanisms such as RNA-binding protein (RBP) interactions or RNA secondary structure, features that most current models cannot reliably predict, but that MPRAs can directly measure. At the same time, MPRAs provide a means of uncovering the functional consequences of disease-associated and rare variants beyond the scope of many existing training datasets.

Throughout my thesis work, I aimed to generate and characterize MPRA datasets to investigate how sequence variation impacts co- and post-transcriptional regulation. For AS, I focused on the intersection of *cis*-regulatory elements and their associated trans-acting factors to

decipher how RBP motifs can impact splicing decisions. For both AS and APA, I further examined cell type-specific regulation by implementing MPRAs in multiple cell lines, providing insight into how genetic variation influences RNA processing in different biological contexts. I also analyzed disease-associated variants to uncover how disruptions in splicing or polyadenylation are implicated in human disease. To complement these analyses, I applied existing ML models to validate these datasets, identify contexts where predictions break down, and highlight regulatory effects that models cannot yet capture. Together, this work establishes a framework for understanding the sequence and regulatory logic of RNA processing and provides insight into the effects of disease-associated variants, which may aid in the identification of new therapeutic targets, especially in diseases where splicing or polyadenylation is known to play a critical role (e.g., spinal muscular atrophy, Duchenne muscular dystrophy, cancer, Autism Spectrum Disorder, Alzheimer's disease).[1–15]

Briefly, the topics this work will cover are:

Chapter 2: Development of COMPASS (cell type Oriented Massively Parallel reporter Assay of Splicing Signatures)

In this chapter, I introduce Cell-type Oriented Massively Parallel Assay of Splicing Signatures (COMPASS), an MPRA used to decipher the *cis*-regulatory drivers of disease-associated splicing regulation. COMPASS measures splicing outcomes for 87,546 single and double variants across more than 1,700 genes in five human cell lines of diverse tissue origin. COMPASS targets disease-relevant gene sets, including ACMG actionable genes and SFARI autism-associated genes, enabling systematic dissection of splicing impacts in health and

disease. Our measurements reveal numerous splice-disrupting variants (SDVs), including ClinVar variants currently classified as variants of uncertain significance. Using prime editing, we validate variant effects in the genome for ClinVar variants in *BIN1*, a gene implicated in Alzheimer's disease, cancer, and cardiac pathology. Benchmarking COMPASS against state-of-the-art predictive models shows that while these approaches capture broad variant effects, important gaps still remain.

Chapter 3: Using COMPASS to map splicing outcomes mediated by RBP motifs and cell type-specific contexts

Using COMPASS data, we examine the role of RBPs in shaping splicing regulation across diverse human cell lines (e.g., HEK293, HeLa, K562, MCF7, and HMC3). By systematically mapping putative RBP motifs and quantifying the effect sizes of their disruption, we dissect the contribution of *cis*-regulatory elements to splicing and highlight potential modes of regulation mediated by their trans-acting factors. These analyses reveal mechanisms by which RBPs contribute to both global and cell type-specific splicing regulation. We also compare splicing outcomes across cell lines, identifying a distinct subset of variants that drive cell type-specific splicing programs. Finally, we benchmark tissue-specific predictive models against COMPASS measurements, highlighting gaps where model predictions fail to capture experimentally observed differences.

Chapter 4: An MPRA of clinically relevant variants in multiple cell lines to validate a residual neural network for predicting 3′ cleavage and polyadenylation

In this chapter, I present the use of MPRAs to interrogate the sequence determinants of APA. Using the sequence-to-function model APARENT2, we computationally predict the effects of candidate APA variants and benchmark these predictions against human variation data as well as models designed for tissue-specific predictions. We then experimentally tested the impact of APA variants linked to human disease, with a particular focus on autism spectrum disorder (ASD). Portions of this work have been adapted from our published study in *Genome Biology* (Linder J, Koplik SE, et al., 2022). We further extend this analysis by developing a larger MPRA that assays over 10,000 disease-associated sequences predicted to affect polyadenylation by APARENT2. This MPRA was tested in two human cell lines (HEK293 and K562), enabling systematic evaluation of disease-associated APA variants across cellular contexts and identifying gaps in how predictive models capture APA regulation.

Chapter 2: Development of COMPASS (Cell-type Oriented Massively Parallel reporter Assay of Splicing Signatures)

2.1 Motivation

RNA splicing removes non-coding introns from pre-mRNA and joins the protein-coding exons to enable translation of mRNA into functional proteins. An embedded *cis*-regulatory code determines when and where emerging transcripts are spliced. Splice donors (SDs) and splice acceptors (SAs) define exon-intron boundaries and engage the core splicing machinery.[16] Through the use of competing splice sites, a single gene can produce multiple isoforms, thereby substantially increasing protein diversity in eukaryotes.[17,18]. Splicing outcomes vary depending on the interplay between RBPs and their binding to *cis*-regulatory RNA elements, which collectively guide spliceosome assembly and exon inclusion.[19,20] Further, heterogeneity in splicing outcomes arises from cell type differences in RBP expression, with nearly half of alternatively spliced isoforms displaying tissue-restricted patterns.[19–21] However, quantifying splicing in heterogeneous tissues is challenging because bulk tissue measurements average signals across all cell types, thereby masking cell type-specific effects.[22]

Furthermore, variants that disrupt the *cis*-regulatory splicing code can change RNA and subsequent protein isoform ratios, leading to a variety of human diseases.[23,24] Although much progress has been made toward deciphering the *cis*-regulatory code governing AS, we still cannot accurately predict the impact of variants on splicing, beyond those that directly modify SD and SA sequences. Splicing quantitative trait loci (sQTL) studies can link genetic variants to splicing outcomes at the population level, but they remain correlative and offer limited mechanistic insight into how individual variants alter splicing.[14,25–28] The relatively recent development of accurate splicing models has enabled considerable progress towards the

identification of splice disrupting variants (SDVs).[29–37] Still, such models are not yet equipped to fully capture cell type-specific variant impact and, with few exceptions,[36,37] are not designed to be easily interpretable and suggest mechanistic explanations for observed splicing phenotypes.

Synthetic splicing reporters are an attractive tool for studying sequence-to-function relationships, overcoming some of the limitations of traditional transcriptomics studies, enabling us to characterize less commonly observed sequences, including rare, pathogenic, or *de novo* variants. Prior work demonstrated the utility of splicing massively parallel reporter assays (MPRAs) to quantify variant impact on splicing for thousands of variants across hundreds or even thousands of exon contexts (MFASS, MapSY, Vex-seq).[38–40] Most studies using synthetic reporters examine both references and single nucleotide variants (SNVs) in introns or exons to measure the variant's impact on isoform abundance. Complementary research employed MPRAs with designed or fully randomized sequence elements, which can be cheaply scaled to hundreds of thousands of sequences to systematically characterize all possible *cis*-regulatory elements of a given length (e.g., hexamers) or even train predictive sequence function models.[30,36,41–47] Alternatively, saturation mutagenesis of single disease-relevant exons has been used to assay a few hundred to more than a thousand SNVs (eg, FAS exon 6, WT1 exon 5, RON exon 11, CD19 exons 1–3, and POU1F1 exon 2).[48–53]

These challenges and advances together motivate COMPASS (Cell-type Oriented Massively Parallel reporter Assay of Splicing Signatures), which scales splicing MPRAs to nearly 90,000 variants across more than 2,000 exons in five cell lines. In this context, cell lines provide a powerful system: they are homogeneous at the population level yet distinct enough to resolve regulatory differences that would be obscured in heterogeneous tissues, while still experimentally accessible for large-scale functional assays such as MPRAs. We test both single

and double variants in the same sequence background to understand positional and combinatorial variant effects. By comparing the impact of *de novo* variants and variants of uncertain significance (VUS) with that of pathogenic variants in the same exons and introns, we can also identify candidate disease-causing variants. Our dataset provides a foundation for further validation and potential improvement of computational splicing models.

2.2 Overview of existing technologies

2.2.1 Usage of MPRAs to decode regulatory mechanisms

Variants that disrupt the *cis*-regulatory splicing code can change RNA and protein isoform ratios and identities, leading to a variety of human diseases.[23,24] Although much progress has been made toward deciphering the *cis*-regulatory code governing AS, we still cannot accurately predict the impact of variants on splicing, beyond those that directly modify SD and SA sequences.

MPRAs are an attractive approach for studying sequence-to-function relationships, overcoming some of the limitations of traditional population studies.[54–56] Array-based DNA synthesis enables the construction of libraries that are orders of magnitude larger than the number of variants typically identified in genomic studies. Libraries can range from hundreds to over 100,000 designed sequences, enabling systematic exploration of sequence space at a scale far beyond naturally occurring variation. [57,58] High-throughput RNA sequencing then quantifies the activity of thousands to millions of reporter constructs in parallel. In an MPRA, sequence variation is confined to a region of interest (e.g., an exon or intron), while the remainder of the reporter is held constant, thereby isolating the impact of sequence changes on the process being studied.

This framework allows systematic exploration of both common and rare variants, including those present in dbSNP or annotated in ClinVar as pathogenic or of uncertain significance, as well as *de novo* sequences.[59–61] By capturing both population-observed and hypothetical variation, MPRAs expand the scope of functional analysis beyond what is accessible through GWAS alone and complement computational predictors with direct functional measurements. To date, MPRAs have been applied to a range of transcriptional and post-transcriptional processes, including transcription, AS ,APA, mRNA stability, and translation.[39,40,58,59,61–68]

Our lab has previously developed an MPRA to interrogate AS, constructed from two synthetic libraries each containing two fully randomized 25-nucleotide sequences. One library targeted alternative 5′ SD sites and comprised 265,137 unique reporters, while the second targeted alternative 3′ SA sites and comprised 2,211,789 unique reporters. Data from these libraries were used to train HAL, a predictive model that estimates splicing outcomes by assigning weights to 6-mer sequence features and summing their contributions to predict changes in percent spliced in (PSI). This multinomial linear regression framework demonstrated strong performance in predicting the effects of human single-nucleotide variants within exons and at donor sites in the assay.[61] Importantly, this represented one of the first applications of MPRA-derived functional data to train a quantitative splicing model, establishing that empirically measured sequence-function relationships can be leveraged to infer splicing regulatory rules. This multinomial linear regression model performed well at the task of at predicting the effect of human SNVs in exons and at SDs in this MPRA, outperforming other models at the time of publication.[69]

2.2.2 Machine learning for the prediction of AS outcomes

Other groups have since developed splicing models that leverage large-scale genomics datasets. MMSplice, created by the Gagneur lab, is composed of separately trained neural network modules scoring exons, introns, and splice sites. These modules generate predictions for PSI, splicing efficiency, and variant pathogenicity.[32] From this same group, Cheng et al. extended this framework with MTSplice, a neural network that predicts tissue-specific variant effects by combining MMSplice outputs with outputs from a model called TSplice trained on AS catalog of the transcriptome (ASCOT) data derived from ENCODE and GTEx.[28,70,71] However, relatively few exons with tissue-specific effects are represented in ASCOT, and the dataset is biased toward brain tissues, limiting model generalizability.[71] Similarly, GTEx provides comprehensive transcriptomic data at the level of bulk human tissues, but, aside from lymphoblastoid lines, it does not include the immortalized cell lines most commonly used in experimental genomics. This mismatch limits the direct applicability of models trained on tissue-derived datasets to research carried out in standardized cell line systems such as those established by ENCODE (e.g., K562, HeLa, MCF7, HEK293).[72,73]

SpliceAI represents another class of model, using deep residual neural networks trained on large-scale human genome annotations to predict SD and SA usage from raw sequence.[74] Pangolin builds on a similar deep learning framework but incorporates training across multiple tissues, using publicly available RNA-seq transcriptome data from heart, liver, brain, and testis to provide splice site usage and variant effect predictions in a cell type aware manner.[31] While these models achieve strong performance on benchmark datasets, their predictive power is constrained by the biases of their training data and by the limited availability of experimentally validated variant-level splicing outcomes. For example, SpliceAI is trained on reference genome

annotations rather than on variant perturbations, while Pangolin leverages transcriptome data from a limited set of tissues and individuals. As such, neither model has been trained directly on the full diversity of human sequence variation. As a result, they are well-suited for predicting general *cis*-regulatory features of splicing but are less likely to capture the effects of rare or disease-associated variants.[31,74] More recently, Smith & Kitzman (2023) benchmarked multiple splicing effect predictors using MPRA data as experimental ground truth. They found that deep learning models such as SpliceAI and Pangolin achieved the highest overall sensitivity, but notable challenges remained, particularly in predicting the splicing impact of certain exonic variants.[75]

Although these models achieve strong performance on benchmark datasets, they remain limited by biases in their training data and the small number of experimentally validated splicing outcomes available. MPRAs have begun to address this gap. MaPSy measured 4,964 published disease-causing exonic mutations and similarly, Vex-Seq has measured 2,059 variants, providing a focused but relatively small set of outcomes.[39,59] On a larger scale, MFASS assayed 27,733 variants using a fluorescence-based sort-seq strategy that enables high-throughput screening but was applied primarily in a single cell type. In MFASS, exon skipping was quantified categorically rather than continuously, which may reduce sensitivity to variants with modest effects. [40] Collectively, these studies highlight both the promise and the limitations of existing high-throughput assays. Given the breadth of human genetic variation and the high proportion of mutations predicted to affect splicing, there remains a critical need for larger, cell type-resolved resources to support the training and evaluation of predictive models and to further elucidate the rules of AS in health and disease.[69,76]

To address these gaps, we developed COMPASS, a massively parallel splicing reporter assay with measurements in five human cell lines spanning more than 87,000 sequences. This work enabled systematic measurement of sequence effects on exon skipping and represents the largest splice variant MPRA performed to date.

2.3 Workflow for the COMPASS and overview

2.3.1 COMPASS design and variant selection

COMPASS is an exon skipping MPRA comprised of short human exons and their flanking introns (**Figure 2.1A**). We selected exons from the human genome 90 nt or shorter and further filtered their nucleotide sequence length to be a multiple of three to avoid destabilizing reporter transcripts through nonsense mediated decay. Exons matching these requirements were included even if they are naturally constitutively spliced, in order to broadly learn the relationship between nucleotide sequence and exon inclusion. For each exon, we also included 20 nt of the corresponding downstream intron and a variable stretch of the upstream intron such that the total length of the assayed region is 161 bp. The variable regions are flanked by constant sequences derived from SMN2 introns 6 and 7 to ensure the exon can be recognized by the splicing machinery.[30,77,78] Random 3'UTR barcodes were associated with the variable cassette exon and introns through DNA sequencing. The barcode enables mapping of each transcript to a reporter gene of origin, even when the entire variable intronic and exonic region is excised.

We mined publicly available databases to identify disease-associated and common human variants that fall within these introns and exons (**Figure 2.1B**). COMPASS contains 61 variants from Geuvadis, 5,948 from the Exome Aggregation Consortium (ExAC), 1,286 from ClinVar, and an additional 29,758 single and 48, 290 double variants randomly targeted to introns and

exons that contain at least one ClinVar variant.[79–81]

We delivered the pooled plasmid library to five different human cell lines: HEK293 (embryonic kidney), K562 (leukemia), HeLa (Adenocarcinoma, uterus), MCF-7 (Adenocarcinoma, breast), and HMC3 (microglia). We collected and sequenced the library-specific mRNA and bioinformatically determined the splicing profiles of each sequence

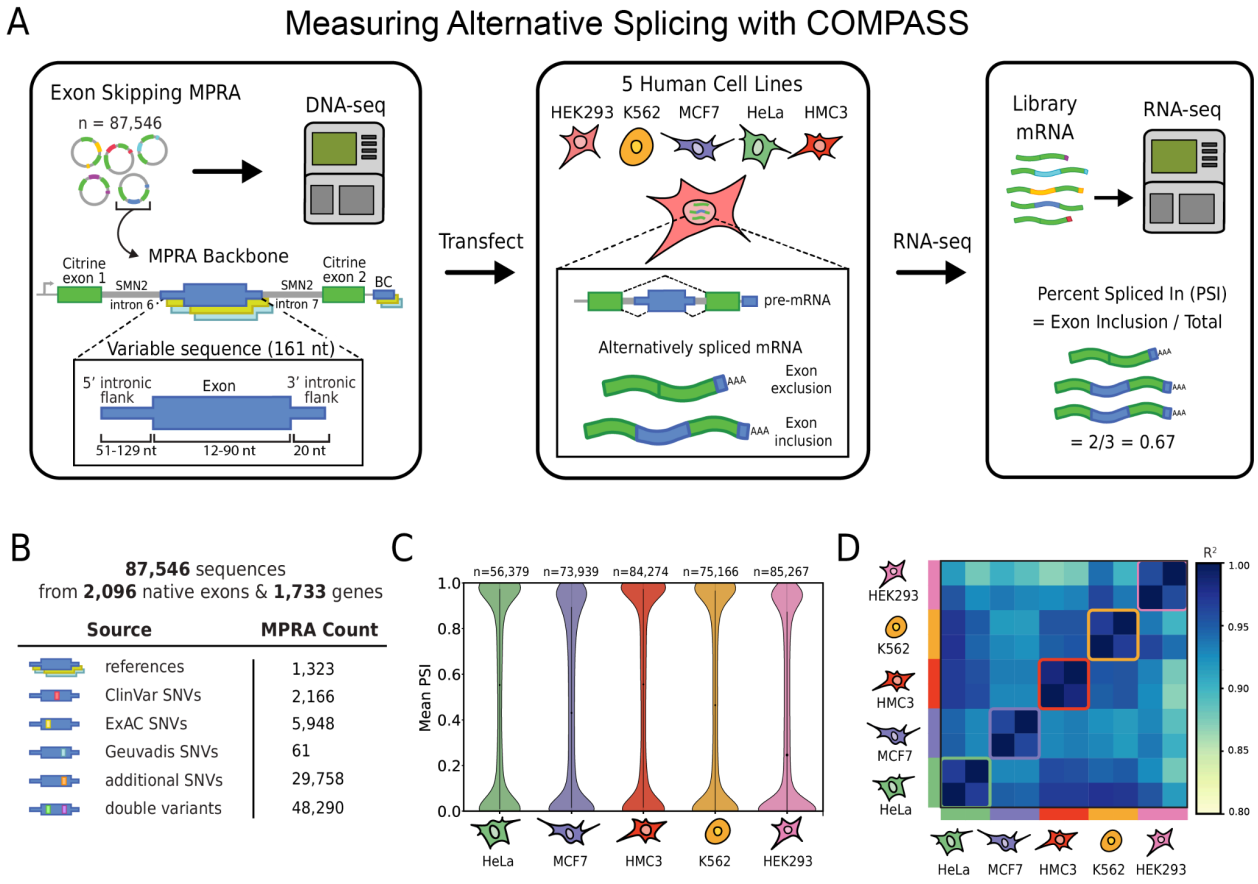


Figure 2.1 | COMPASS, an exon skipping MPRA, was performed on over 87,000 sequences in 5 human cell lines

(A) A workflow diagram of the COMPASS showing the structure of the splicing reporter containing variable and constant regions, which are transfected into 4 human cell lines (HEK293, HeLa, K562, MCF7, and HMC3), with resulting mRNA sequenced via RNA-seq. **(B)** The breakdown of the sources from which the MPRA sequences containing short exons and intronic flanks were derived. **(C)** Measured PSI distribution across library sequences in 5 cell lines. **(D)** Pearson R2 correlation of COMPASS from 5 cell lines (HEK293, HeLa, K562, MCF7, and HMC3).

as outlined in Methods, with the aid of the barcode sequence in the 3′ UTR. Sequencing across the variable exon and both intron-exon junctions enabled us to map not only expected splice isoforms but also to quantify unexpected splice events typically due to variants that create new SDs or SAs (**Figure 2.2**).

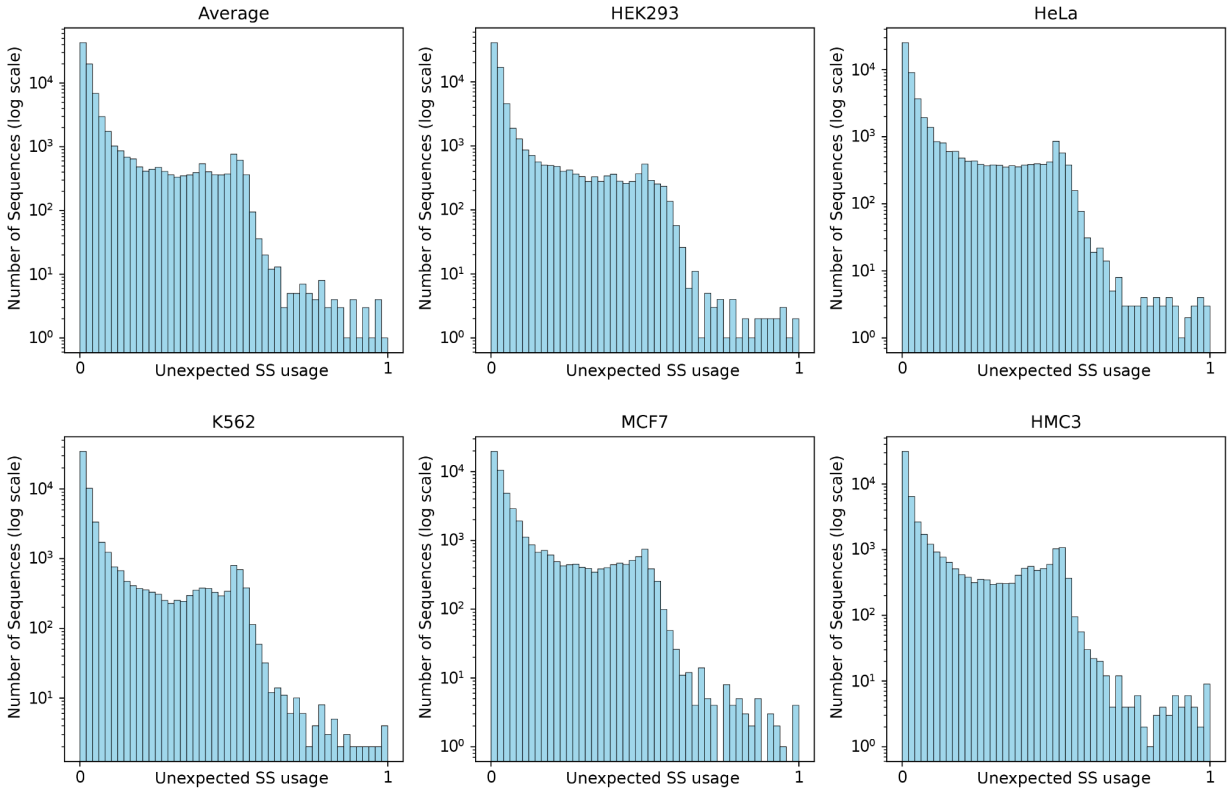


Figure 2.2 | Frequency of unexpected splice site usage

Distribution of unexpected splice site usage across COMPASS. For each sequence, the fraction of reads mapping to noncanonical splice sites was calculated within each cell line. Histograms depict the number of sequences ($\log_{10}$ scale) at a given frequency of unexpected splicing. Shown are the overall average across cell lines together with distributions for HEK293, HeLa, K562, MCF7, and HMC3.

2.3.2 Mapping of splice junctions and quantification of isoform abundance

Next, we calculate PSI of expected splice junctions, defined as the number of RNA molecules containing the expected exon divided by the total number of RNA molecules with the corresponding barcode (**Figure 2.1 A**). A majority of reporters exhibit either dominant exon inclusion or exclusion, i.e. non-alternatively spliced, with relatively fewer reporters exhibiting moderate inclusion levels consistent with AS (**Figure 2.1 C**). Splicing outcomes between cell lines are highly correlated though less than replicates from the same cell line (**Figure 2.1 D; Figure S1 A**)

2.3.3 Validation of splicing COMPASS with control experiments

To confirm that our plasmid design does not bias splicing, we tested four disease-associated exons with well-characterized skipping behavior: *SMN2* exon 7 (54 nt), *MAPT* exon 10 (93 nt), *CFTR* exon 12 (87 nt), and *DMD* exon 29 (150 nt). Exon inclusion levels in our assay closely matched published values, both by RT-PCR (r = 0.83) and RNA-seq (r = 0.85), with near-perfect agreement between the two readouts (r = 1.0) (**Figure 2.3A**). These results validate the plasmid backbone for high-throughput MPRA measurements.

A mini MPRA library of 13 sequences sampled from the full COMPASS DNA library was used to optimize the computational pipeline for PSI calculation as further outlined in methods (**Figure 2.3B**). After optimization, PSI values from the mini library correlated closely with those obtained for the same sequences in COMPASS. We further benchmarked COMPASS against the MFASS dataset, another large-scale splicing assay. For more than 350 overlapping variants, PSI values showed strong concordance between the two studies (r = 0.74) (**Figure 2.3C**), despite differences in plasmid backbone design (*SMN1* versus *SMN2* intronic sequence).

These results demonstrate that splicing MPRAs, such as COMPASS, are robust and generalizable across sequence contexts. Measurements showing weaker concordance between the two MPRAS, are likely caused by the lower accuracy and precision of the other MFASS assay, which uses binned measurements rather than a continuous PSI scale, reducing measurement resolution.

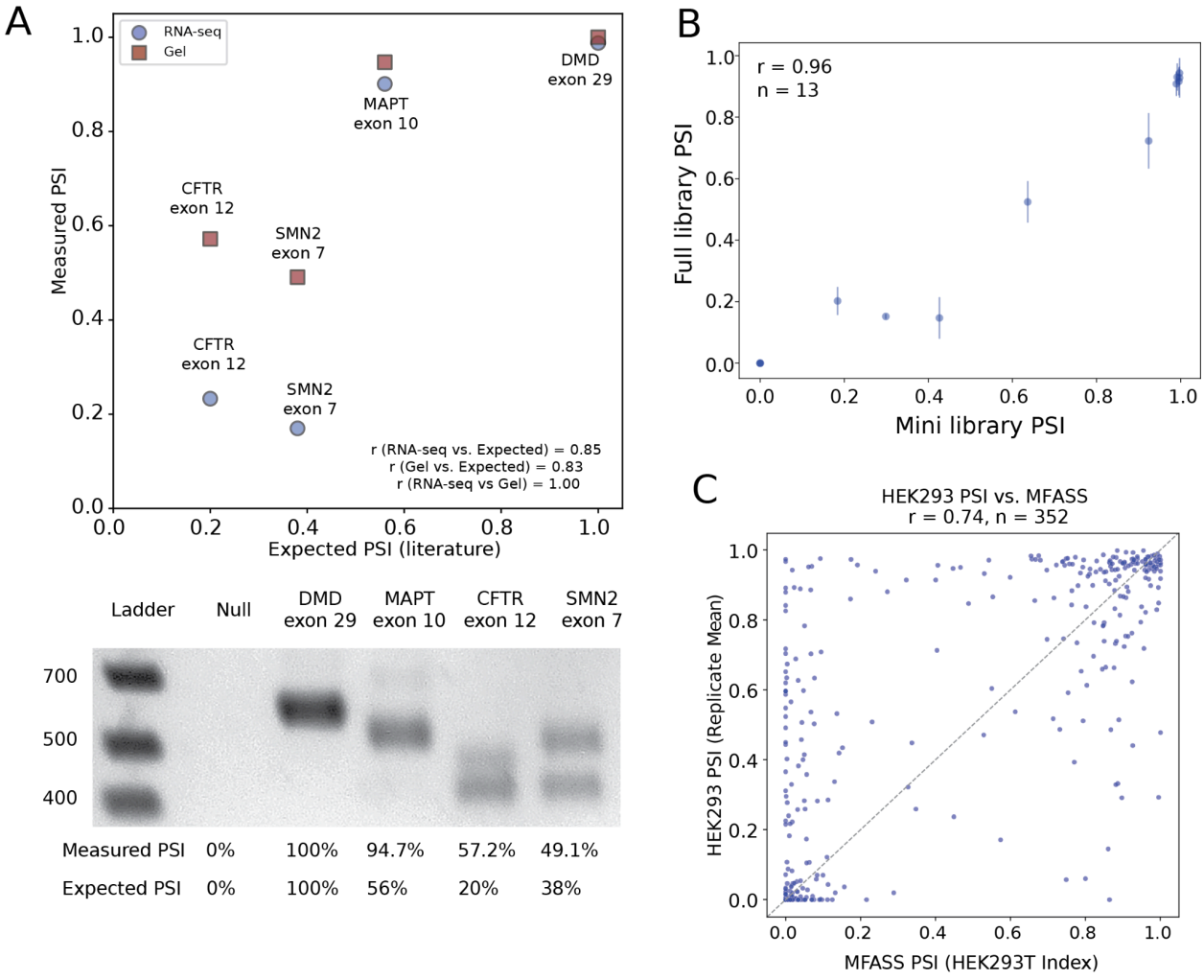


Figure 2.3 | Validation of assay design and reproducibility of splicing measurements
(A) Disease-associated exons with well-characterized splicing outcomes (*SMN2* exon 7, *MAPT* exon 10, *CFTR* exon 12, *DMD* exon 29) were tested in the COMPASS reporter backbone. PSI values measured by RT-PCR and gel electrophoresis correlated with literature values (r = 0.83), and RNA-seq of the same products further improved concordance (r = 0.85). **(B)** A 13-plasmid mini-library in HEK293 verified by RNA-sequencing, reproduced PSI values observed in COMPASS (r = 0.96). **(C)** PSI values for >350 shared variants were correlated between COMPASS and the independent MFASS assay (r = 0.74), despite differences in plasmid backbones, demonstrating robustness and generalizability of the approach.

2.4 Quantifying and predicting variant impact on splicing.

Next, we turn to characterizing the impact of SNVs (**Figure 2.4A**). We found 87,547 sequences with measurements in at least one cell type. Of these, 2,559 lacked a matched reference, precluding calculation of variant-associated changes. In total, 83,852 variants remained with an associated reference, allowing calculation of ΔPSI, defined as the difference between the variant and reference PSI. Replicates were highly correlated not only at the PSI level but also at the ΔPSI level (**Figure S1A**). PSIs and subsequently ΔPSIs were determined from the.pooled included and excluded reads across replicates, requiring a minimum of 10 reads per replicate and at least two replicates per cell line. In this chapter, values were averaged across cell lines to emphasize sequence-level effects, whereas individual cell line effects are examined separately in Chapter 3 to characterize cell type specificity.

A variant that increases exon inclusion, thus is associated with a positive ΔPSI. While easy to interpret, ΔPSI is highly sensitive to the reference PSI value (**Figure 2.4B**). Measured ΔPSI values tend to be small if the reference exon is fully spliced in or out, as in these cases even a large perturbation may not be enough to dislodge the exon from its preferred state. Conversely, for alternatively spliced exons where the competing SDs and SAs are relatively balanced, comparably weaker perturbations can result in large PSI changes. This dependence on the starting inclusion level means that a large ΔPSI in the reporter context may not correspond to a similarly large ΔPSI in the native gene context. Although the splice regulatory signals in the central exon and near intronic context are unchanged, the competing distal splice signals, i.e. those used when the exon of interest is spliced out, vary between reporter and native context.

For most of our analysis, we use the delta log odds ratio (Δlogit) as a more robust and generalizable metric for variant impact (**Figure 2.4 C**).[30,32,33,82–85] Variants that increase exon

inclusion have a positive Δlogit(PSI). As noted by Baeza-Centurion *et al.*, the odds ratio has an intuitive biophysical interpretation: it can be thought of as a multiplicative factor that enhances or attenuates the rate of inclusion in a simple model of competition between exon inclusion and exclusion.[85] Consistent with the biophysical model described previously, the logit values are less sensitive to starting PSI and are more evenly distributed (**Figure 2.4C**). While the Δlogit(PSI) is thus a useful measure of variant impact, practical applications ultimately require knowledge of the variant ΔPSI in the native gene context. These values can be estimated with a biophysical model given knowledge of the reference PSI obtained from RNA sequencing data and the corresponding variant logit measured in the reporter construct.[83]

Still, we note that in practice logits are an imperfect metric, as they are highly sensitive to sequencing depth, as discussed in **Methods**. To mitigate this and avoid artifacts introduced by logit transformation, PSI values were clipped prior to transformation, as is standard in Δlogit(PSI) analyses.[32,33,84,86,87] A clipping interval of [0.01, 0.99] was selected empirically as outlined in the **Methods** and is consistent with FRASER.[84] Without clipping, PSI estimates become strongly dependent on read depth, particularly when the reference and variant used in the calculation are not sequenced to comparable coverage, as outlined in **Methods**.

To examine how positional context influences variant effects, we compared Δlogit(PSI) values relative to the position of the SD and SA sites. For each variant, the Δlogit(PSI) was plotted at its position with respect to the nearest splice site. Variants located farther from the splice sites thus reflect longer exon lengths, since the most distant positions occur only in longer exons (**Figure 2.4D**). This visualization highlights that variants disrupting SDs and SAs tend to have the most pronounced and most negative effect (**Figure 2.4D**). However, variants with large negative or positive effect sizes can be found anywhere in the assayed introns and exons,

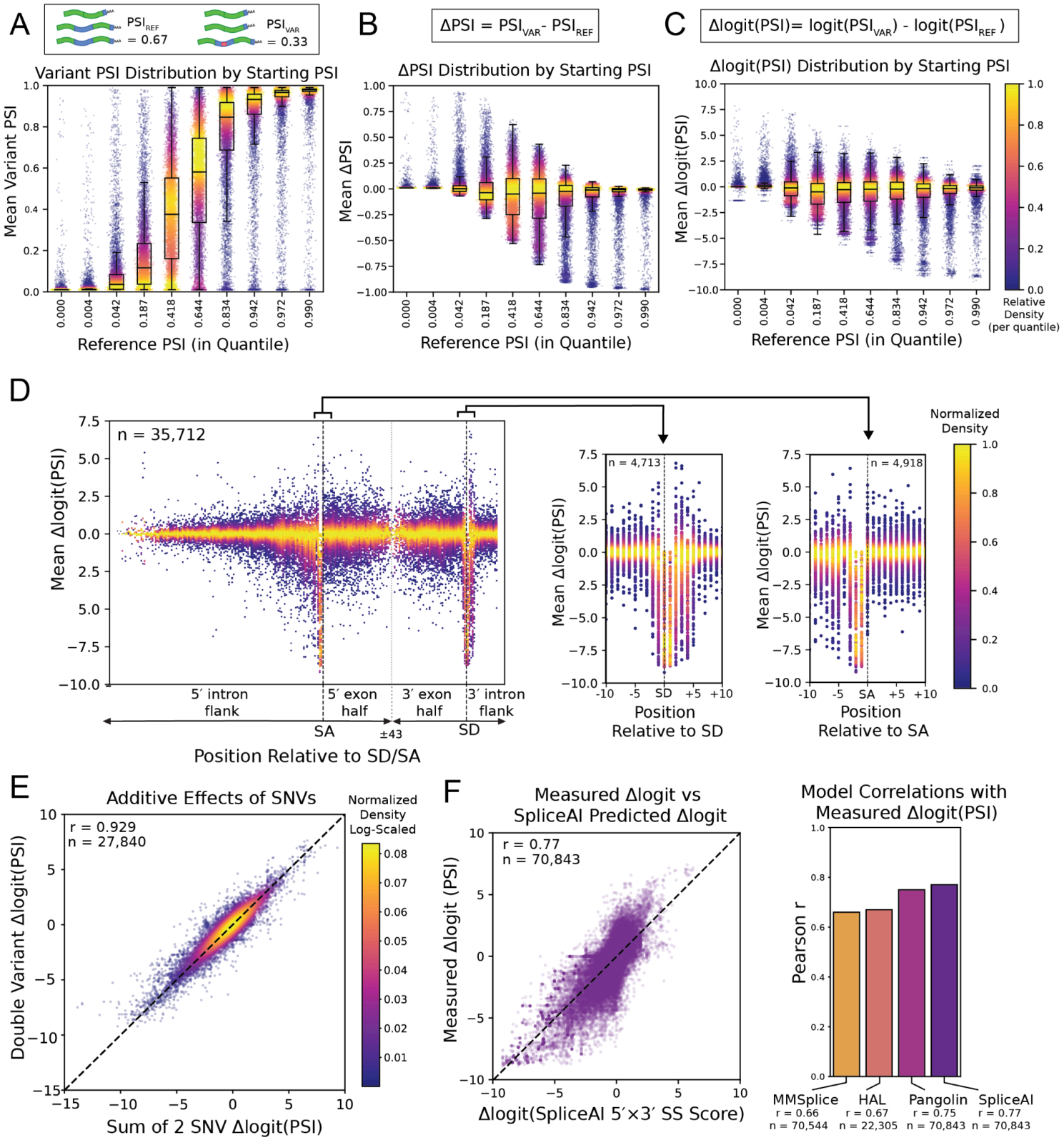

A
PSI_REF = 0.67
PSI_VAR = 0.33
Variant PSI Distribution by Starting PSI
Mean Variant PSI
Reference PSI (in Quantile)
B
ΔPSI = PSI_VAR - PSI_REF
ΔPSI Distribution by Starting PSI
Mean ΔPSI
Reference PSI (in Quantile)
C
Δlogit(PSI)= logit(PSI_VAR) - logit(PSI_REF)
Δlogit(PSI) Distribution by Starting PSI
Mean Δlogit(PSI)
Reference PSI (in Quantile)
Relative Density (per quantile)
D
n = 35,712
Mean Δlogit(PSI)
5′ intron flank
5′ exon half
3′ exon half
3′ intron flank
SA
±43
SD
Position Relative to SD/SA
n = 4,713
Position Relative to SD
n = 4,918
Position Relative to SA
Normalized Density
E
Additive Effects of SNVs
r = 0.929
n = 27,840
Double Variant Δlogit(PSI)
Sum of 2 SNV Δlogit(PSI)
Normalized Density Log-Scaled
F
Measured Δlogit vs SpliceAI Predicted Δlogit
r = 0.77
n = 70,843
Measured Δlogit (PSI)
Δlogit(SpliceAI 5′×3′ SS Score)
Model Correlations with Measured Δlogit(PSI)
Pearson r
MMSplice r = 0.66 n = 70,544
HAL r = 0.67 n = 22,305
Pangolin r = 0.75 n = 70,843
SpliceAI r = 0.77 n = 70,843

Figure 2.4 | Global splicing effects of single and double variants
(A) Variant PSI values grouped into ten quantiles based on the reference (starting) PSI, showing the distribution of single and double variants within each quantile. **(B)** ΔPSI values grouped into the same reference PSI quantiles, highlighting how variant-induced splicing changes scale with the baseline inclusion level. **(C)** Δlogit(PSI) values grouped by reference PSI quantiles, showing that the logit transformation attenuates the dependence of variant effects on the starting PSI. In all panels, densities are normalized within quantiles to facilitate comparison of relative effects. **(D)** Distribution of Δlogit(PSI) as a function of SNV position relative to the SA or SD (n = 35,712). Densities are normalized per position and show an enrichment of large-magnitude effects in proximity to the SD and SA. Insets to the right highlight the regions spanning 10 nucleotides upstream and downstream of each splice site. **(E)** Additive effects of SNVs, comparing the summed Δlogit(PSI) values of two individual variants with the measured Δlogit(PSI) of the corresponding double variant carrying both changes (n = 27,840, Pearson r = 0.929). **(F)** Predictive performance of computational models, comparing experimental Δlogit(PSI) with SpliceAI predictions. For each sequence, the SD probability at the SD site and the SA probability at the SA site were multiplied to obtain a splicing probability, which was then treated as a predicted PSI. Predicted Δlogit(PSI) was computed using the same logit transformation applied to experimental data, allowing direct comparison. SpliceAI achieved the strongest performance (n = 70,843, Pearson r = 0.77), capturing the greatest number of variants and showing the best agreement with measured effects. Pangolin performed to a similar level (n=70,843 Pearson r = 0.75) and captured the same number of variants as Splice AI. For comparison, MMSplice, which is limited to variants within 100 nucleotides of the variable exon, achieved Pearson r = 0.66 (n = 70,544), and HAL, which is limited to exonic variants, achieved Pearson r = 0.67 (n = 22,305).

consistent with a *cis*-regulatory code that extends beyond core elements.

We next turned to analysis of double variants and asked whether double variant effects could be explained by those of the two variants individually. We find that variant effects are almost perfectly additive in logit space and therefore independent (**Figure 2.4E**). This result is consistent with an earlier observation that effect sizes associated with short k-mer motifs are additive in log space.[30]

Next, we asked whether our measurements could be predicted with existing splicing models. For this analysis, we excluded variants where the associated reference has exactly 0 or 1 PSI. Variants that disrupt SDs or SAs in these cases have no measurable impact and therefore,

the measured Δlogit is exactly zero. Even though these measurements are zero, models will predict a non zero effect because their predictions are agnostic to the reference PSI. Experimentally, these effects are not observable, but they could become observable in alternative reporter designs, or if sequenced to extremely high sequencing depth.

After excluding variants where the associated reference PSI is 0 or 1, we found that splicing predictors HAL, MMSplice, Pangolin, and SpliceAI all could predict variant impact with high accuracy (**Figure 2.4F**). [30–32,34] SpliceAI in particular performed well on both intron and exonic single and double variants with a Pearson r of 0.77 (**Figure 2.4F, left**). For SpliceAI the predicted PSI was calculated as the product of the donor probability at the expected SD and the acceptor probability at the expected SA on the variable exon.[88] We observed performance with Pangolin comparable to SpliceAI, as expected given their similar model architecture, with the main distinction that Pangolin was trained on quantitative data from four different tissues.[31] Older and simpler models such MMSplice and HAL also achieved good accuracy, albeit HAL can only predict exonic variants

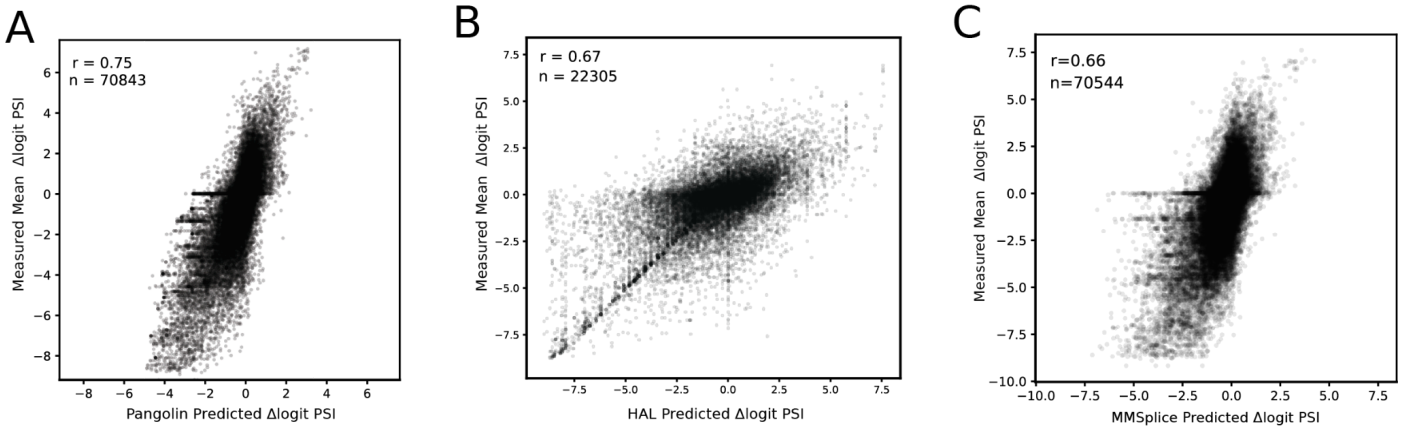


Figure 2.5 | Correlation of experimental Δlogit(PSI) with predictions from AS models
Scatter plots compare measured Δlogit(PSI) values to model predictions for **(A)** Pangolin (r = 0.75, n = 70,843), **(B)** HAL (r = 0.67, n = 22,305), and **(C)** MMSplice (r = 0.66, n = 70,544).[30,31,89]

(Figure 2.4F, right). [30,32] These results suggest that current splice predictors have learned a fairly accurate representation at least of those aspects of the splicing code that are similar across cell types. The full correlation plots for Pangolin, HAL, and MMSplice are provided in **Figure 2. 5**.

2.5 COMPASS measurements identify splice-disrupting variants and suggest potential modes of pathogenicity for variants of uncertain significance.

2.5.1 Splicing impact of ClinVar- and dbSNP-annotated variants in COMPASS

We next turn to the analysis of measured variants with annotations in ClinVar (**Figure 2.6A**). We observe that variants classified as benign or likely benign (B/LB) generally have a limited impact on splicing, with most Δlogit(PSI) values being close to zero. In contrast, several pathogenic or likely pathogenic (P/LP) variants have a large negative Δlogit(PSI), suggesting that they might act by disrupting splicing; the two highlighted variants in the pathogenic category both disrupt core splice signals (**Figure 2.6A**). The majority of variants of conflicting or uncertain significance (CL/VUS) have small effect sizes, indicating that they are likely benign or at least do not act by disrupting splicing. However, there are also several CL or VUSs variants with large negative or positive effect sizes (**Figure 2.6A**). Because of their strong molecular phenotypes, these variants are prime candidates for further investigation as they suggest a plausible mode of pathogenicity through altered splicing. The highlighted, large effect size variants occur in disease-associated genes and are either intronic or disrupt core splicing signals (**Figure 2.6B**). Finally, **Figure 2.6C** shows a subset of measured variants that have annotations in ClinVar and that also occur in dbSNP and for which we thus have minor allele frequency (MAF) information. As expected, most pathogenic and likely pathogenic variants are rare while most common variants are benign. It is also notable that only few rare variants are labeled as

benign though many are likely benign. P/LP variants were observed only in the rare group, and CL/VUS variants were also enriched among rare alleles. B/LB classifications made up the highest relative proportion of sequences in the common group, consistent with pathogenic and uncertain annotations being concentrated in rare variants.

2.5.2 Near-site saturation mutagenesis of disease-relevant exons

For exons and flanking introns from disease-relevant genes, we generated near-site saturation mutagenesis (NSSM) maps. For this purpose, we define an exon family as a reference sequence (an exon with its flanking introns) together with all associated variants. Each NSSM map is represented by at least 150 SNVs in at least one cell line. In total, 107 exon families met this threshold, with variant coverage for each sequence spanning nearly one-third of all possible substitutions and likely perturbing most or all *cis*-regulatory elements in the exon family. Position-based heatmaps were used to visualize the effects of SNVs on exon inclusion across all measured cell lines. Heatmaps display the average Δlogit(PSI) values across measured cell lines for each SNV, with axes indicating the SNV position and nucleotide change. We highlight exons from the ACMG list of clinically actionable genes, including exon 42 of MYH11 (familial thoracic aortic aneurysm 4), exon 10 of LMNA (dilated cardiomyopathy 1A), and exon 24 of SCN5A (cardiac arrhythmia syndromes) (**Figure 2.6D-F**). These three exons are featured because they are naturally alternatively spliced and are thus susceptible to SDVs. A larger set of NSSM maps includes additional ACMG exons such as *MUTYH* exon 6, *TNNT2* exons 4 and 5, *RYR1* exon 94, *TSC2* exon 32, and *MYH11* exon 6 as well as exons with a large number of pathogenic or VUS variants (**Figure 2.7**). All maps span 161 nucleotides, with truncation applied

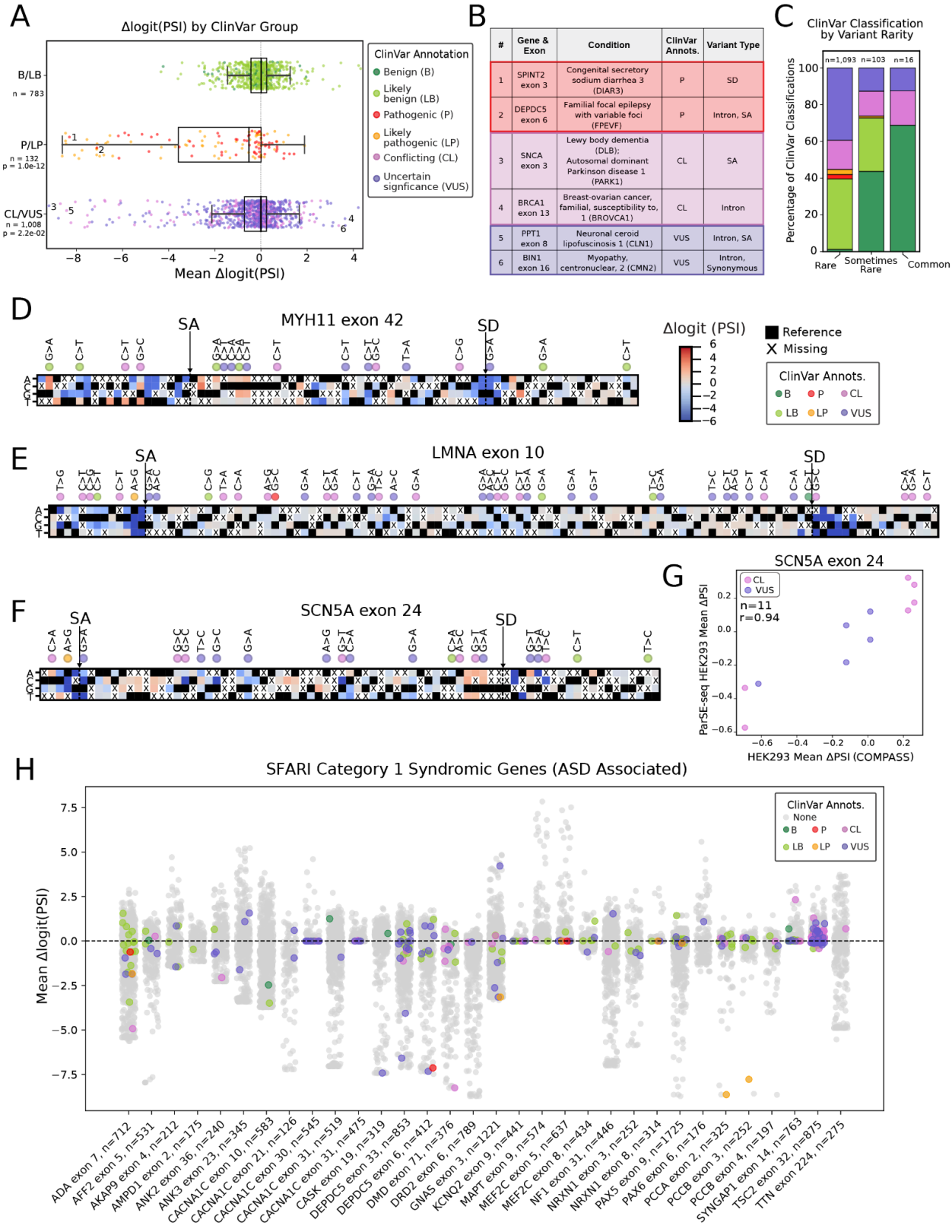

A
Δlogit(PSI) by ClinVar Group
B/LB
n = 783
P/LP
n = 132
p = 1.0e-12
CL/VUS
n = 1,008
p = 2.2e-02
Mean Δlogit(PSI)
ClinVar Annotation
Benign (B)
Likely benign (LB)
Pathogenic (P)
Likely pathogenic (LP)
Conflicting (CL)
Uncertain signficance (VUS)
B
| Gene & Exon | Condition | ClinVar Annots. | Variant Type
1 | SPINT2 exon 3 | Congenital secretory sodium diarrhea 3 (DIAR3) | P | SD
2 | DEPDC5 exon 6 | Familial focal epilepsy with variable foci (FPEVF) | P | Intron, SA
3 | SNCA exon 3 | Lewy body dementia (DLB); Autosomal dominant Parkinson disease 1 (PARK1) | CL | SA
4 | BRCA1 exon 13 | Breast-ovarian cancer, familial, susceptibility to, 1 (BROVCA1) | CL | Intron
5 | PPT1 exon 8 | Neuronal ceroid lipofuscinosis 1 (CLN1) | VUS | Intron, SA
6 | BIN1 exon 16 | Myopathy, centronuclear, 2 (CMN2) | VUS | Intron, Synonymous
C
ClinVar Classification by Variant Rarity
Percentage of ClinVar Classifications
n=1,093
n=103
n=16
Rare
Sometimes Rare
Common
D
MYH11 exon 42
SA
SD
Δlogit (PSI)
Reference
Missing
ClinVar Annots.
B
P
CL
LB
LP
VUS
E
LMNA exon 10
SA
SD
F
SCN5A exon 24
SA
SD
G
SCN5A exon 24
CL
VUS
n=11
r=0.94
ParSE-seq HEK293 Mean ΔPSI
HEK293 Mean ΔPSI (COMPASS)
H
SFARI Category 1 Syndromic Genes (ASD Associated)
Mean Δlogit(PSI)
ClinVar Annots.
None
B
P
CL
LB
LP
VUS
ADA exon 7, n=712
AFF2 exon 5, n=531
AKAP9 exon 4, n=212
AMPD1 exon 2, n=175
ANK2 exon 36, n=240
ANK3 exon 23, n=345
CACNA1C exon 10, n=583
CACNA1C exon 21, n=126
CACNA1C exon 30, n=545
CACNA1C exon 31, n=519
CACNA1C exon 31, n=475
CASK exon 19, n=319
DEPDC5 exon 33, n=853
DEPDC5 exon 6, n=412
DMD exon 71, n=376
DRD2 exon 6, n=789
GNAS exon 3, n=1221
KCNQ2 exon 9, n=441
MAPT exon 9, n=574
MEF2C exon 5, n=637
MEF2C exon 8, n=434
NF1 exon 31, n=446
NRXN1 exon 3, n=252
NRXN1 exon 8, n=314
PAX5 exon 9, n=1725
PAX6 exon 6, n=176
PCCA exon 2, n=325
PCCB exon 3, n=252
PCCB exon 4, n=197
SYNGAP1 exon 14, n=763
TSC2 exon 32, n=875
TTN exon 224, n=275

Figure 2.6 | Splicing-associated effects of ClinVar-annotated population variants
(A) Distribution of average Δlogit(PSI) for variants present in all lines. Variants were grouped by ClinVar classification: benign/likely benign (B/LB, n = 783), pathogenic/likely pathogenic (P/LP, n = 132), and conflicting/uncertain significance (CL/VUS, n = 1008). B/LB variants exhibit consistently lower Δlogit(PSI) magnitudes, whereas P/LP and CL/VUS variants span broader ranges, with many high-magnitude outliers. Fisher's exact test confirmed significant differences between B/LB and P/LP (p value = 1.0×10^{-12}) and between B/LB and CL/VUS (p value = 2.2×10^{-2}), consistent with splicing disruption contributing to pathogenic and uncertain classifications. These results confirm that P/LP variants are frequently associated with aberrant splicing and suggest that a subset of CL/VUS variants may exert pathogenic effects through similar mechanisms. **(B)** Six representative variants from 3A are highlighted: two pathogenic, two VUS, and two CL. Each is associated with a splicing disruption, including SD/SA disruption or intronic variants annotated in ClinVar. **(C)** ClinVar classifications stratified by allele frequency categories from ClinVar dbSNP155: rare (n = 1,093), sometimes rare (n = 103), and common (n = 16). Rare variants in dbSNP155 are defined by having a MAF of less than 1%. Common variants are those with a MAF of at least 1%. Sometimes rare variants have a MAF of less than 1% in some, but not all, reporting projects. **(D, E, F)** Near-site saturation mutagenesis (NSSM) plots span the assayed sequence, with the three possible substitutions at each position colored by their measured Δlogit(PSI), alongside the reference allele (black). Positions marked with an "X" indicate variants absent from the dataset. Shown are sections of NSSM maps for **(D)** *MYH11* exon 42, **(E)** *SCN5A* exon 24, and **(F)** *LMNA* exon 10 **,** all of which are ACMG-listed genes. ClinVar annotations displayed above the heatmaps for variants of known clinical significance. **(G)** PSI values from our assay were compared with ParSE-seq, which measures splicing of *SCN5A* exons in HEK293 cells using a rat insulin minigene reporter. For ClinVar CL and VUS variants in *SCN5A* exon 24 (n = 11), the two assays were highly concordant (Pearson r = 0.94). **(H)** Variants from SFARI category 1 syndromic genes (1S). Of the 240 1S genes in the SFARI Gene database, 32 exons from 24 of these genes were represented in our dataset. For these exons, we show the distribution of variant effects on splicing, with ClinVar-annotated SNVs (n = 309) highlighted. This overlap emphasizes the contribution of splicing disruption to pathogenic mechanisms in autism-associated genes.

in **Figure 2.6** for easier visualization; the full-length NSSM maps are shown in **Figure 2.7.**

The *MYH11* gene encodes smooth muscle myosin heavy chain (SM-MHC) and gives rise to two principal isoforms, SM1 and SM2, through AS at exon 42.[90,91] Exon 42 is included in the SM2 protein isoform but not in SM1, resulting in a shorter, alternative protein C terminus as the exon encodes for only 9 amino acids followed by a stop codon in the *MYH11* reading frame (albeit not in that of the reporter).[92] SM1 expression is broadly detected from fetal development through adulthood, whereas SM2 appears more restricted to well-differentiated smooth muscle

cells and to specific developmental contexts such as fetal blood vessle. formation.[93,94] A number of pathogenic splice variants have been reported across multiple exons in this gene.[90,91,95,96] In COMPASS, we observe multiple variants with conflicting and uncertain significance in *MYH11* exon 42 with variable effects on splicing. Mutations with conflicting annotations that disrupt a polyC stretch starting at position 105 of the NSSM map (**Figure 2.6D**) are particularly intriguing because they have a strong positive effect size. However, these variants are interspersed with benign mutations of similar effect size suggesting that they are likely benign themselves in spite of their potential to modulate exon inclusion, possibly because the C terminal is not the only distinguishing feature between isoform SM1 and SM2 and both have similar functions. Consistent with prior reports, exon 42 harbors putative binding motifs for the poly(C)-binding proteins PCBP1 and PCBP2, which may underlie its sensitivity to mutations in this region.[97] Taken together, the splicing of *MYH11* exon 42 highlights both isoform diversification in smooth muscle biology and a potential mutational hotspot.

Nuclear envelope proteins lamin A and C are both derived from the *LMNA* gene through AS. A third common splice isoform, lamin AΔ10 results from skipping of exon 10 but is otherwise identical to lamin A and has no defined function.[98] In our measurements, we show *LMNA* exon 10 contains a large number of VUSs but measurements reveal that most have relatively small effect sizes, suggesting that they are unlikely to result in exon skipping **(Figure 2.6E**). Still, it cannot be excluded that some of the (non-synonymous) variants are pathogenic but directly modify protein stability or folding rather than splicing, similar to the G>C mutation at position 68 of the NSSM heatmap. We also highlight our near-saturation results for *SCN5A* exon 24, extending the recent ParSE-seq MPRA of this arrhythmia-associated gene. *SCN5A* encodes

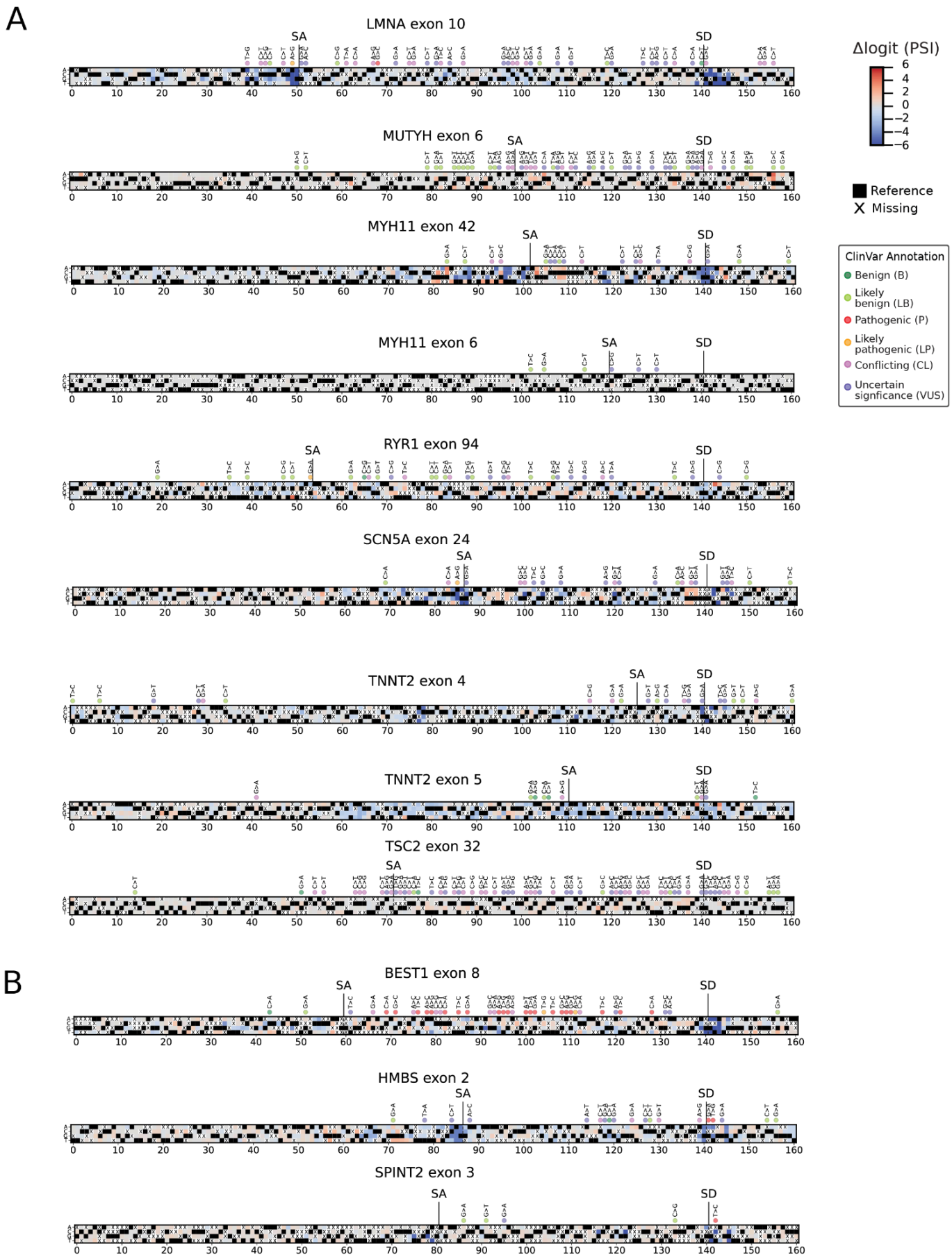

A
LMNA exon 10
MUTYH exon 6
MYH11 exon 42
MYH11 exon 6
RYR1 exon 94
SCN5A exon 24
TNNT2 exon 4
TNNT2 exon 5
TSC2 exon 32
B
BEST1 exon 8
HMBS exon 2
SPINT2 exon 3
SA
SD
Δlogit (PSI)
Reference
Missing
ClinVar Annotation
Benign (B)
Likely benign (LB)
Pathogenic (P)
Likely pathogenic (LP)
Conflicting (CL)
Uncertain signficance (VUS)

Figure 2.7 | Complete near-saturation mutagenesis (NSSM) maps of exons in disease-relevant genes
(A) Shown are NSSM maps spanning the full assayed sequence, with each position represented by the three possible nucleotide substitutions colored by their measured Δlogit(PSI) alongside the reference allele in black. Positions marked with an "X" indicate variants absent from the dataset. ClinVar annotations are displayed above the heatmaps for variants of known clinical significance. The full NSSM maps are shown here for *MYH11* exon 42, *SCN5A* exon 24, and *LMNA* exon 10, which are displayed in truncated form in Figure 3D-F. Additional exons from ACMG SF v3.2 listed genes are shown, such as *MUTYH* exon 6, *MYH11* exon 6, RYR1 exon 94, *TNNT2* exon 4 and 5, and *TSC2* exon 32. (**B)** Three exons not on the ACMG list but containing a large number of pathogenic or VUS variants are likewise displayed: *BEST1* exon 8, *HMBS* exon 2, and *SPINT2* exon 3.

the cardiac sodium channel NaV1.5 and its loss-of-function is the leading monogenic cause of the arrhythmia disorder Brugada syndrome (BrS) as well as other cardiac arrhythmias.[99–101] Prior low-throughput studies in the *SCN5A* gene have identified intronic and exonic SDVs, which have been associated with BrS.[99,102–104] COMPASS revealed a spectrum of SDVs, encompassing both variants overlapping ClinVar annotations, including a pathogenic variant in the SA, and additional SDVs not previously annotated in ClinVar (**Figure 2.6F**). COMPASS measurements also showed strong concordance with the same measurements from the ParSE-seq MPRA in HEK293 cells (**Figure 2.6G**).

Although many NSSMs do capture nearly the entire mutational landscape, by examining predictors that perform well on our measured data, such as SpliceAI, we can interpolate the missing positions to generate complete saturation mutagenesis maps (**Figure 2.8**) However, as illustrated in other cases (**Figure 2.9**), the model sometimes predicts spurious effects or miss true splicing effects seen in experiments, underscoring that they are not yet a substitute for direct experimental measurement. In this regard, MPRAs provide a powerful tool for systematically validating and refining computational predictions.

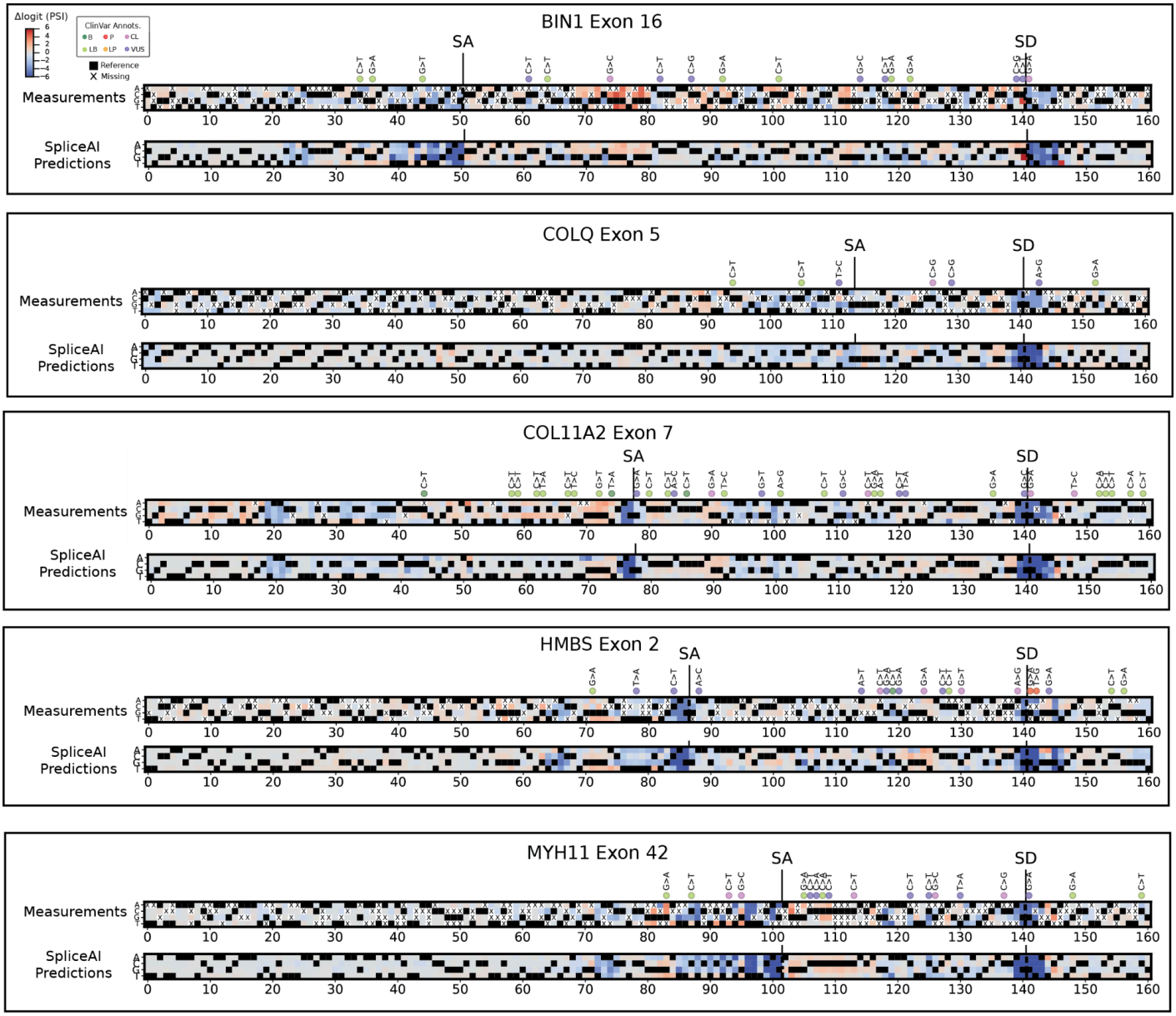


Figure 2.8 | Predictions of SpliceAI compared to COMPASS NSSM maps
Shown are some NSSM plot examples spanning the full assayed sequence, with each position represented by the three possible nucleotide substitutions colored by their measured Δlogit(PSI), alongside the reference allele in black. Positions marked with an "X" indicate variants absent from the COMPASS dataset. ClinVar annotations are displayed above the heatmaps for variants of known clinical significance. Below each experimental map, SpliceAI predictions for the same sequence are shown. In these examples, SpliceAI closely recapitulates the position- and allele-specific effects observed experimentally while providing complete coverage across all possible substitutions.

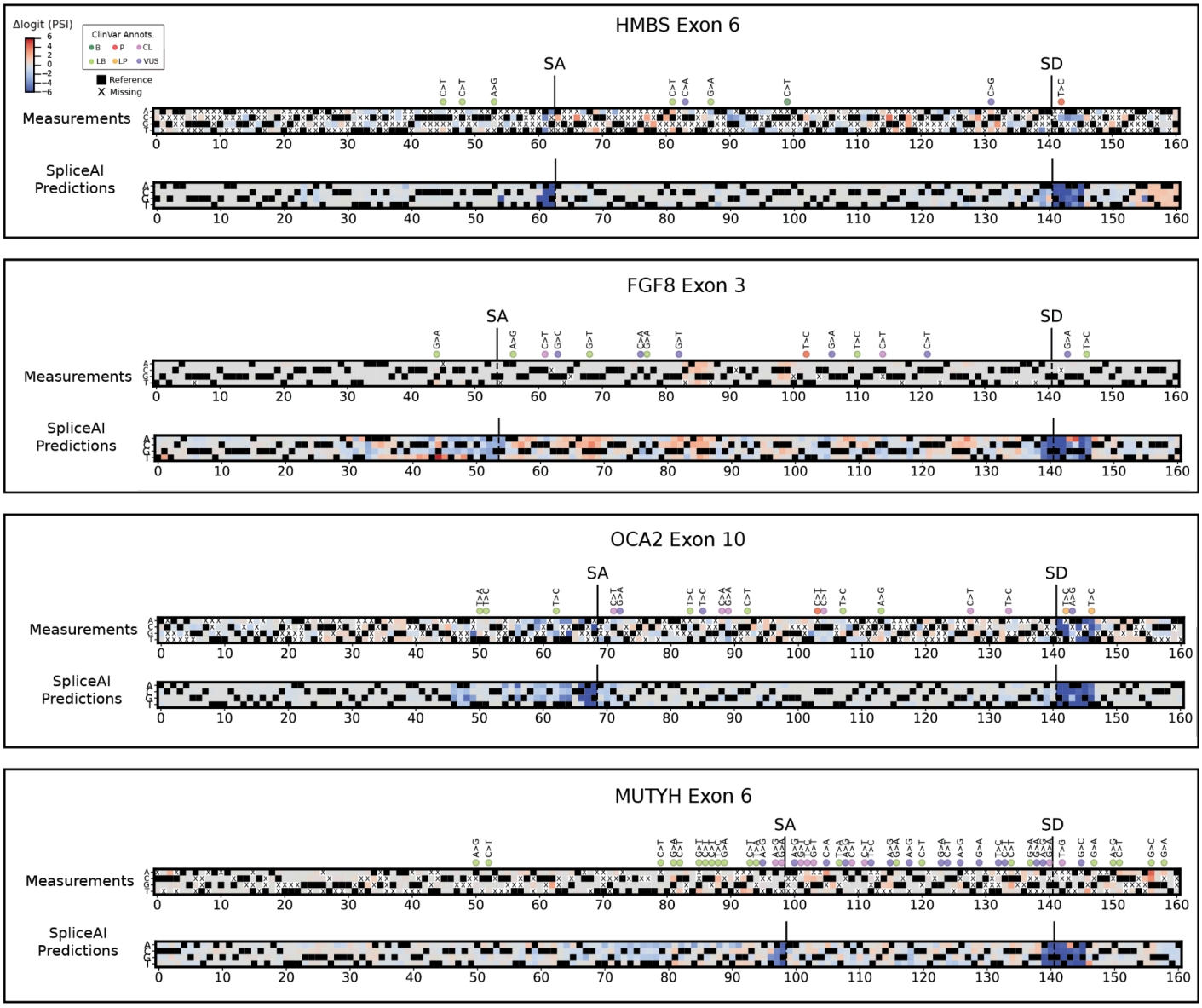


Figure 2.9 | Limitations of model-based predictions shown by NSSM maps
Shown are NSSM plot examples spanning the full assayed sequence, with each position represented by the three possible nucleotide substitutions colored by their measured Δlogit(PSI), alongside the reference allele in black. Positions marked with an "X" indicate variants absent from the COMPASS dataset. ClinVar annotations are displayed above the heatmaps for variants of known clinical significance. Below each experimental heatmap, SpliceAI predictions for the same sequence are shown. In these cases, the model predicts spurious SD or SA effects not supported by the data, or fails to capture variant-driven splicing changes that are observed experimentally. These results illustrate that models can help bridge gaps in mutagenesis efforts but are not yet a substitute for direct experimental measurement.

Finally, we turn to a set of genes strongly associated with Autism Spectrum Disorder (ASD). We focus on exons within genes that have been clearly implicated in ASD and are categorized as syndromic category 1 genes in the SFARI gene database.[105] In COMPASS, we found that 32 exons from 24 genes fall into this category, with the number of variants per exon ranging from 126 to 1,221 (**Figure 2.6H**). Across the gene panel, we observe a large number of SDVs that have not yet been annotated in ClinVar but that are promising targets for further investigation because of their strong splicing phenotypes. Of those variants with existing ClinVar annotations, a pathogenic variant in *DEPDC5* exon 6 has a strong negative impact on splicing. Notably, a VUS in the same gene has a similar impact suggesting a possible reclassification. Likely pathogenic variants that are strongly splice-disrupting are found in *PCCA* exon 2 and *PCCB* exon 3.

Conversely, pathogenic variants in *ADA* exon 7 and *MEF2C* exon 5 have minimal impact on splicing and likely act through different mechanisms. Notably, *MEF2C* exon 5 has a reference PSI near zero, meaning that only positive effects that increase exon inclusion can be practically detected without unlimited sequencing depth. Similar circumstances occur in other exons, and we document these cases (**Supplementary Table 1)**, which lists near-saturation exons alongside their starting PSI and the range of observed Δlogit values. The large scale of COMPASS enables similar visualizations for many additional gene categories beyond the SFARI ASD set. Genes from the ACMG SF v3.2 list, COSMIC Tier 1 cancer genes, transcription factors, and cytokines/cytokine receptors are shown in **Appendix 1**.[106–109] Additionally, we highlight "hotspot" exons, which are exons enriched with and susceptible to SDVs (**Appendix 1**).[110]

An obvious question raised by this variant analysis is whether we can identify putative mechanisms that might explain impacts on splicing. One hypothesis that I explore further in

Chapter 3 is that variants have a large effect because they disrupt a *cis*-regulatory element important for splicing. For example, it is plausible that the polyC stretch in *MYH11* exon 42 is a binding motif for an RNA binding protein (RBP). However, outside of SDs and SAs, recognizing functional RBP binding sites and their cognate RBPs can be challenging.

2.5.3 ClinVar variant validation by prime editing in native genomic context

To validate MPRA-derived splicing effects in an endogenous setting, we implemented a prime editing (PE) workflow in HEK293 cells. In this workflow, a genomic target is selected, the desired variant is introduced by prime editing, and editing efficiency is quantified by sequencing genomic DNA. In parallel, splicing outcomes are measured by sequencing mRNA from the same population. These measurements are compared directly to the native unedited control PSI to determine the splicing change driven by the variant in its genomic context (**Figure 2.10A**).

BIN1 exon 16 (referred to as exon 13 in some transcripts) undergoes alternative splicing and has been shown to play a role in the onset of cancer, Alzheimer's disease, and cardiac pathology.[111–114] To further validate functional splicing consequences at this locus, we selected two ClinVar variants in *BIN1* exon 16 family with conflicting pathogenicity annotations, based on their strong but opposite splicing effects in the MPRA (**Figure 2.10B**). The chr2:127,051,220:C>G variant in BIN1 has conflicting ClinVar classifications of LB and VUS, the latter reflecting insufficient evidence to either confirm or rule out pathogenicity. This variant (variant 1) disrupts a poly-G tract and increases exon inclusion in our MPRA. The chr2:127,051,153:C>T variant in BIN1, affecting the donor splice site of intron 16, has conflicting ClinVar classifications ranging from VUS to LP, with the latter submission noting splice site disruption. This variant (variant 2) reduces exon inclusion in COMPASS. PE

introduced each variant at the endogenous locus, and genomic sequencing confirmed high editing efficiency for variant 1 (>90%) and lower efficiency for variant 2 (~28%) (**Figure 2.10 C,D**).

To understand RNA splicing of this exon and the effects of the targeted variants in the native context, we first measured the reference PSI from unedited HEK293 cells to establish the endogenous inclusion level of *BIN1* exon 16. These measurements can be compared to the reference PSI already obtained from COMPASS in HEK293 cells (**Figure 2.10E**). We then introduced each variant with PE and directly measured the variant PSI in the edited HEK293 populations. To make the COMPASS values directly comparable to the PE measurements, which occur in a mixed population of edited and unedited alleles, we also generated an edit-corrected COMPASS variant PSI, representing the expected COMPASS PSI value under the observed PE editing efficiency. This corrected value provided an accurate benchmark for evaluating whether the splicing effects observed through PE matched those measured by COMPASS (**Figure 2.10F**). We also compared ΔPSI values in both contexts (**Figure 2.10G).** However, because the reference PSIs differ between COMPASS and endogenous measurements (**Figure 2.10E**), we computed Δlogit(PSI) to correct for differences in starting PSI. The results showed that variant 1 produced a strong increase in exon inclusion in COMPASS, and this effect was reproduced in PE-edited cells with similar magnitude. Variant 2 showed a detectable effect on lowering inclusion in COMPASS, and this effect was even more pronounced in the PE experiment, consistent with the

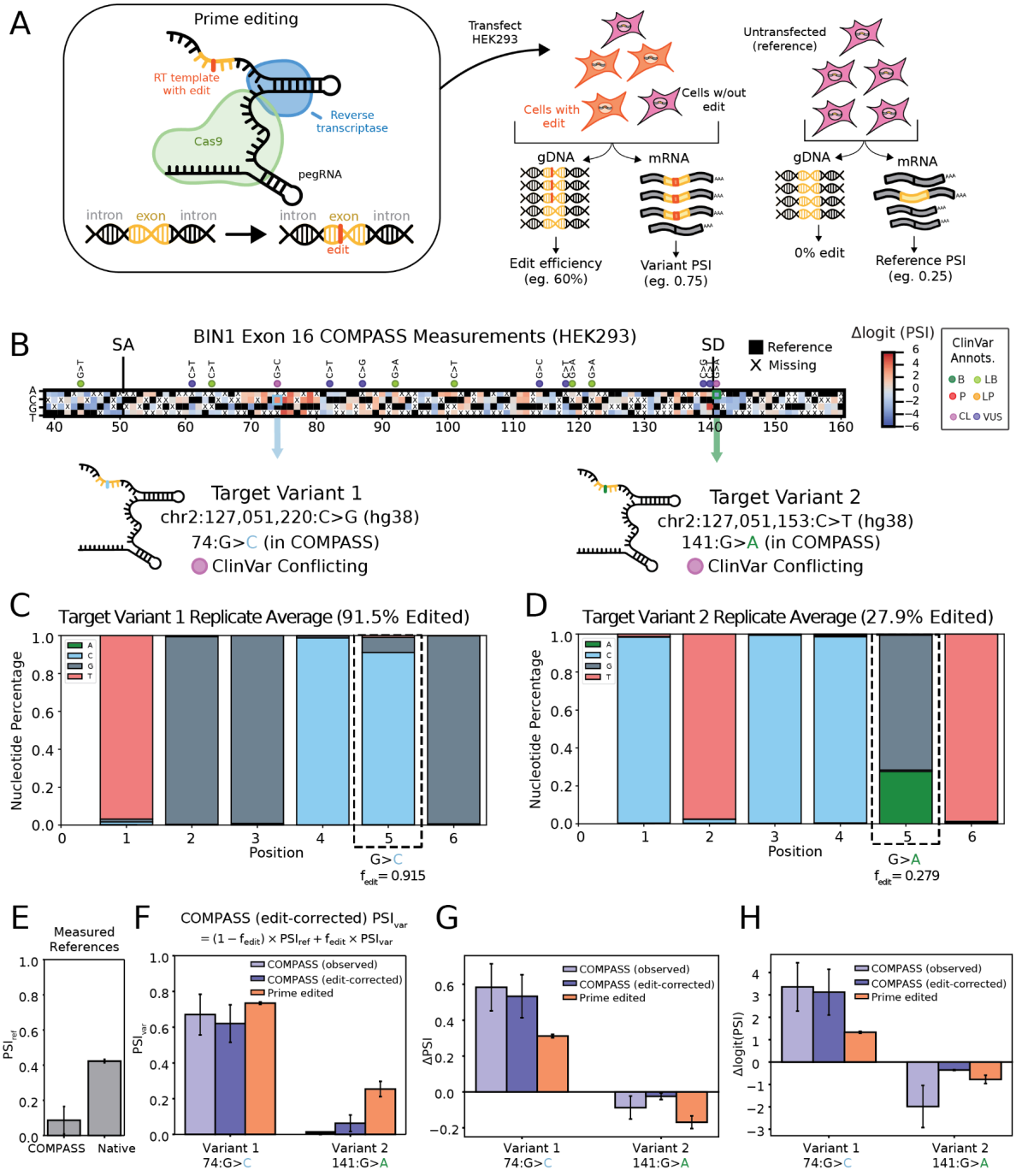


Figure 2.10 | Prime editing validation of BIN1 exon 16 ClinVar variants

(A) Schematic of the prime editing (PE) system and experimental workflow. PE enables the introduction of precise base edits in an exon or intron of interest. HEK293 cells were transfected with pegRNAs targeting *BIN1* exon 16, where the reverse transcriptase (RT) template specifies the desired substitution (e.g., G>C). Editing efficiency was measured from genomic DNA (gDNA) sequencing, and splicing outcomes were quantified from mRNA sequencing. An unedited population of HEK293 cells served as the control reference. **(B)** NSSM map of COMPASS results for *BIN1* exon 16 in HEK293 cells, highlighting the two ClinVar variants

chosen for validation. target variant 1 is chr2:127051220:C>G (hg38), corresponding to position 74:G>C in the COMPASS reporter. Target variant 2 is chr2:127051153:C>T (hg38), corresponding to position 141:G>A in the COMPASS reporter. **(C)** Editing efficiencies for variant 1 measured from gDNA sequencing in two biological replicates of PE-edited HEK293 cells (fraction edited = 0.919 in replicate 1 and 0.910 in replicate 2; average fraction edited = 0.915). **(D)** Editing efficiencies for variant 2 measured from gDNA sequencing (fraction edited = 0.281 in replicate 1 and 0.277 in replicate 2; average fraction edited = 0.279). **(E)** Reference PSI values (unedited controls) measured in COMPASS and in endogenous HEK293 cells. **(F)** Variant PSI values measured in COMPASS, the edit-corrected expectation from COMPASS, and the observed PE PSI in HEK293 cells. Edit-corrected values were calculated as shown where f_{edit} is the average fraction of edited alleles across both PE replicates for that variant. **(G)** ΔPSI values in COMPASS and PE contexts for both variants, comparing variant PSI to the respective reference PSI. (H) Δlogit(PSI) values calculated as demonstrating concordance of variant effects across COMPASS and PE contexts after accounting for baseline PSI differences.

reduced exon inclusion observed in COMPASS. Even in the native locus, variant 2 did not fully ablate exon inclusion despite changing the canonical U1 SD recognition sequence, suggesting U12 may partially rescue this SS.[115] Importantly, the concordance between COMPASS and PE was strongest on the Δlogit(PSI) scale, demonstrating that Δlogit(PSI) provides a more robust measure than linear ΔPSI for comparing contexts with different baseline inclusion levels.

Overall, these experiments demonstrate that the variant effects measured in COMPASS are consistent with those observed in the endogenous context. Both *BIN1* exon 16 variants, despite being annotated in ClinVar with conflicting interpretations of pathogenicity, produced reproducible splicing changes in the same direction in COMPASS and PE experiments. Importantly, the effect magnitudes closely matched once editing efficiency and differences in reference PSI were taken into account, thereby validating these conflicting ClinVar variants as having clear splicing effects in their native genomic context. Moreover, the finding that disruption of the poly-G tract leads to such a strong increase in exon inclusion in both contexts is consistent with these motifs functioning as splicing silencers and further suggests the involvement of RBPs that recognize this element.[30]

2.6 Conclusions

Understanding how sequence variation alters splicing is central to uncovering the functional mechanisms that drive human disease. Here, we report COMPASS, the largest splicing MPRA to date, screening over 87,000 single and double variants across 2,096 exons to map their effects on splicing in disease-relevant contexts. Much of the library was focused on disease-associated genes, including ACMG-listed genes and exons containing pathogenic ClinVar variants, enabling systematic measurement of SDVs in disease-relevant contexts. SDVs are well-recognized drivers of Mendelian and complex disease[116,117], but while identifying variants with strong splicing phenotypes is now feasible, establishing their causal connection to disease remains challenging.[118] To aid in reclassification of VUSs, or even variants not yet annotated in ClinVar, we propose a simple heuristic: variants in exons already harboring known P/LP alleles should be considered higher priority if they cause comparable or more severe splicing disruption, whereas variants with effects similar to B/LB alleles should be deprioritized. Even in the absence of known pathogenic reference variants, Δlogit(PSI) values provide a robust prioritization metric by enabling comparisons within exons, particularly when splicing alterations in other exons of the same gene have established disease relevance.

However, reclassification of many of these variants will still require additional evidence directly linking splicing changes to disease causality. The ACMG guidelines emphasize that functional assays can provide supportive evidence but are rarely sufficient on their own for definitive classification.[119,120] This highlights both the promise and the limitation of large-scale splicing assays: they can powerfully identify molecular phenotypes, but clinical interpretation must integrate genetic, clinical, and population-level data to determine pathogenicity.[121]

Care must also be taken in interpreting MPRA measurements themselves. Although ΔPSI captures the change in isoform abundance, it is highly dependent on the reference exon's baseline inclusion level. Since this baseline may differ between the reporter assay and the endogenous gene context, estimates of variant impact can vary accordingly. Prior work has shown that variant effects scale with baseline inclusion levels and that logit-transformed values provide a more stable metric than raw ΔPSI.[83,85] For this reason, we use Δlogit(PSI) as our primary metric in variant analyses and suggest that this approach may also be broadly useful for other reporter-based splicing MPRAs.

While Δlogit(PSI) provides a more stable measure than raw ΔPSI, interpretation of variant effects and their clinical relevance still depends on the baseline properties of the reference exon in its endogenous context. Naturally, alternatively spliced exons, for example, tend to be more sensitive to disruption, and recent work has highlighted "hotspot" exons defined by weak splice signals, dense RBP binding motifs, and prior evidence of alternative isoform usage as especially susceptible to variation.[110] We also note that reporter assays cannot always capture true variant effects: if an exon is constitutively excluded in a reporter, variants that weaken splicing are not easily detected. Similarly, if an exon is constitutively included in a reporter, measuring splice-enhancing variants remains challenging. These caveats emphasize the need to integrate reporter-based functional data with endogenous splicing data.

In this regard, our prime editing experiments provide independent validation that the splicing effects measured by COMPASS are preserved in the endogenous context. By introducing conflicting ClinVar variants into the genome of HEK293 cells and directly measuring splicing outcomes, we confirmed that MPRA-derived variant effects are consistent

with those observed in native settings, thereby establishing both the accuracy and biological relevance of COMPASS.

Finally, measured Δlogit(PSI) values in COMPASS are generally well predicted by current splicing models, underscoring their utility for SDV discovery. For example, we demonstrate that predictors such as SpliceAI can fill in gaps in NSSM maps when experimental coverage is incomplete.[74] More broadly, COMPASS provides a resource for training and benchmarking next-generation splicing ML models. These models are particularly valuable for generalizing beyond experimental datasets, enabling saturation mutagenesis predictions at a scale that cannot realistically be achieved experimentally, even with MPRAs. For example, they could be applied to generate predictions at genome scale.[15]

Together, COMPASS establishes a framework for systematically quantifying variant impacts on splicing. The next chapter extends this foundation to investigate how splicing outcomes vary across cell types, and how RBP motifs, including those with cell-type-specific activity, may shape these effects.

Chapter 3: Using COMPASS to map splicing outcomes mediated by RBP motifs and cell type-specific contexts

3.1 Motivation

AS is a fundamental post-transcriptional mechanism that enables a single gene to generate multiple transcript isoforms through the regulated inclusion or exclusion of coding and non-coding sequences. This process is controlled by a combination of *cis*-regulatory elements embedded in the pre-mRNA and trans-acting RBPs that recognize these sequences. The canonical splicing machinery relies on conserved features, including the 5′ SD, 3′ SA, branch point, and polypyrimidine tract, to define exon-intron boundaries and direct intron removal.[20,122–124]

Beyond these core motifs, splice regulatory elements distributed across exons and introns provide an additional layer of regulatory information. RBPs that bind these elements can act as splicing enhancers or silencers, and their activity is often tissue- or cell type-specific and, when misregulated, can contribute to disease.[21,125] Variation in RBP expression, localization, and activity across cellular contexts is a major driver of this specificity.

For example, neuronal RBPs such as NOVA can act as either splicing enhancers or repressors depending on motif position, and their misregulation has been linked to neurological disease.[126–128] Cell type-specific AS regulated by RBPs such as PTBP1 and Rbfox governs neuronal fate in the developing cortex, and disruption of this program can cause cortical malformations.[129] In epithelial tissues, ESRP1/2 promote inclusion of epithelial-specific exons that contribute to cell identity, and altered ESRP activity is linked to cancer progression.[130–132] In muscle, MBNL1 coordinates splicing programs required for normal development, and its sequestration underlies the pathogenic exon mis-splicing observed in myotonic dystrophy.[133–135]

More recent work also captures the differences in splicing between cell types and identifies regulators of this specificity.[70,136,137] Despite these advances, we remain far from accurately predicting cell type-specific splicing outcomes or pathogenic variant effects from sequence alone, motivating the development of experimental datasets and computational models that bridge this gap.

3.2 Overview of existing work

Previous splicing MPRAs compared relatively small sets of sequences across a limited number of cell types and generally reported only modest evidence for cell type-specific regulation.[39,40,59] This likely reflects the fact that truly cell type-specific splicing events are relatively uncommon compared to broadly conserved regulation, and thus require large-scale assays to detect with confidence. Similarly, an MPRA measuring mRNA stability across six different cell lines found that differential regulation was the exception rather than the rule, suggesting that many regulatory features are widely shared.[65] This apparent homogeneity may also be explained by the widespread activity of RBPs that are expressed and functional in most cellular contexts,[138] as well as technical aspects of plasmid-based reporters, which can dampen apparent differences by saturating splicing signals or failing to capture the influence of lowly expressed trans factors.[45]

In contrast, transcriptome-wide analyses have provided strong evidence for widespread cell type-specific splicing. For example, studies in primary neurons identified hundreds of alternatively spliced exons with highly cell type-specific inclusion patterns,[129,139] and large-scale resources such as GTEx and ASCOT have cataloged thousands of tissue-restricted splicing

events across human tissues.[28,70,140] These findings emphasize that cellular context plays a central role in shaping splicing regulation and variant effects.

Bulk transcriptome datasets such as GTEx and ASCOT, while powerful for cataloging tissue-specific splicing events, cannot resolve the *cis*-regulatory motifs or RBPs driving these changes and often average signals across heterogeneous cell populations.[28,70,140] Most predictive models are therefore trained in a cell type-agnostic manner and capture only broad patterns of variant impact.[61,74] Some predictors attempt to incorporate tissue-specific information using GTEx datasets, but these are limited to selected tissues and are not always comparable to cell line-based research.[31,71] Predicting cell type-specific splicing is particularly challenging because RBP activity is context-dependent: the same factor can enhance or repress exon inclusion depending on motif position, and RBPs often act combinatorially in ways that vary across cell types.[129,138,141,142]

These limitations highlight the need for functional datasets that perturb sequence motifs directly and measure their effects across distinct regulatory environments, rather than inferring patterns from bulk transcriptomes alone. COMPASS therefore takes a step toward filling this gap by systematically characterizing splicing in ENCODE reference cell lines HEK293 (embryonic kidney), K562 (leukemia), HeLa (uterine adenocarcinoma), MCF-7 (breast adenocarcinoma), and HMC3 (microglia).[72,73] By measuring splicing at scale in nearly 100,000 human-derived sequences that systematically perturb many possible motifs within the same sequence context, this assay provides the statistical power to infer RBP motif-driven regulation and to detect cell type-specific splicing outcomes in parallel.

3.3 Effect sizes can be associated with RBP binding motifs

To identify putative RBP binding sites, we assembled a database of RBP-associated motifs from ATtRACT, CISBP-RNA, and oRNAment, which integrates data from *in vitro* binding assays (SELEX, RNAcompete) with select curated *in vivo* CLIP-derived motifs. [143–145] The combined database contains 1,007 motifs represented as position weight matrices (PWM). Using FIMO, we scanned all library sequences with the PWM collection to identify statistically significant matches corresponding to putative RBP binding sites (**Figure 3.1A**). [146]

Next, we asked whether we could infer the importance of an RBP binding site from the impact of variants that disrupt it. To reduce redundancy of similar motifs across and within databases, and to streamline subsequent analysis, we clustered individual motifs into 209 consensus motif clusters, a common approach used for transcription factor motifs that has also been successfully applied to RBP motifs.[147–149] For each pair of motif cluster and exon family, we then divide the family into members that contain a motif from that cluster and those that do not. Here, an exon family is defined as a reference exon (and flanking introns) and all associated variants. A sequence matching a motif in the reference sequence may be sufficiently altered by an overlapping variant such that it no longer constitutes a significant match to the motif PWM in FIMO. Conversely, a variant might introduce a motif match even if the reference does not match the motif. We then calculate an average logit(PSI) for the two groups within each cell line and define the difference of these averages as the motif effect size (**Figure 3.1B**). The example shows exon 16 of *BIN1* which contains a good match for the FUS (fused in sarcoma) binding motif from oRNAment. This motif overlaps variant 1, which we characterized using PE (**Figure 2.10**). The *BIN1* exon 16 family members that do contain the FUS motif (green, n=551) on average have a more negative logit than those that do not contain it (red, n=35), resulting in an

overall negative effect size for the motif. This negative effect size is consistent with the cognate RBP, likely FUS, being an exonic splice silencer (ESS), consistent with literature.[150]

We note that splice site variants were excluded from this analysis because their effects are typically large, direct, and mediated through canonical splicing machinery, rather than through auxiliary RBPs. Double variants, however, were retained to increase the number of sequence perturbations considered, with the caveat that PSI values for a given motif family member may deviate from the reference due to a second variant distal to the motif of interest; this bias is minimized because both motif-containing and motif-lacking groups include such variants.

We generalized motif effect size calculations to all motif-exon family pairs in each cell line, provided that within that cell line, there were at least ten family members containing the motif and at least ten lacking it. **Figure 3.1C** shows data for HEK293 as a representative example, with all cell line data provided (**Figure S2A, B)**. We calculated effect sizes separately for motifs in the leading intron, central exon, and trailing intron, since RBPs are known to exert opposite regulatory effects depending on where they bind.[151–153]

Motifs belonging to Cluster 23, which includes motifs associated with ESRP1 in addition to FUS, have putative binding partners in five exon families (**Figure 3.1A)**. All observed instances are associated with a negative effect size, with the most negative effect size observed for *BIN1* exon 16 (**Figure 3.1B, C**). Conversely, effect sizes can vary widely when calculated for the same motif cluster but for different exon families. For example, motif Cluster 79 which includes motifs associated with ERI1, PCBP2, PCBP1 and HNRNPK has a mixture of negative and positive effect sizes when comparing different exon families (**Figure 3.1C**). There may be several reasons for this variation. First, each motif cluster aggregates multiple similar motifs,

each of which may be recognized by a different RBP (**Figure 3.1A**); in turn, different RBPs may be responsible for the observed splicing effects in different exon contexts even if the motifs belong to the same cluster. Still, we chose to calculate effect sizes at the cluster level because accurately assigning a particular instance of the motif to a particular RBP is challenging. Second, even for identical motifs recognized by the same RBP, local sequence context including distance to splice sites, RNA secondary structure, and the presence of neighboring RBP motifs can influence the regulatory outcome.[154] Finally, technical factors such as uneven sequencing depth across exon families may introduce additional noise, which can make RBP-associated effects present differently across families. For further analysis we focus on motifs for which the effect sizes have predominantly the same sign across all exon families (**Figure S2C**). Examples of motifs with concordant effect sizes are shown in **Figure 3.1D,E**. **Figure 3.1D** shows examples of exonic splice regulators. These include a motif associated with SRSF1/7 that is found in four exon families and has a positive effect size across all cell lines, consistent with the known role of SR proteins as exonic splice enhancers (ESEs).[155–158] Conversely, polyG motif clusters associated with multiple RBPs including multiple HNRNP family members act as ESSs as previously described.[150,159–164] **Figure 3.1E** shows intronic splice enhancers (ISEs) and silencers (ISSs). A polypyrimidine motif found to recruit PTBP1 and RBM5 is found across multiple introns and broadly acts as ISE, consistent with literature showing that PTB binds CU-rich elements and mediates context-dependent regulation of exon skipping.[165,166] Additional examples of motifs further show that variation of effect sizes between cell types tend to be more subtle than those between different exon contexts (**Figure 3.1D,E, Figure S2D,E**). Additional RBP effects are provided in **Appendix 2** (exonic motifs) and **Appendix 3** (intronic motifs), which present all motif plots analogous to **Figure 3.1D,E** in all cell lines.

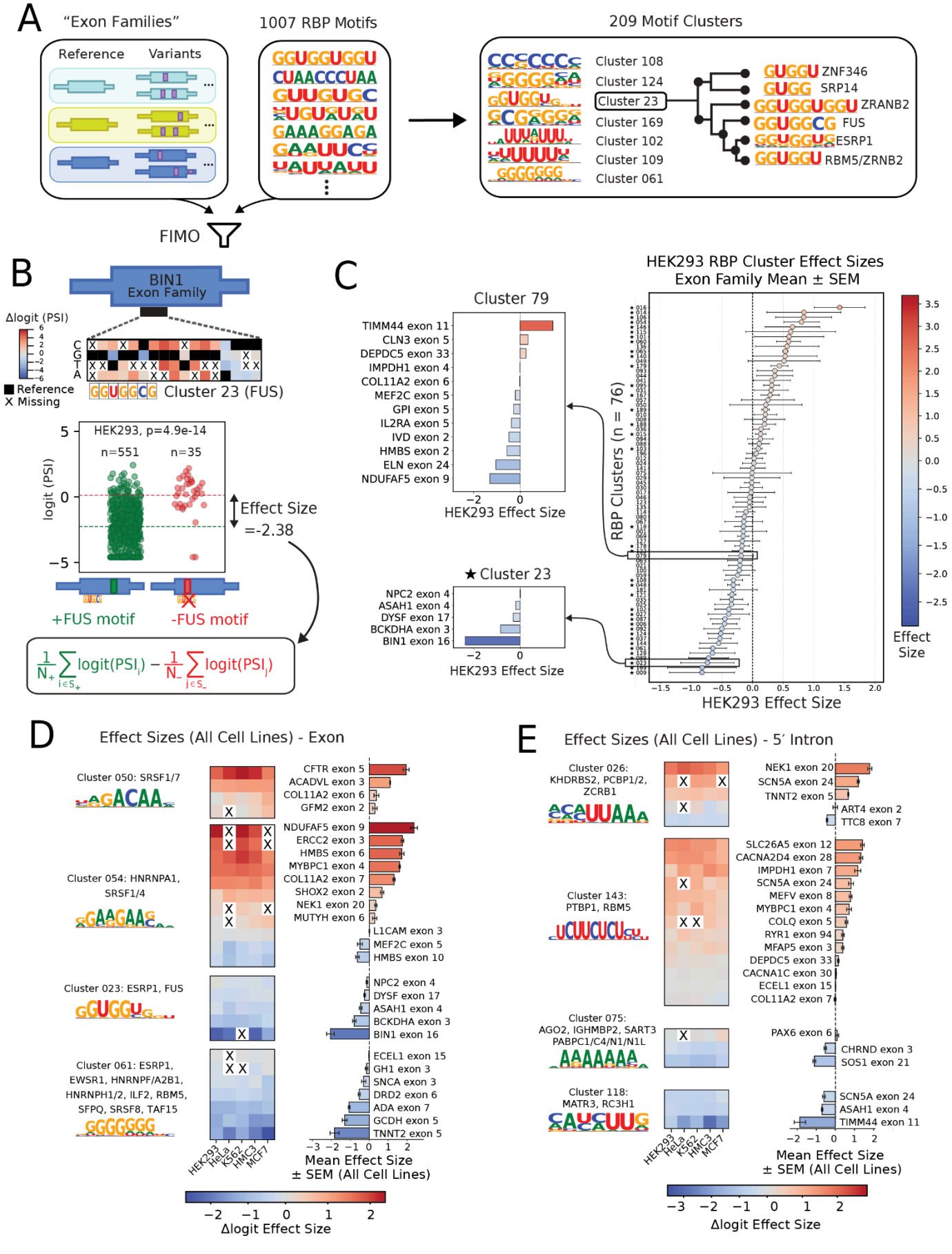

A
"Exon Families"
Reference
Variants
1007 RBP Motifs
209 Motif Clusters
Cluster 108
Cluster 124
Cluster 23
Cluster 169
Cluster 102
Cluster 109
Cluster 061
GUGGU ZNF346
GUGG SRP14
GGUGGUGGU ZRANB2
GGUGGCG FUS
GGUGGUG ESRP1
GGUGGU RBM5/ZRNB2
FIMO
B
BIN1
Exon Family
Δlogit (PSI)
Reference
Missing
Cluster 23 (FUS)
HEK293, p=4.9e-14
n=551
n=35
logit (PSI)
Effect Size
=-2.38
+FUS motif
-FUS motif
C
HEK293 RBP Cluster Effect Sizes
Exon Family Mean ± SEM
Cluster 79
TIMM44 exon 11
CLN3 exon 5
DEPDC5 exon 33
IMPDH1 exon 4
COL11A2 exon 6
MEF2C exon 5
GPI exon 5
IL2RA exon 5
IVD exon 2
HMBS exon 2
ELN exon 24
NDUFAF5 exon 9
HEK293 Effect Size
★ Cluster 23
NPC2 exon 4
ASAH1 exon 4
DYSF exon 17
BCKDHA exon 3
BIN1 exon 16
HEK293 Effect Size
RBP Clusters (n = 76)
HEK293 Effect Size
Effect Size
D
Effect Sizes (All Cell Lines) - Exon
Cluster 050: SRSF1/7
CFTR exon 5
ACADVL exon 3
COL11A2 exon 6
GFM2 exon 2
Cluster 054: HNRNPA1, SRSF1/4
NDUFAF5 exon 9
ERCC2 exon 3
HMBS exon 6
MYBPC1 exon 4
COL11A2 exon 7
SHOX2 exon 2
NEK1 exon 20
MUTYH exon 6
L1CAM exon 3
MEF2C exon 5
HMBS exon 10
Cluster 023: ESRP1, FUS
NPC2 exon 4
DYSF exon 17
ASAH1 exon 4
BCKDHA exon 3
BIN1 exon 16
Cluster 061: ESRP1, EWSR1, HNRNPF/A2B1, HNRNPH1/2, ILF2, RBM5, SFPQ, SRSF8, TAF15
ECEL1 exon 15
GH1 exon 3
SNCA exon 3
DRD2 exon 6
ADA exon 7
GCDH exon 5
TNNT2 exon 5
HEK293
HeLa
K562
HMC3
MCF7
Mean Effect Size
± SEM (All Cell Lines)
Δlogit Effect Size
E
Effect Sizes (All Cell Lines) - 5′ Intron
Cluster 026: KHDRBS2, PCBP1/2, ZCRB1
NEK1 exon 20
SCN5A exon 24
TNNT2 exon 5
ART4 exon 2
TTC8 exon 7
Cluster 143: PTBP1, RBM5
SLC26A5 exon 12
CACNA2D4 exon 28
IMPDH1 exon 7
SCN5A exon 24
MEFV exon 8
MYBPC1 exon 4
COLQ exon 5
RYR1 exon 94
MFAP5 exon 3
DEPDC5 exon 33
CACNA1C exon 30
ECEL1 exon 15
COL11A2 exon 7
Cluster 075: AGO2, IGHMBP2, SART3 PABPC1/C4/N1/N1L
PAX6 exon 6
CHRND exon 3
SOS1 exon 21
Cluster 118: MATR3, RC3H1
SCN5A exon 24
ASAH1 exon 4
TIMM44 exon 11
HEK293
HeLa
K562
HMC3
MCF7
Mean Effect Size
± SEM (All Cell Lines)
Δlogit Effect Size

Figure 3.1 | *Cis*-regulatory effects of RBP motifs across exon and intron regions
(A) Workflow of data processing prior to motif effect size analysis. Exon families were defined as groups of sequences containing a reference sequence and its corresponding variants. We compiled a set of 1,007 RBP motifs from the ATtRACT, CISBP-RNA, oRNAment, and Ray et al. (2013) databases. These motifs were collapsed into 209 representative clusters to reduce redundancy, and the clusters were used for downstream effect size analyses. All sequences were scanned for motif matches using FIMO, and the resulting clusters were used for downstream effect size analyses. **(B)** Example of motif effect size analysis. Effect sizes were calculated within exon families and defined regions. For each motif identified in a sequence by FIMO, we compared the set of sequences with a significant hit ($q \leq 0.01$) to those without a detectable hit. In the example of *BIN1* exon 16 measured in HEK293, variants (single or double) were grouped into motif-present and motif-absent sets. The equation defines the effect size as the difference in mean logit(PSI) between the motif-present and motif-absent groups, yielding a Δlogit(PSI) motif effect size. This analysis was performed within exon families and within specific regions (exonic and intronic) and repeated across all cell types. Further details are provided in the **Methods**. **(C)** Concordance filtering of motif effect sizes in HEK293. Motif clusters with effect size calculations in at least three exon families are shown. Each point represents the mean Δlogit(PSI) across exon families in HEK293, with error bars denoting the SEM. Clusters marked with a star pass the concordance filter (≥75% agreement in effect direction across exon families). Cluster 23 exemplifies consistent effects across contexts (including *BIN1* as in panel B), whereas cluster 79 does not pass concordance, reflecting differential effects of this motif across exon families. **(D)** Concordant RBP motifs in exonic regions. Examples of motif clusters with consistent effects across exon families are shown. SR protein motifs act as exonic splicing enhancers (ESEs) that promote exon inclusion, whereas G-rich motifs act as exonic splicing silencers (ESSs), in line with their previously reported repressive activity in MPRAs.[30] **(E)** Concordant RBP motifs in intronic regions. Shown are examples of motif clusters with consistent effects across multiple exon families. PTBP1/RBM5 motifs function as intronic splicing enhancers (ISEs) that promote exon inclusion, consistent with PTBP1 binding to extended CU-rich elements, and have been hypothesized to regulate splicing in a location dependent manner.[167–170] Poly(A)-associated motifs, recognized by PABPs, are shown as intronic splicing silencers (ISSs) which are often known to regulate splicing on terminal introns. [171–173] Other motifs are shown as ISSs or ISEs for comparison.

3.4 COMPASS maps cell-type-specific splicing programs

3.4.1 Cell type-specific exon inclusion

As noted above, splicing patterns are highly correlated between cell types (**Figure 2.1D**) and, consequently, RBP effect sizes are also broadly similar (**Figure S2D,E**). Still, there are many individual sequences that exhibit cell type specific splicing. To identify sequences that are most differentially spliced in one cell type relative to all others, we calculated a mingap score defined as the absolute difference between the PSI in the target cell type and the next closest PSI in any of the other cell types. We performed this analysis across the five cell lines (K562, HeLa, MCF7, HEK293, HMC3) profiled in this study.[72] We then ordered mingap scores by size to identify the sequences that are most differentially spliced in (**Figure 3.2A**). In this figure, we retained only the top 20 sequences per cell type based on their mingap ranking. For four of the five cell lines (HeLa, K562, MCF7, HMC3), we were able to identify at least one sequence with strong cell type specificity, here defined by a mingap ≥0.25 between the most and second most included cell line. In total, including those shown in **Figure 3.2A**, this included 21 sequences for HeLa, 8 for K562, 88 for MCF7, and 99 for HMC3 that are preferentially included in the target cell line but largely excluded across the others. This complete set of sequences is shown in **Figure 3.3B.**

To evaluate the robustness of observed cell type-specific splicing effects, we next asked whether the same shift recurs across multiple independent sequences within the same exon family. Variants associated with exon 8 of myocyte enhancer factor 2C (*MEF2C)* provide an intriguing example. This exon is referred to as exon β in the literature and is included in transcripts in the brain and striated muscle but excluded elsewhere and is a known oncogene. *MEF2C* encodes a transcription factor, and exon β inclusion has been shown to enhance its

transcriptional activation capacity.[174–176] In our reporter, most members of the *MEF2C* exon 8 family are excluded across all cell lines. However, 9 of 292 family members are differentially included (mingap ≥0.25) in MCF7 cells.

Intriguingly, we identified a recurrent mutation at the 24th position (T>A) that promotes exon inclusion in MCF7 cells (**Figure 3.2B, C**). This effect was observed across five independent sequences, including one single-nucleotide change and four double mutants that also carried additional substitutions elsewhere in the sequence. The reproducibility of the 24T>A change across sequences provides a high-confidence example of cell type-specific variant activity. The variant position overlaps a motif cluster containing several SR proteins and hnRNPs (HNRNPA1, SRSF1, SRSF4, SRSF5, SRSF6 and TRA2B), implicating altered RBP binding as a potential driver of the observed specificity. *MEF2C* has been identified as a known oncogene and also has reports of its expression in epithelial tumor cells of breast cancers; however, its precise role in this context remains unclear. More broadly, frequent alterations in SR protein family members across breast tumors underscore their central role in tumor-associated splicing programs, though whether *MEF2C* is directly affected has yet to be determined.[174,177–180] Additional representative examples of cell type-specific exon inclusion are *OCA2* exon 10 (high in HMC3), *USP28* exon 6 (high in K562), *DIAPH1* exon 2 (high in HeLa), and *TNNT2* exon 5 (high in MCF7) (**Figure 3.2D**). Many of these exons are alternatively and, in certain cases, cell-type-specifically spliced.[181–184] In *DIAPH1* exon 2, two of the three cell type-specific variants occur at base 30 and overlap cluster 44 (HNRNPK, PCBP1). The RBPs listed in the parentheses are the best matches from all motifs in the cluster for this particular sequence. In *TNNT2* exon 5, 52 variants exceed the mingap threshold, including 5 at position 12 that overlap cluster 026 (ZCRB1, PCBP1/2) and cluster 138 (KHDRBS3), both of which function as ISEs **(Appendix 2**).

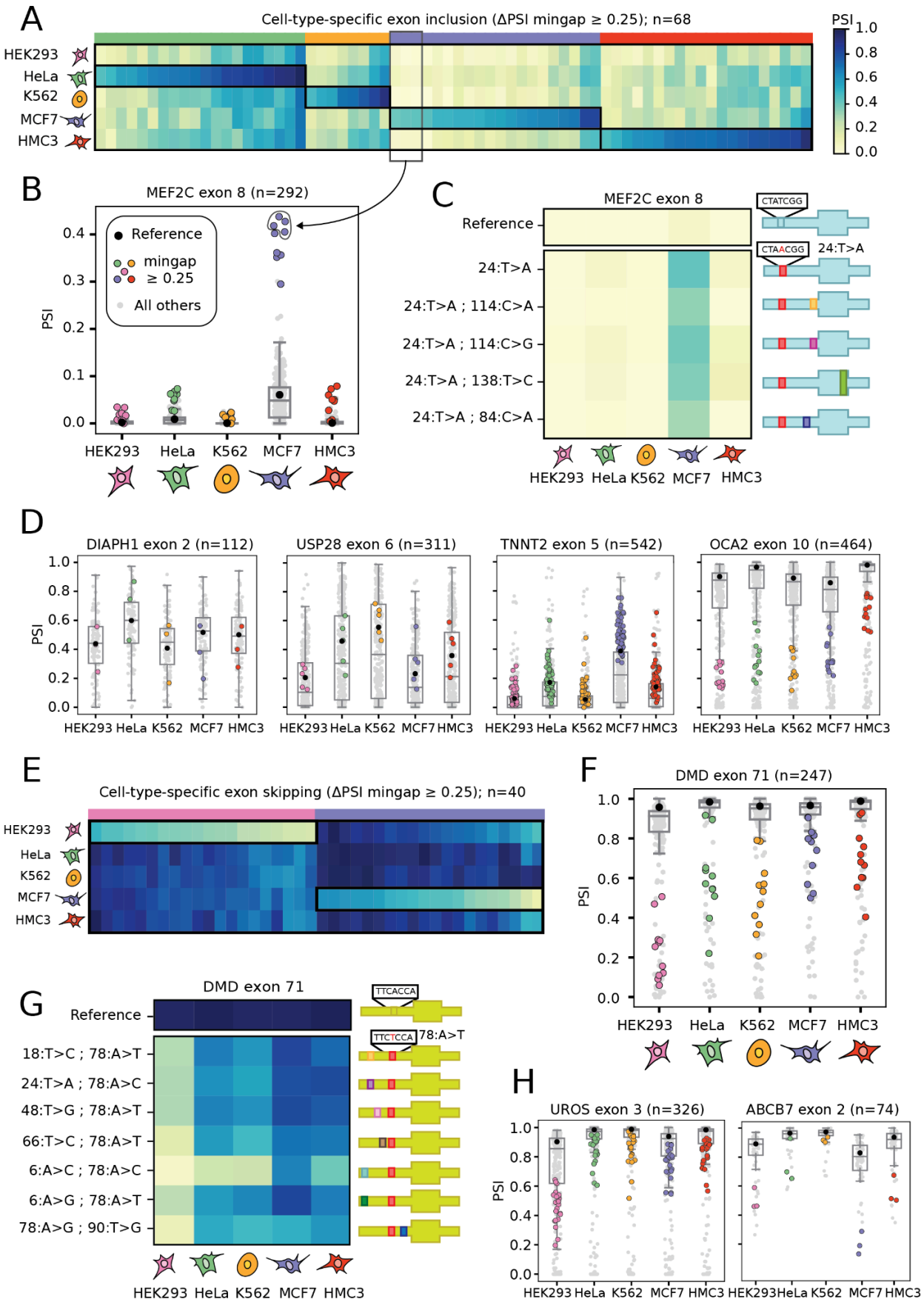

A
Cell-type-specific exon inclusion (ΔPSI mingap ≥ 0.25); n=68
HEK293
HeLa
K562
MCF7
HMC3
PSI
1.0
0.8
0.6
0.4
0.2
0.0
B
MEF2C exon 8 (n=292)
Reference
mingap ≥ 0.25
All others
PSI
C
MEF2C exon 8
Reference
24:T>A
24:T>A ; 114:C>A
24:T>A ; 114:C>G
24:T>A ; 138:T>C
24:T>A ; 84:C>A
CTATCGG
CTAACGG
24:T>A
D
DIAPH1 exon 2 (n=112)
USP28 exon 6 (n=311)
TNNT2 exon 5 (n=542)
OCA2 exon 10 (n=464)
E
Cell-type-specific exon skipping (ΔPSI mingap ≥ 0.25); n=40
F
DMD exon 71 (n=247)
G
DMD exon 71
Reference
18:T>C ; 78:A>T
24:T>A ; 78:A>C
48:T>G ; 78:A>T
66:T>C ; 78:A>T
6:A>C ; 78:A>C
6:A>G ; 78:A>T
78:A>G ; 90:T>G
TTCACCA
TTCTCCA
78:A>T
H
UROS exon 3 (n=326)
ABCB7 exon 2 (n=74)

Figure 3.2 | Cell type-specific splicing regulation revealed through ΔPSI mingap analysis
(A) Heatmap of sequences with cell type-specific exon inclusion, identified using the ΔPSI mingap ≥ 0.25 threshold, defined as the absolute difference between the highest PSI value and its next closest value. Up to the top 20 events per cell line were retained (HeLa, 20; MCF7, 20; HMC3, 20; K562, 8). **(B)** Swarm plot of MEF2C exon 8 sequences across cell lines. 9 variants pass the mingap ≥ 0.25 threshold in MCF7 and are colored for MCF7, with the reference in black. The same 9 variants are colored in the other four cell lines for comparison. **(C)** Example of a recurrent event: a single 24T>A variant in MEF2C exon 8 repeatedly observed to drive MCF7-specific inclusion. Statistical testing confirms reproducible differences across sequences. **(D)** Representative exon families where many variants drive cell type-specific inclusion, including DIAPH1 exon 2 (HeLa; n = 3), USP28 exon 6 (K562; n = 5), TNNT2 exon 5 (MCF7; n = 52), and OCA2 exon 10 (HMC3; n = 13). Variants passing the ΔPSI mingap ≥ 0.25 threshold in the indicated cell line are shown in color, with the reference PSI in black. These variants are also colored in the other four cell lines for comparison. The total number of sequences in each exon family is shown. **(E)** Heatmap of sequences with cell type-specific exon skipping, identified using the ΔPSI mingap ≥ 0.25 threshold, defined as the absolute difference between the lowest PSI value and its next closest value. Up to the top 20 events per cell line were retained (HEK293, 20; MCF7, 20). **(F)** Swarm plot of DMD exon 71 sequences across cell lines, a representative example of cell type-specific skipping (n = 11 variants with ΔPSI mingap ≥ 0.25). Variants meeting the threshold are shown in color, with the reference PSI in black. The same 11 variants are colored in the other four cell lines for comparison. **(G)** Recurrent variants in DMD exon 71 passing the ΔPSI mingap ≥ 0.25 threshold exhibit HEK293-specific exon skipping and MCF7-specific inclusion. **(H)** Representative exon families where many variants drive cell type-specific skipping, including UROS exon 3 (HEK293; n = 23) and ABCB7 exon 2 (MCF7; n = 3). Variants passing the ΔPSI mingap ≥ 0.25 threshold in the indicated cell line are shown in color, with the reference PSI in black. The same variants are colored in the other four cell lines for comparison

For *USP28* exon 6, no clusters provided sufficient evidence to explain the observed specificity, suggesting contributions from additional or uncharacterized regulatory motifs.

Families with cell type-specific variant effects are generally alternatively spliced in the reporter construct and, therefore, more sensitive to variant impacts than exons that are constitutively spliced in or out. Most of the exon families that support highly cell type-specific variants also exhibit cell type-selectivity at the family level. Global exon family-level differences in PSI are accentuated by specific variants, yielding pronounced cell type-specific splicing

outcomes. These effects tend to be smaller on the logit scale, which corrects for some of the non-linear effects observed when working with PSI.

3.4.2 Cell type-specific exon skipping

Applying the same approach to identify sequences that have cell type-specific exon skipping **(Figure 3.2E**), we identified strong evidence of preferential exon exclusion in two of the five cell lines (HEK293 and MCF7), defined by a mingap ≥ 0.25. In total, 614 sequences in HEK293 and 37 in MCF7 were selectively skipped in the target cell line while being broadly included in the others. In **Figure 3.2E**, we show only the top 20 sequences per cell type by mingap ranking. The complete set of all 614 cell type specific skipping sequences is shown in **Figure 3.3A**. A relevant example is *DMD* exon 71, where recurrent variants at position 78 reproducibly yielded HEK293-specific skipping and MCF7-specific inclusion (**Figure 3.2G**). This site overlaps clusters 092 (PCBP2/4) and 063 (HNRNPK), and effect sizes for Cluster 063 match the MCF7 inclusion phenotype, consistent with disruption of an ISE. Notably, *DMD* exon 71 is very short and shares characteristics with neuronal microexons, which are highly conserved short exons strongly regulated by RBPs.[2,185–187] Additional representative examples of cell type-specific skipping are *UROS* exon 3 (low in HEK293) and *ABCB7* exon 2 (low in MCF7) (**Figure 3.2H**). In *UROS* exon 3, 23 variants surpass the mingap threshold; 5 map to position 108 and overlap cluster 144 (TIA1L1), which functions as a strong ESS with slightly stronger effects in HEK293 as observed in the effect size analysis (**Appendix 2**). Additional variants overlap cluster 201 (SRSF1) and cluster 048 (ELAVL2). We were not able to calculate effect sizes for these clusters because they were not observed sufficiently frequently. For *ABCB7* exon 2, which is known to be alternatively spliced,[181] no clusters provided sufficient evidence to explain the

observed specificity, suggesting contributions from additional or uncharacterized regulatory motifs. In both cell type-specific inclusion and skipping, most of the exon families that support highly cell type-specific variants also exhibit cell type-selectivity across the entire family. Global exon family differences are accentuated by specific variants, yielding pronounced cell-type-specific splicing outcomes.

3.4.3 Other forms of cell type-specific splicing

So far, we've focused on the two most dramatic types of cell type-specific splicing where an exon is included (or excluded) in one cell type versus all others in COMPASS. However, there is a broader, more general class of splice events where differential splicing occurs between two cell lines with intermediate PSI values for the other three. A representative example of that group is *COLQ* exon 5 where a large fraction of family members exhibit markedly higher exon inclusion in HeLa compared to HEK293 (**Figure 3.3C**). For further characterization of this differential splicing we selected one member of the *COLQ* exon 5 family (**Figure 3.3D**). *COLQ* exon 5 contains motif hits for motif clusters 61 (HNRNPH1/2/3/F), 80 (SRSF3), and 166 (SRSF9), and we next asked whether disrupting any of these motifs would change the observed cell type specificity. We created three additional splicing reporters with two or three mutations each disrupting one motif at a time (**Figure 3.3E**). We first confirmed using a gel assay that the PSI of the original COMPASS variant is approximately twice as high in HeLa cells compared to HEK293 cells. (**Figure 3.3F**). Motif disruption experiments revealed distinct contributions of motifs from clusters 80 (SRSF3), 166 (SRSF9), and 61 (HNRNPH1/2/3/F2). The putative SRSF3 binding motif acted as a strong enhancer: disruption of the motif reduced PSI to zero in

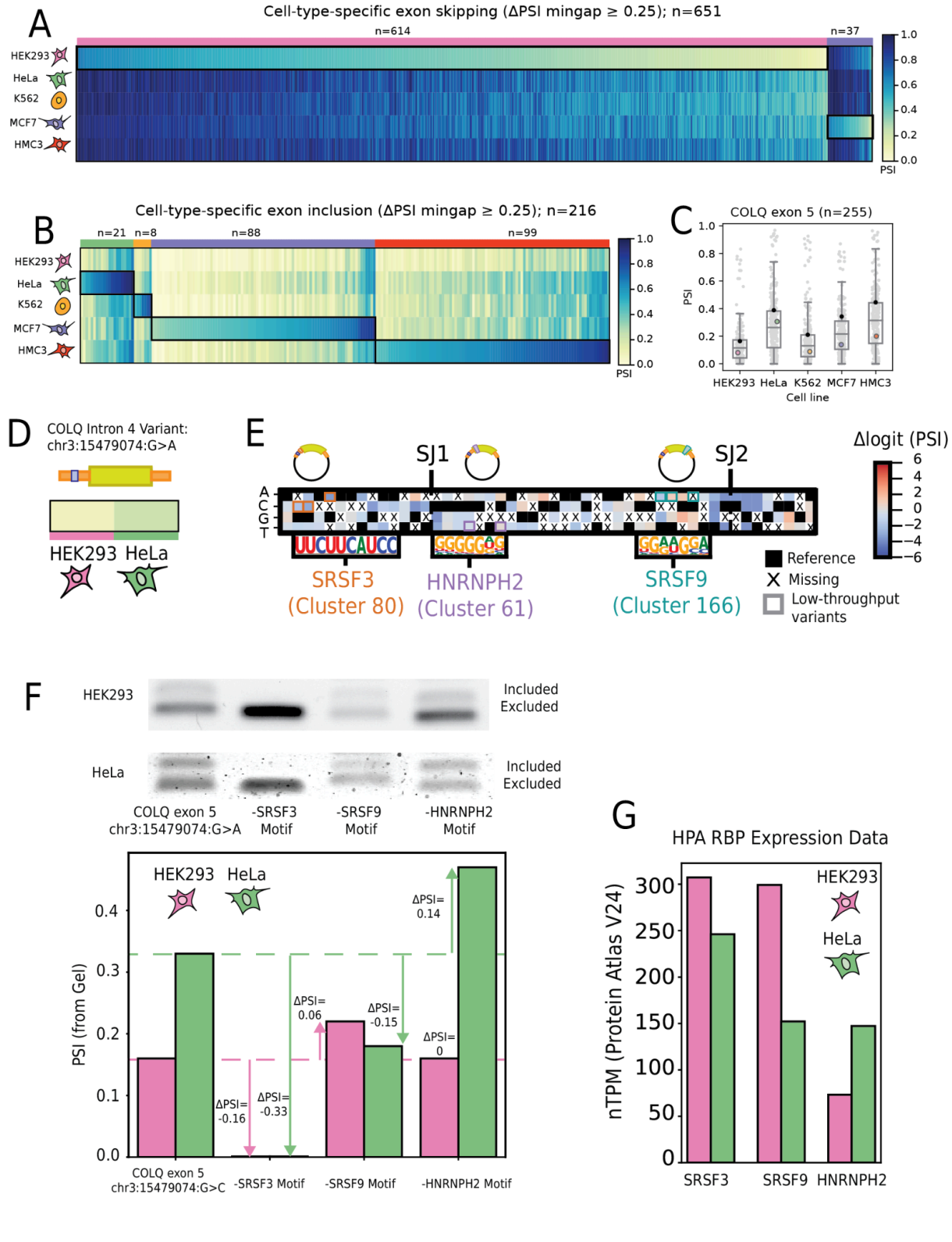

A
Cell-type-specific exon skipping (ΔPSI mingap ≥ 0.25); n=651
n=614
n=37
HEK293
HeLa
K562
MCF7
HMC3
PSI
B
Cell-type-specific exon inclusion (ΔPSI mingap ≥ 0.25); n=216
n=21
n=8
n=88
n=99
C
COLQ exon 5 (n=255)
PSI
Cell line
D
COLQ Intron 4 Variant:
chr3:15479074:G>A
E
SJ1
SJ2
Δlogit (PSI)
SRSF3
(Cluster 80)
HNRNPH2
(Cluster 61)
SRSF9
(Cluster 166)
Reference
Missing
Low-throughput
variants
F
Included
Excluded
COLQ exon 5
chr3:15479074:G>A
-SRSF3
Motif
-SRSF9
Motif
-HNRNPH2
Motif
PSI (from Gel)
ΔPSI=
-0.16
ΔPSI=
-0.33
ΔPSI=
0.06
ΔPSI=
-0.15
ΔPSI=
0
ΔPSI=
0.14
COLQ exon 5
chr3:15479074:G>C
G
HPA RBP Expression Data
nTPM (Protein Atlas V24)
SRSF3
SRSF9
HNRNPH2

Figure 3.3 | Cell type-specific splicing regulation revealed through differential splicing analysis and motif perturbation.
(A) Heatmap of sequences with cell type-specific exon inclusion, identified using the mingap ΔPSI ≥ 0.25 threshold, defined as an average replicate absolute ΔPSI between the highest-PSI cell line and the next highest (HeLa, 21; MCF7, 88; HMC3, 99; K562, 8). **(B)** Heatmap of sequences with cell type-specific exon skipping, identified using the mingap ΔPSI ≥ 0.25 threshold, defined as an average replicate absolute ΔPSI between the lowest-PSI cell line and the next lowest (HEK293, 614; MCF7, 37). **(C)** Swarm plot of *COLQ* exon 5 PSIs for all cell lines which shows moderate HEK293-HeLa differences. One variant passed the mingap ΔPSI ≥ 0.25 threshold and is shown in color. The reference PSI is marked in black. **(D)** One representative sequence with a variant upstream of *COLQ* exon 5 was selected for motif disruption experiments based on its large PSI difference (0.16) between HEK293 and HeLa, and a tractable number of candidate motifs. **(E)** NSSM plot of *COLQ* exon 5 and portions of intron 4. The plot is shown on Δlogit(PSI) scale, with RBP motif binding sites for SRSF3, SRSF9, and HNRNPH2 identified by FIMO. Targeted mutagenesis of each motif was performed in three separate reporters to assess the regulatory and cell type-specific roles of these motifs; mutated regions are highlighted. **(F)** RT-PCR gels showing alternatively spliced isoforms for the original variant (chr3:15479074:G>C) and each of the three motif-disrupting constructs, tested in HEK293 and HeLa. Isoform proportions reflect the splicing impact of each perturbation. PSI values were derived from the quantification of gel band intensities for each construct in both cell types. ΔPSI values are calculated between each motif-disrupting construct and the original variant (chr3:15479074:G>C), shown separately for HEK293 and HeLa. **(G)** Expression (nTPM) of SRSF3, SRSF9, and HNRNPH2 in HEK293 and HeLa based on data from the Human Protein Atlas (version 24).

both cell lines (as measured on a gel**)**. Disruption of the SRSF binding motif had a very slightly positive effect in HEK293 but a fairly strong negative impact in HeLa where it reduced PSI by about half. Finally, HNRNPH1/2/3/F functioned as a HeLa-specific silencer as deletion of the corresponding motif resulted in increased PSI in HeLa but not HEK293 (**Figure 3.3F)** This example demonstrates how local motif architecture and differential RBP activity could impact cell type-specific splicing regulation. Expression data from the Human Protein Atlas (proteinatlas.org) (**Figure 3.3G**) showed that HNRNPH2 is more highly expressed in HeLa, consistent with it acting as a stronger silencer in this cell line, whereas the effects of SRSF9 were not fully consistent with their expression patterns.[188]

3.6 Benchmarking cell type-specific splicing predictors

A central goal of this work was to evaluate whether splicing prediction models can capture the variant effects measured in COMPASS across diverse cell lines. In particular, we asked whether models that explicitly incorporate tissue or cell type or tissue information outperform those that rely solely on sequence only. In our initial benchmarking of cell type-agnostic models (**Figure 2.4F**), SpliceAI achieved the strongest predictive performance on our dataset, whereas Pangolin, despite being trained on tissue-resolved data, performed worse. SpliceAI and Pangolin share the same underlying deep residual convolutional neural network (CNN) architecture and both use long sequence contexts (±10 kb) to predict splice site usage probabilities.[31,74] By contrast, Pangolin integrates tissue-specific annotations from GTEx across five representative tissues (heart, liver, brain, testis) to model tissue-specific splice usage.[31] However, GTEx profiles are derived from bulk tissues composed of heterogeneous cell populations, which may not map precisely to the regulatory programs active in immortalized cell lines. Immortalized lines used in COMPASS represent unique cellular contexts, where we showed that tissue-specific models such as Pangolin did not offer an advantage over tissue-agnostic approaches.

This observation raised the broader question of whether any model explicitly incorporating tissue or cell type information could improve predictions on COMPASS data. To address this, we evaluated MTSplice, a model designed to integrate tissue-specific regulatory activity.[71] For comparison, we previously used MMSplice, a tissue-agnostic model that splicing predictions based on sequence alone. As described in Chapter 2, we applied a filtered evaluation in which sequences with reference PSI values of 0 or 1 were excluded.

To generate MTSplice predictions, we mapped each of our five assayed cell lines to the most relevant tissue categories provided by MTSplice. Specifically, HEK293 was mapped to Cortex-Kidney (index 36), MCF7 to Mammary Tissue-Breast (index 22), K562 to Leukemia (CML)-Cells (index 24), HeLa to Ectocervix-Cervix, Endocervix-Cervix, and Uterus (indices 26, 27, 53; averaged), and HMC3 to thirteen brain regions (indices 9-21; averaged). For each mapping, we generated an average Δlogit prediction for the corresponding tissue and compared it directly to the measured Δlogit values for that cell line. For MMSplice, a single tissue-agnostic prediction was used for all five lines.

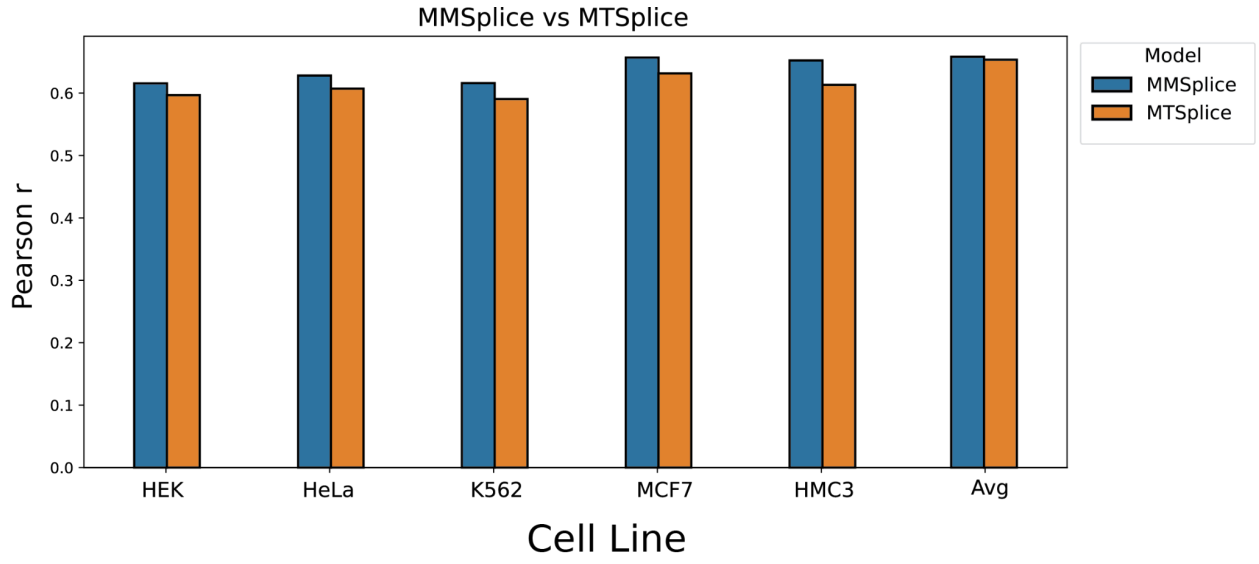


Figure 3.4 | Benchmarking tissue-specific vs tissue-agnostic splicing models on COMPASS
Bar plots show Pearson correlation (r) between measured Δlogit(PSI) values in each cell line and predictions from MTSplice (orange) or MMSplice (blue). For each of the five cell lines (HEK293, HeLa, K562, MCF7, HMC3), measured Δlogit values were compared to MTSplice predictions generated from the most closely matched GTEx tissues. For MMSplice, measured Δlogit values were compared to the tissue-agnostic MMSplice predictions. Additional bars show correlations between the average measured Δlogit across all cell lines and the corresponding average MTSplice or the tissue-agnostic MMSplice predictions. Across all comparisons, MTSplice performed slightly worse than MMSplice, indicating that tissue-specific modeling trained on bulk GTEx data does not improve predictive accuracy of COMPASS data.

The results are summarized in **Figure 3.4**. Each bar shows the Pearson correlation between measured Δlogit values for a given cell line and predictions from either MTSplice (orange) or MMSplice (blue). We also included correlations between the average measured Δlogit across all five cell lines and the corresponding average MTSplice predictions (orange) or the tissue-agnostic MMSplice predictions (blue). Across all comparisons, MTSplice performed slightly worse than MMSplice, failing to improve prediction accuracy for any individual cell line or for the averaged data.

Together with the reduced performance of Pangolin compared to SpliceAI (**Figure 2.4F**), these results indicate that, in the context of our cell line-based MPRA dataset, tissue-specific models do not provide a predictive advantage. Models trained on bulk tissue profiles may not accurately capture the regulatory environment of cell lines. While splicing is undoubtedly shaped by cell type-specific context, the models tested here appear to derive more predictive power from generalizable features such as splice site strength and motif composition than from tissue-specific information.

3.7 Conclusions

While many RBPs and their corresponding cis-regulatory binding motifs have been identified and compiled in databases such as ATtRACT, CISBP-RNA, oRNAment, and Ray et al. (2013)[143–145], we have yet to fully characterize the quantitative relationships between *cis*-regulatory motif variation, cellular context and splicing outcomes. Here, we took a step toward addressing this challenge by using motif perturbations to quantify the impact of RBP binding motifs on exon inclusion. Earlier work, such as MaPSy, established important precedent by comparing reference to SNVs in specific contexts.[82] However, that approach was limited to

pairwise comparisons. By assaying thousands of SNVs across many motif occurrences, COMPASS enables a large-scale and robust quantification of motif effects on splicing regulation. We note that further work is needed to establish causal links between motifs and the RBPs that recognize them. Mapping from RBP to motif is usually not unique and many splice regulatory proteins are associated with multiple, often divergent motifs determined using different measurement methods or biological contexts. Conversely, a single putative binding site is often compatible with motifs for multiple different RBPs. A further confounder is that RBP levels are often not known as they may differ from those of the corresponding RNAs because of differences in stability or post-translational modifications required for protein activity.

We systematically compared splicing of tens of thousands of reporter sequences across five cell lines of diverse tissue origins. We found that splicing patterns are highly correlated across different cell types but also identified subsets of sequences with cell type-specific splicing patterns. Strongly differentially spliced exons in COMPASS (mingap of $\geq$0.25) are rare with fewer than 1% of sequences identified as cell type specific. At the exon family level, fewer than 6% of families contained even a single sequence classified as cell type-specific by this criterion. These sequences that did show cell type-specific effects represent promising candidates for follow-up characterization in their native genomic context to determine whether they reflect true cell type-specific events outside a plasmid reporter context. These findings agree with earlier splicing MPRAs comparing smaller numbers of cell types and sequences.[39,40,82] A recent MPRA measuring mRNA stability across six different cell lines similarly found that differential regulation is the exception rather than the rule.[189] This relative homogeneity of splicing patterns across cell lines may be explained by the fact that the features underlying splicing decisions are largely shared across cell types, reflecting the widespread activity of RBPs that are broadly

expressed and functional across cellular contexts.[138] Moreover, the use of plasmid-derived splicing reporters may result in some saturation effects that could dampen the observed cell type-specificity, and it is also possible that certain splicing factors are not sufficiently present in the cell lines used. Thus, additional validation of these exons in endogenous settings is suggested. Nonetheless, these results are consistent with findings from primary neurons which identified hundreds of differentially spliced exons for a given cell type of interest when assaying exon inclusion at transcriptome scale.[129,139] Models that treat splicing in a cell type-agnostic manner capture broad patterns of variant impact, but they may not fully account for events that depend on cell type context. Incorporating cell type-specific regulatory information could aid in the interpretation of the subset of variants with cell type-specific splicing outcomes, particularly in disease contexts where pathogenicity is observed in certain cell types.

3.8 Ongoing and future work for COMPASS

Importantly, COMPASS was uniquely conducted across five distinct cell lines, providing a compendium of splicing regulation in different cellular contexts. Building on this resource, we are developing new ML models to predict splicing regulation, with the goal of capturing cell type-specific patterns and disease variant effects directly from COMPASS data. Preliminary CNN models trained on COMPASS data show promise in predicting PSI values across cell types. Moreover, by retraining MMSplice on COMPASS data, we substantially improved its performance over the baseline model, reaching accuracy comparable to SpliceAI. This success provides a foundation for extending the same strategy to MTSplice, with the aim of more effectively capturing cell type-specific variant effects.

A key next step is to make these models interpretable, particularly by linking predictions to RBP motif presence. This will be important whether we pursue newly developed CNN architectures or adapt existing models such as MMSplice. We can use Scrambler, which learns to mask unimportant bases and highlight sequence positions essential for prediction. For example, Scrambler can detect the loss of a presumed RBP binding site or other important motif introduced by a variant, directly linking model predictions to interpretable sequence features.[190] This can allow us to connect learned features to known RBP motifs and validate their roles in splicing regulation, in a manner similar to our effect size analyses.

Following model development, we aim to use it for *in silico* design of new cell type-specific sequences, in addition to experimentally testing sequences already identified in COMPASS across multiple cell types for validation. As a proof of concept, we are testing cell type-specific splicing using three COMPASS variants with strong effects (ΔPSI mingap ≥ 0.25), in which the alternative exon introduces an in-frame stop codon. In our minigene reporter, which contains constitutive citrine exons, inclusion of this stop codon makes protein reporter output directly dependent on cell type-specific splicing. We will use flow cytometry to measure these effects.

Existing approaches to cell type-specific expression typically rely on promoters or enhancers, which can be large, difficult to characterize, and prone to leaky activity.[191,192] An alternative strategy, exemplified by the SLED system, uses naturally occurring tissue-specific exons identified from the ASCOT database.[70,193] Building on this idea, our approach would leverage MPRA data to identify sequences with cell type-specific splicing activity directly in the context in which they would be deployed. Future models trained on COMPASS could extend this strategy by designing additional sequences *in silico* for experimental testing, providing a more

targeted and potentially higher-success strategy for programmable, post-transcriptional control of protein expression. Beyond discovery of additional cell type-specific sequences, COMPASS, along with the development of new or existing models, could also be used to guide the design of splice-modulating therapeutics, such as splice-switching oligonucleotides or synthetic splice factors, to correct aberrant splicing in human disease.[194–196]

Chapter 4: An MPRA of clinically relevant variants in multiple cell lines to validate a residual neural network for predicting 3′ cleavage and polyadenylation

4.1 Motivation

Almost all human mRNA transcripts undergo cleavage and polyadenylation. The position and efficiency of 3′ end processing are governed by a complex *cis*-regulatory code centered on the polyadenylation signal (PAS). The PAS consists of a core hexamer, typically AATAAA, and surrounding upstream and downstream sequence elements, which together recruit the core processing machinery (CFIm, CstF, CPSF, and hFIP1) (**Figure 4.1A**).[197–200] Moreover, more than 70% of human genes contain multiple PASs (alternative polyadenylation, or APA), resulting in RNA isoforms with distinct 3′ ends (**Figure 4.1B)**.[201–203] The most common form of APA is the occurrence of two or more competing PASs in the 3′ untranslated region (3′ UTR).[197]

Numerous genetic variants are known to cause or contribute to human disorders by disrupting the *cis*-regulatory code of polyadenylation signals. Yet, due to the complexity of this code, variant interpretation remains challenging.[204] Assessing the impact of genetic variation on polyadenylation is important in both research and clinical settings, as several mutations that disrupt APA isoform abundances have been implicated in disease.[205–207]

A quantitative understanding of the APA code would improve our ability to identify such pathogenic variants. However, experimentally characterizing every possible variant is not feasible even with high-throughput measurement techniques, such as MPRAs [47,53,60,63,208–210], which are still limited to tens of thousands of variants and mostly targeted to specific genes.

Statistical methods, such as GWAS or and mapping of APA QTLs (3′ aQTLs) have had success in linking genetic variation to disease but require large sampling to characterize rare variants and do not provide information for *de novo* variants.[211] Furthermore, GWAS cannot predict the mechanisms for which a variant is pathogenic.[212–215]

These limitations highlight the need for predictive models that can infer functional effects directly from sequence. Such models can also score disease variants for their impact on RNA processing, including APA, by estimating their likelihood of disrupting isoform usage. Applying these predictions to variants identified in population studies and validating them with functional assays will maximize the impact of MPRAs by focusing experimental efforts on the most informative and clinically relevant variants.

4.2 Overview of existing technologies

4.2.1 Machine learning for predicting APA

Evaluating the influence of genetic variation on polyadenylation holds significance in both research and clinical settings, given the implication of several mutations that disrupt APA isoform abundances and can contribute to disease.[207,216–218] Even single PASs in 3′ UTRs, devoid of competing signals, may contain mutations that can cause aberrant polyadenylation or cause an imbalance of mRNA abundance.[219]

In a complementary approach to GWAS or aQTLs, deep learning models that infer the functional impact of genetic variants directly from sequence have shown strong performance in distinguishing disruptive mutations, independent of their population frequency.[66,74,89,220–224] Several of these sequence-based models have also been extended to polyadenylation.[216–218,225] In particular, our lab has previously trained a CNN called APARENT for APA prediction.[226]

Inspired by the recent success of deep residual networks applied to splicing and transcription factor binding prediction[74,227], we have developed APARENT2, a sequence-based residual neural network for 3′ cleavage prediction at base-resolution.

4.3 APARENT2: A residual neural network for predicting 3′ cleavage and polyadenylation

Inspired by the recent success of residual neural networks applied to splicing and transcription factor binding prediction, our lab developed APARENT2, a sequence-based residual neural network for 3′ cleavage prediction at base resolution.[15,74,227] This model generalizes to the case of APA for a variable number of polyadenylation signals.

APARENT2 is a deep residual network on a re-processed version of the APA MPRA of Bogard et al which contain over 3.3 million APA reporters with randomized proximal PAS sequence measured within 12 diverse 3′ UTR contexts.[226] Briefly, the MPRA data was re-processed to map 3′ cleavage reads at base-pair resolution for some missing UTR contexts. The network, which is illustrated in (**Figure 4.1C)** and is referred to as APARENT2, is architecturally similar to SpliceAI and BPNet.[74,227]

Through a sequence of 28 residual blocks, the network transforms a one-hot coded representation of the input PAS (205 nt) into a predicted 3′ cleavage distribution. The last output layer of the network predicts the total isoform proportion of a competing distal PAS. To evaluate performance, we tested the network's ability to infer total proximal isoform abundance on a held-out set of 1085 native human PASs (measured in the MPRA[226]) (**Figure 4.1D**). APARENT2 had a very high correlation (r^2 =0.84) with this held-out set, with improved performance compared to other existing APA models.

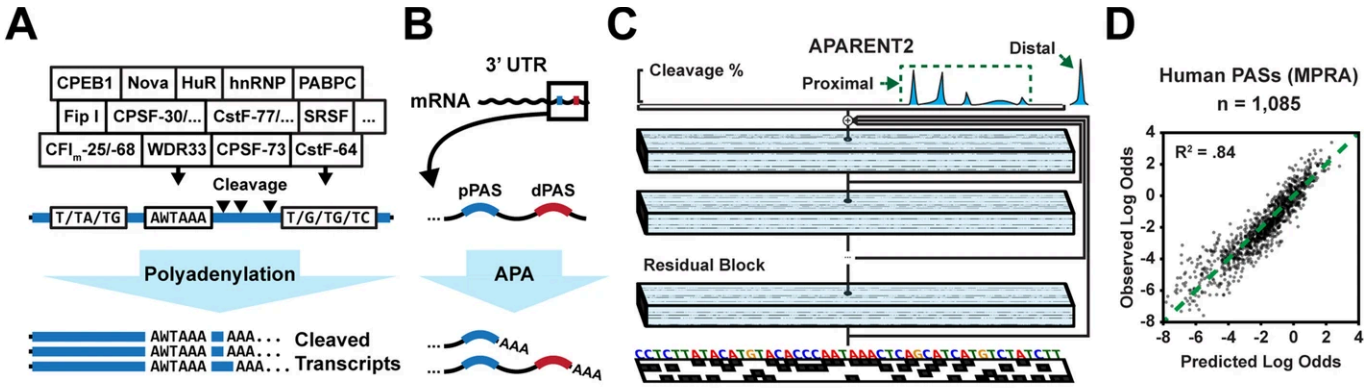


Figure 4.1 | Deep residual neural network model for predicting APA from sequence
(A) Core processing elements, auxiliary RBPs, and other determinants influence polyadenylation signal affinity. **(B)** Illustration of tandem 3′ UTR APA in pre-mRNA. **(C)** Residual neural network architecture. A one-hot coded representation of the PAS is used to predict the 3′ cleavage distribution. **(D)** Predicted vs measured proximal isoform log odds of native human 3′ UTR PASs measured in an MPRA (n=1,085). Figure 4.1 is adapted from Linder J, **Koplik SE**, et al. (2022) *Genome Biol.*

4.3.1 Disrupted polyadenylation variants (predicted by APARENT2) are selected against in the human population

We next sought to understand the connection between the functional impact of genetic variation on polyadenylation and human health. Using APARENT2, we performed full *in silico* saturation mutagenesis of every annotated PAS in PolyADB V3[228] and imputed the effect size (odds ratio) of every possible SNV (n > 43.8 million). For each PAS, we calculated the average wildtype isoform usage across all tissues in PolyADB. We then recalculated the isoform usage in the presence of each mutation by using the APARENT2 prediction to scale the isoform odds. Given these two quantities, we estimated the change in isoform proportion (Δuse) due to each variant. When cross-referencing our predictions against the > 2.8 million PAS SNVs curated from the > 71,000 genomes sequenced in gnomAD v3[229] (**Figure 4.2A**), we found that disruptive loss-of-function variants (resulting in downregulated polyadenylation) are depleted in common

variants (AF > 0.1%) compared to singletons (wilcoxon p= 2.1 x10^{-76}; **Figure 4.2B**). . Disruptive loss-of-function variants (Δ use < −0.15) occur ~2.5-fold less frequently among common variants (AF >10%) than singletons and they occur ~1.4-fold more frequently in unobserved variants (AF = 0%) compared to singletons (**Figure 4.2C**). These results suggest a negative selection pressure on disruptive variants in human polyadenylation signals.

4.3.2 Gain-of-function mutations at the 3′ end are enriched in autism spectrum disorder (ASD)

Most of the known deleterious polyadenylation variants are highly disruptive CSE mutations.[205–207] However, while we found in the previous section that highly disruptive loss-of-function variants are generally selected against, they also frequently occur as common variants. This suggests that we cannot use the inferred effect on polyadenylation alone as a predictor for variant pathogenicity. To highlight this phenomenon, we intersected our predictions against CAUSALdb[230], a database containing fine-mapping results from over 3000 GWAS summary statistics (including UK Biobank[231] and GWAS Catalog[232]). We first noticed that SNPs with a large posterior inclusion probability (PIP) from UK Biobank are enriched for disruptive APA variants (**Figure 4.3B**).[233] We then identified 96 SNPs with PIP >90% and many of these are known deleterious variants that act through APA (**Figure 4.2D**). As expected, the predicted effect size of these known APA mutations varies considerably. For example, the variant rs1799963 in the F2 gene increases polyadenylation efficiency only modestly (<1.5-fold) but is responsible for thrombophilia.[219] In contrast, the cancer-associated variant rs78378222 disrupts the PAS of the TP53 gene >10-fold. Clearly, the downstream consequence of disrupted polyadenylation depends on the importance of the affected APA isoforms, not to mention the

gene itself. However, we can assume that a mutation is likely not deleterious if it occurs in a PAS with common variants that have even larger effect sizes. Thus, we can eliminate PAS mutations and classify them as likely benign when they co-occur with putative functional common variants in gnomAD with high impact on polyadenylation. For example, the pathogenic variant rs1799963 would not be eliminated, since it is the variant with the largest predicted odds ratio of all observed F2 variants in gnomAD (**Figure 4.2E**).

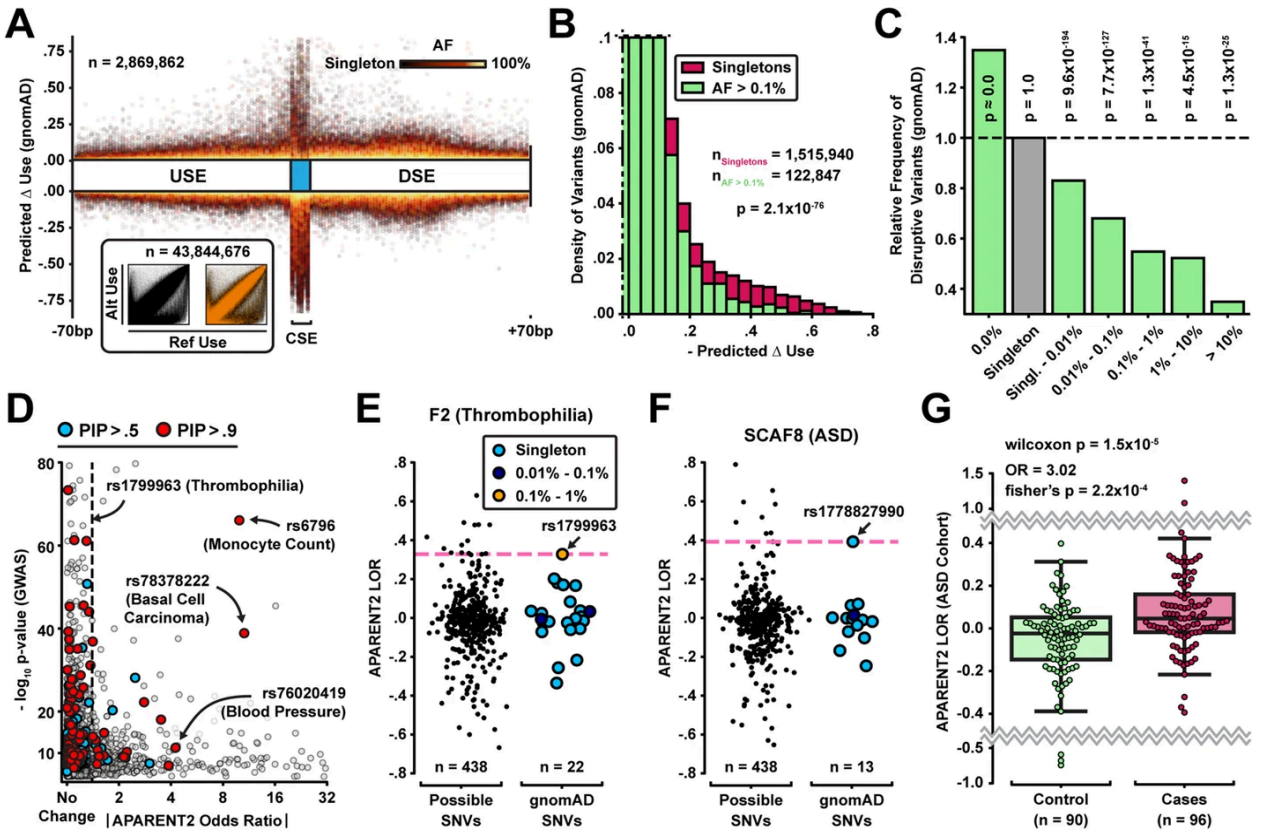


Figure 4.2 | Large-scale analysis of polyadenylation signal mutations and their implication in health and disease

(A) Relative position of mutation vs predicted isoform abundance for all PAS variants in gnomAD (n = 2.8 million). Color intensity represents allele frequency. Inset: Reference vs alternate isoform abundance for all 43.8 million potential PAS SNVs (orange = gnomAD variants). **(B)** Distribution of predicted isoform abundance for common gnomAD variants (AF >0.1%; green) and singletons (magenta). **(C)** Relative enrichment of disruptive variants (isoform abundance < −0.15) with respect to singleton variants. Wilcoxon p-values are shown above each bar. **(D)** Absolute predicted isoform fold change vs p-value of fine-mapped GWAS SNPs from CAUSALdb (95% credible set, n = 4200).[230] **(E)** Distribution of predicted log odds ratios for the F2 PAS. **(F)** Distribution of predicted log odds ratios for the SCAF8 PAS. **(G)**

Predicted log odds ratios among ASD cases and controls from a WGS study.[234] Figure 4.2 is from Linder J, **Koplik SE**, et al. (2022) *Genome Biol.*

Using the stratification process above, we investigated the link between misregulated polyadenylation and autism spectrum disorder (ASD), a relationship which has been suggested before but mainly at the *trans*-regulatory level and less in terms of *cis*-regulatory variation in the 3′ UTR.[7,9,10,15,235] **Figure 4.2F** displays an example rare variant (rs1778827990) associated with ASD.[8] The suspected variant has a considerably higher (positive) effect size than any of the observed variants in gnomAD. Hypothesizing that gain-of-function mutations may be linked to ASD, we ran APARENT2 predictions on whole-genome sequencing (WGS) data from 1902 families[234,236] and found that variants overlapping PASs in cases are enriched for gain-of-function compared to controls (Wilcoxon p = 0.049, n cases = 297, n controls = 296). When removing variants that co-occur with higher-impact common SNPs in gnomAD (AF >0.01%), the significance increased (Wilcoxon p = 2.1×10−4), and when also removing variants that occur in PASs with a protective downstream PAS within 200nt, the significance increased further (Wilcoxon p = 1.5×10−5; (**Figure 4.2G**, **Figure 4.3C**). We observed a 3.02-fold enrichment of gain-of-function mutations in cases (Fisher's p = 2.2×10−4). As additional validation, the predicted effect sizes of variants from the control set were indistinguishable from variants in gnomAD after applying the same filtering (Wilcoxon p = 0.341). When replicating the analysis against a smaller WGS study of 200 families, we again observed an enrichment of gain-of-function mutations in cases relative to the controls from An et al., but the results were only significant with less stringent filtering criteria (Wilcoxon p = 0.039; **Figure 4.3D**). [8,234]

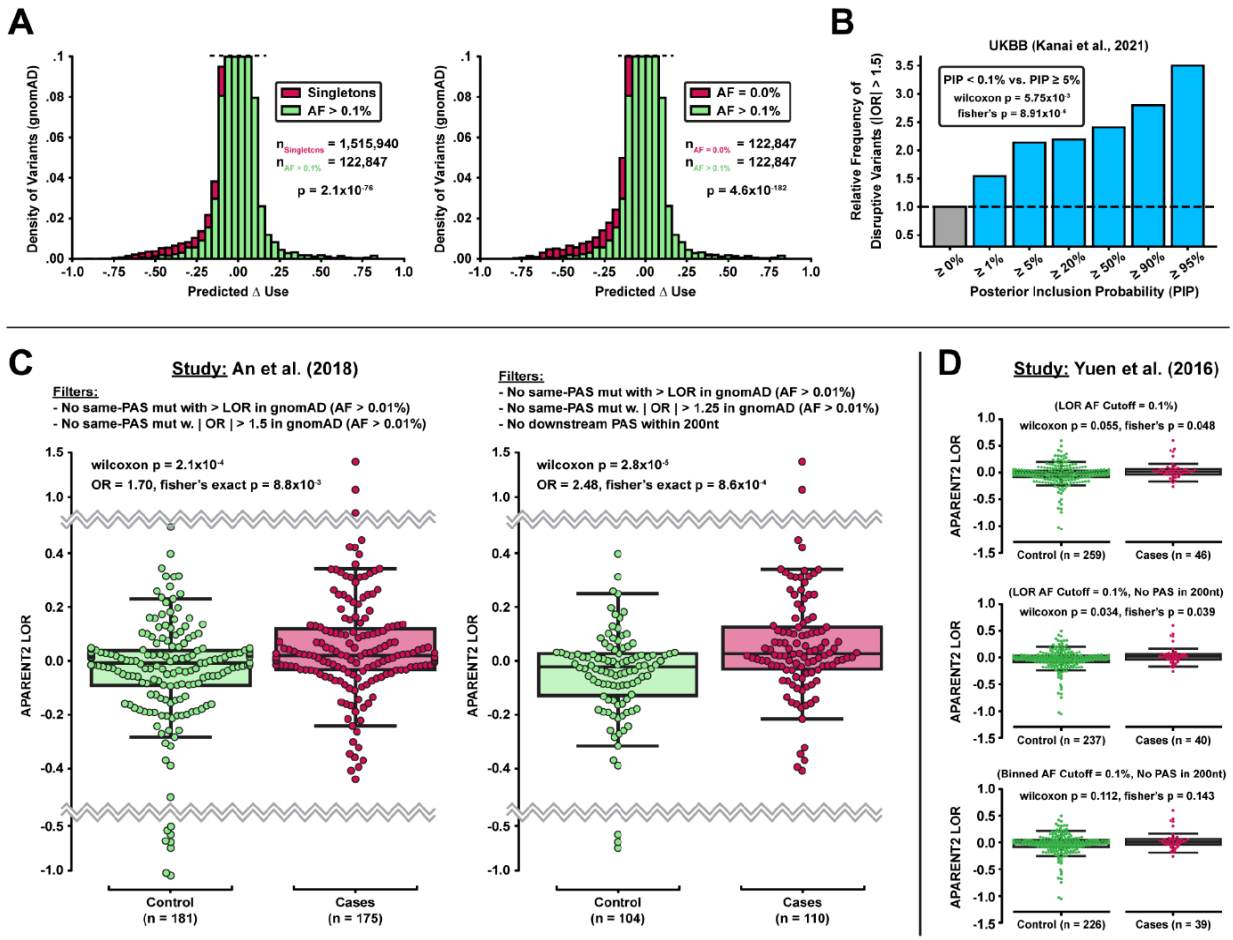


Figure 4.3 | Predicted impact of PAS variants on isoform usage and enrichment in population and disease cohorts

(A) Left: Predicted Δ isoform proportions for singletons (n = 1, 515, 940) and common variants (AF >0.1%; n = 122, 847) from gnomAD v3 that overlap annotated PASs in PolyADB V3. Right: Comparison of predictions for a matched set of unobserved PAS variants (AF = 0.0%; n = 122, 847) and common variants (AF >0.1%) from gnomAD**. (B)** Relative frequency (enrichment) of fine-mapped SNPs from UK Biobank that overlap annotated PASs (n = 90, 356) with a predicted absolute-valued isoform odds ratio >1.5, as a function of more stringent posterior inclusion probabilities (PIP). **(C)** Distribution of predicted isoform log odds ratios among ASD cases and controls from the WGS study of An et al. (2018). Left: Case- and control variants are removed if they occur in PASs that have common variants in gnomAD (AF >0.01%) with larger effect size (log odds ratio) than the investigated variant or common variants that have an absolute odds ratio larger than 1.5 (ncontrol = 181, ncases = 175). Right: Additional removal of variants that occur in PASs with a downstream PAS within 200nt in PolyADB V3 (ncontrol = 104, ncases = 110). A more stringent odds ratio threshold of 1.25 was used. This is the same filtering procedure as in main **Figure 4.2G**, but here the allele count of variants within the same PAS with similar effect sizes have not been aggregated prior to filtering against gnomAD. **(D)** Replicate analysis where the case variants are from the WGS study of 200 families from Yuen et al. (2016) (the control variants come from An et al., 2018). The three filtering steps used in

supplementary **Figure 4.2C** and main **Figure 4.3G** are also used here: (1) filtering variants with neighboring common mutations in gnomAD (AF >0.1%; $n_{control}$ = 259, n_{cases} = 46), (2) additionally removing variants with downstream protective PASs within 200nt ($n_{control}$ = 237, n_{cases} = 40), and (3) same filtering criteria as (2) but gnomAD AFs are re-calculated by aggregating allele counts of similar predicted effect size in the same PAS ($n_{control}$ = 226, n_{cases} = 39). Figure 4.3 is adapted from Linder J, **Koplik SE**, et al. (2022) *Genome Biol.*

In summary, our preliminary data support a functional role for 3′UTR variants in ASD. Variants overlapping PASs in ASD cases were enriched for gain-of-function effects relative to controls. At the same time, mutations are unlikely to be deleterious to polyadenylation if they occur within PASs that already harbor common variants with even larger effects, which are presumed to be benign. Accordingly, such PAS mutations can be classified as likely benign with respect to polyadenylation function when they co-occur with functional common variants in gnomAD that strongly influence cleavage and polyadenylation. After applying this heuristic, we observed an even stronger enrichment of gain-of-function mutations in ASD cases compared to controls.

4.4 An MPRA of clinically relevant variants in multiple cell lines

We ran APARENT2 on WGS as covered in **4.3.2** and found that variants overlapping PASs in ASD cases are enriched for gain-of-function compared to controls (**Figure 4.2G**, **Figure 4.3C-D**).[8,234,236] We found that in ASD cases are linked to gain-of-function mutations and that these gain-of-function effects can be predicted by APARENT2.

Massively parallel reporter assays (MPRAs) provide a powerful approach for assaying molecular phenotypes across hundreds to millions of sequences, and they are well-suited for validating model predictions, including those from APARENT2.[15,226,237] Since MPRAs are well-suited for this task, this motivated us to experimentally validate the enrichment of

gain-of-function mutations for APA in ASD cases suggested by APARENT2 predictions. Using an MPRA to assay APA, we tested 94 ASD-associated PAS variants in HEK293T, SK-N-SH, and microglia-derived HMC3 cells.

4.4.1 MPRA Design and Construction

To assess WGS variants in an MPRA context, we constructed libraries by cloning 250-nt oligo pools 25 nt upstream of a bGH polyadenylation signal in a mCherry reporter plasmid. Each construct contained a shared distal PAS (bGH) and a candidate proximal PAS upstream, such that isoform usage could be quantified by measuring the relative usage of the proximal versus distal site. The libraries included 76 predicted ASD variants (38 case variants and 38 matched controls), 9 GWAS SNPs, 3 hand-picked examples from the F2, SCAF8, and MECP2 genes, 2 aQTLs, 4 matched control SNPs from gnomAD, and 6 control PASs previously measured in the MPRA of Bogard et al.[226] The variants consisted of 38 case and 38 control variants from the ASD data set (19 variants with the largest positive effects and 19 variants with the largest negative effects for both cases and controls, after removing variants that occur in PASs with higher-impact common SNPs in gnomAD; AF> 0.01%), in addition to 9 GWAS SNPs with diverse predicted effects and other disease-relevant examples in the F2 and SCAF8 genes.

Sequences were organized into two matched libraries: (i) a variant library containing 94 case/control variants plus the 6 shared controls, and (ii) a reference library containing the corresponding 94 wild-type sequences plus the same 6 shared controls. Both libraries were transfected separately into HEK293T, SK-N-SH, and HMC3 cells, and reporter mRNA was quantified after Illumina sequencing (**Figure 4.4A**).

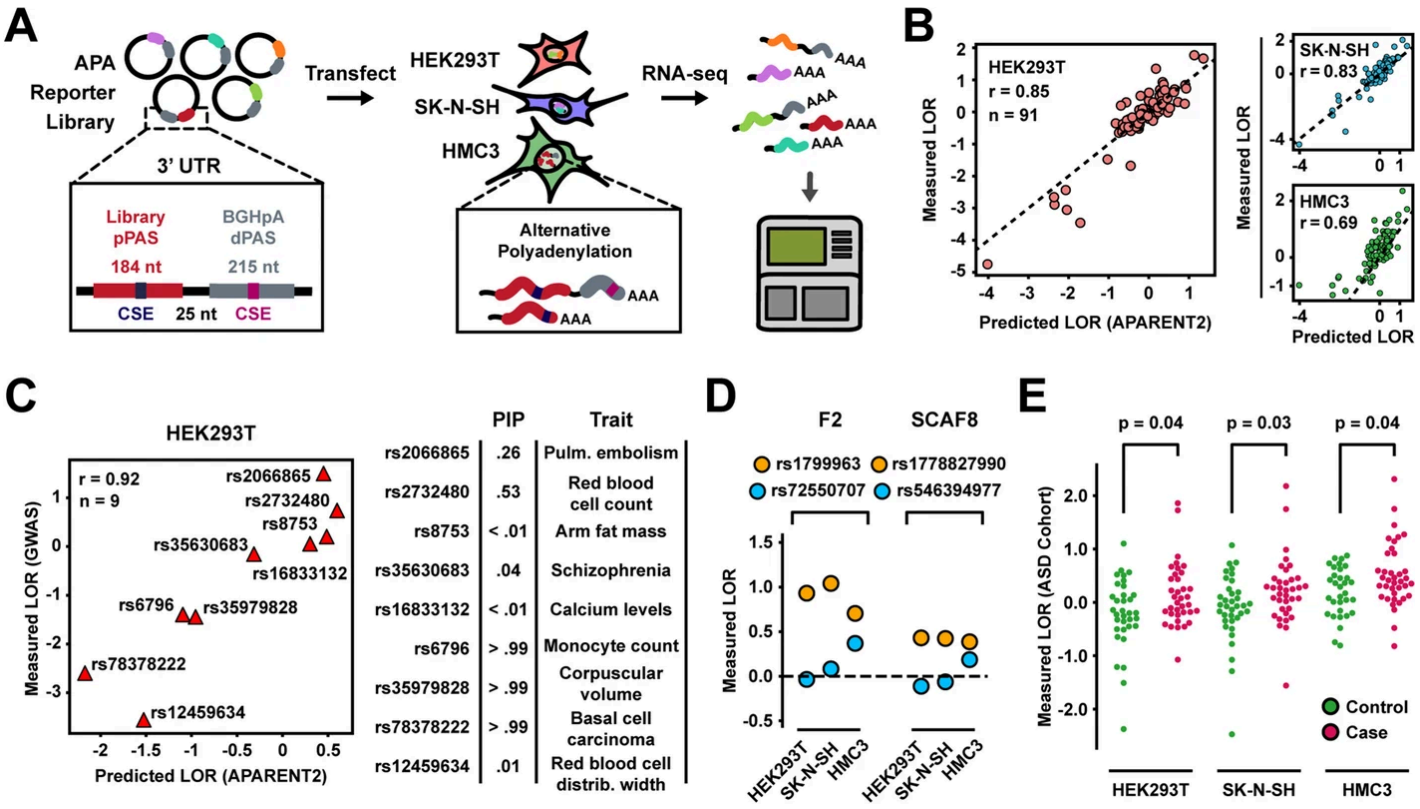


Figure 4.4 | MPRA measures functional impact of PAS variants in disease cohorts
(A) APA reporter system for measuring variant effects in HEK293T, SK-N-SH, and HMC3. **(B)** Predicted vs measured variant effects (LORs) in the three assayed cell lines. **(C)** Predicted vs measured effects of 9 GWAS SNPs (PIP = posterior inclusion probability). **(D)** Measured effects of 2 SNVs in the F2 and SCAF8 genes (orange), alongside common gnomAD SNPs (blue). **(E)** Measured effects of 76 autism variants from An et al. p-values are computed with two-sided Wilcoxon tests.[234]. Figure 4.4 is from Linder J, **Koplik SE**, et al. (2022) *Genome Biol.*

4.4.2 Comparing experimental results to APARENT2 predictions

To validate the predictions made for clinically relevant variants and to assess brain-specific effects, we experimentally measured the reference and variant MPRA libraries in HEK293T, SK-N-SH, and HMC3 cells (**Figure 4.5A-C**). APARENT2's Log odds ratio (LOR) predictions agreed well with the measurements in HEK293T and SK-N-SH (**Figure 4.4B** ; r = 0.85 in HEK293T and 0.83 in SK-N-SH) but were less concordant with HMC3 (r = 0.69), suggesting microglia-specific PAS usage modulation. APARENT2 could accurately predict the

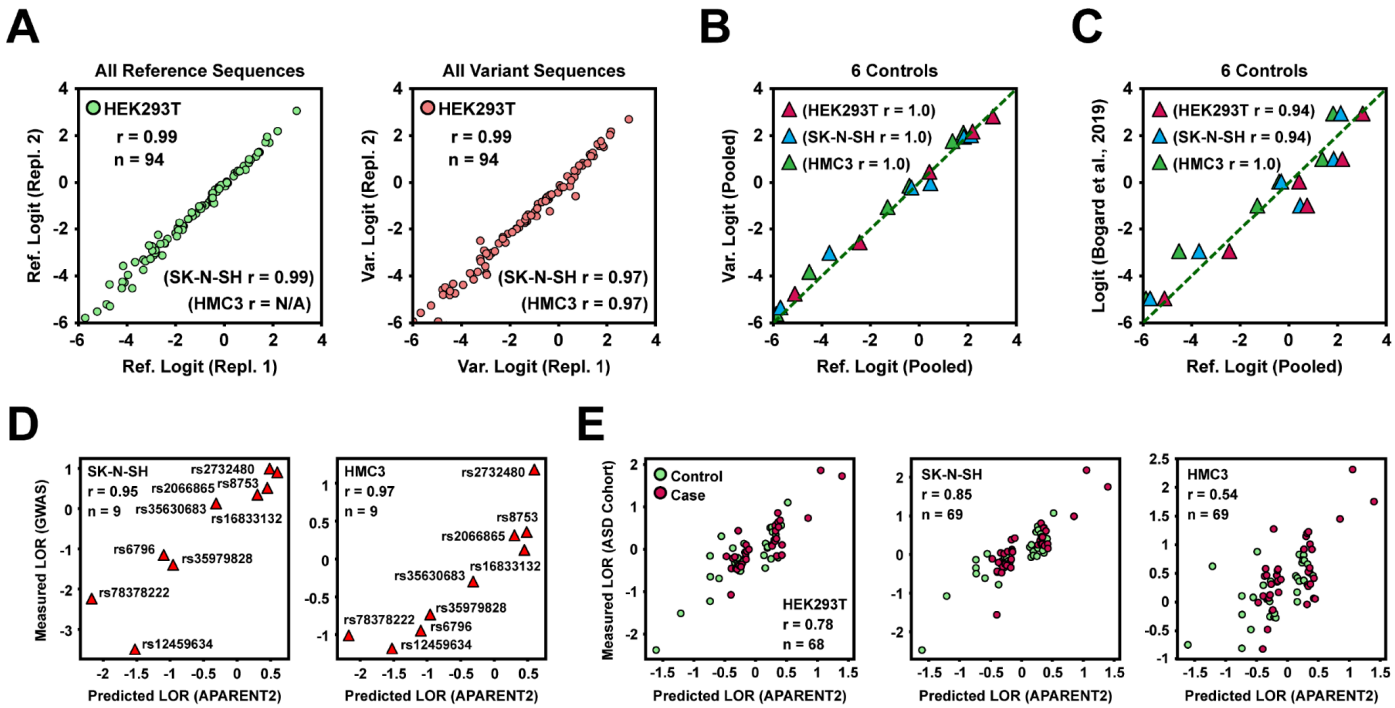


Figure 4.5 | Validation of PAS variant measurements in the plasmid reporter MPRA
(A) Replicate correlation of proximal isoform logits as measured in the plasmid reporter MPRA for the subset of the 100 assayed PASs that had a minimum of 5 supporting reads. Left: Measured logits of reference (wildtype) PASs. Right: Measured logits of variant (alternate) PASs.The scatter plots show the results for HEK293T, while summary correlation metrics for SK-N-SH and HMC3 are annotated in the plots. HMC3 had only one biological replicate with sufficient read depth for the reference library. **(B)** Proximal isoform logits of 6 control PASs that were assayed (without any designed mutations) in both the reference and variant libraries. The isoform logits were estimated by pooling the counts across replicates. **(C)** Correlation between the 6 control PASs of the reference library and their estimates from the MPRA of Bogard et al. (2019), where they had previously been measured. **(D)** Predicted vs measured variant effect sizes (log odds ratios, LORs) of 9 assayed GWAS SNPs. Results are shown for SK-N-SH and HMC3. **(E)** Predicted vs measured variant LORs of the subset of assayed Autism variants with a minimum of 50 reads (pooled replicates). Green = control variants, Magenta = case variants. Figure 4.5 is from Linder J, **Koplik SE**, et al. (2022) *Genome Biol.*

effects of the selected 9 GWAS SNPs (r = 0.92 in HEK293T; **Figure 4.4C**). However, the chosen variants that displayed loss-of-function (the majority of which were associated with cardiovascular traits) had less extreme effects in HMC3 (**Figure 4.5D**). Additionally, we validated the effects of dbSNP variants rs1799963 (F2; thrombophilia) and rs1778827990 (SCAF8; autism case) and found that these variants behave similarly across cell lines (**Fig 4.4D**).

Finally, we recapitulated a significant enrichment of gain-of-function variants in autism cases for all cell lines (**Figure 4.4E**, F**igure 4.5E;** Wilcoxon $p \leq 0.04$). There were ~3-fold more case variants with effect sizes larger than controls in HMC3 than there were in HEK293T and SK-N-SH, suggesting they have potential microglia-specific effects.

4.5 Conclusions

Understanding the regulatory code guiding APA has direct implications for both basic biology and medicine. APARENT2 is a sequence-to-function model that surpasses all prior approaches in predicting the effects of variants on human polyadenylation signals. By computationally assessing the impact of PAS mutations across the genome, intersecting these predictions with human variation data, and experimentally validating high-impact variants, we linked sequence effects on APA to phenotypic traits and conditions. Notably, APARENT2 revealed a three-fold enrichment of gain-of-function PAS variants in ASD cases, which we experimentally confirmed across HEK293, SK-N-SH, and microglia-derived HMC3 cells. Additionally we note the largest effect sizes were observed in HMC3, suggesting potential microglia-specific contributions to APA dysregulation in ASD.

PAS mutations with high predicted effect sizes cannot be definitively classified as causal, since both loss- and gain-of-function variants are common in controls and are also observed in patient cohorts across many diseases. Uniquely, in ASD, we observed a significant enrichment of gain-of-function PAS variants in cases compared to controls, suggesting that a subset of these variants contributes to disease risk. More broadly, this framework can be applied to other disorders to assess whether PAS variants are similarly enriched and may play a role in disease etiology.

4.6 Ongoing and future work for APA

4.6.1 Large-Scale APA MPRA for Systematic Screening of GWAS Variants

While GWAS have identified a number of genomic loci that are associated with many human diseases or disease risk, the majority of these variants lie in non-coding regions, and thus the mechanisms behind disease development remain largely elusive.[238] To address this challenge in the context of alternative polyadenylation (APA), we developed a high-throughput MPRA that assays more than 10,000 disease- and GWAS-linked sequences predicted to affect polyadenylation. Variants were first stratified using predictions from the APARENT2 model, ensuring that the library captured a broad range of predicted regulatory effects. We carried out this assay in HEK293 and K562, generating a dataset that captures how sequence variation influences APA across cell types. In doing so, we also highlight the limitations of the current APARENT2 model, which shows reduced performance on this dataset.

To improve predictive accuracy, we aim to adopt iterative cycles of model design, MPRA measurement, and retraining, a framework that has already been shown to enhance performance in enhancer design.[239,240] By iteratively refining model design with MPRA measurements, we anticipate improved accuracy and greater ability to prioritize variants most likely to disrupt APA. Experimental validation thus provides both mechanistic insight and new training data, enabling improved prioritization and characterization of disease-associated variants. Finally, the MPRA could enable sequence-resolved drug perturbations, including the testing of APA-modulating compounds such as JTE-607,[241] with the longer-term aim of developing treatments to correct aberrant polyadenylation.

4.6.2 Library design and MPRA workflow

Using APARENT2, we predicted thousands of disease-associated variants with potential impact on polyadenylation, which we have begun to validate through large-scale MPRAs encompassing more than 10,000 case and control sequences. These measurements are expected to improve our understanding of how genetic variation regulates polyadenylation, and to serve as rich datasets for refining the APARENT2 model. More than half of these sequences were derived from GWAS studies, enabling us to examine APA variants linked to a broad spectrum of diseases. The libraries also incorporated an expanded set of ASD whole-genome variants from our previous MPRA, as well as synthetic variants generated with the Scrambler architecture.[242]. In this framework, masks stochastically perturb the input sequence while preferentially retaining the most important nucleotides for APARENT2's predictions, thereby revealing regulatory elements that drive polyadenylation isoform choice. Mask-derived variants were included in our MPRA to extend discovery beyond naturally occurring variation. We have also included epidermal-specific polyadenylation site sources from previous studies in human keratinocytes.[243]

Using our established MPRA reporter plasmid, we cloned two separate libraries for variant and reference sequences, each containing over 5,000 constructs. These libraries have already been assayed in HEK293 and K562 cells, with plans to extend measurements to additional cell types including SK-N-SH (neuroblastoma), HaCaT (epidermal), HeLa (cervical cancer), and HMC3 (microglia). Together, these experiments will generate a dataset that supports model training for cell-type-specific APA prediction, enables identification of disease-relevant variant effects, and even provides a platform to test the impact of pharmacological modulators such as JTE-607 (**Figure 4.6**).[241]

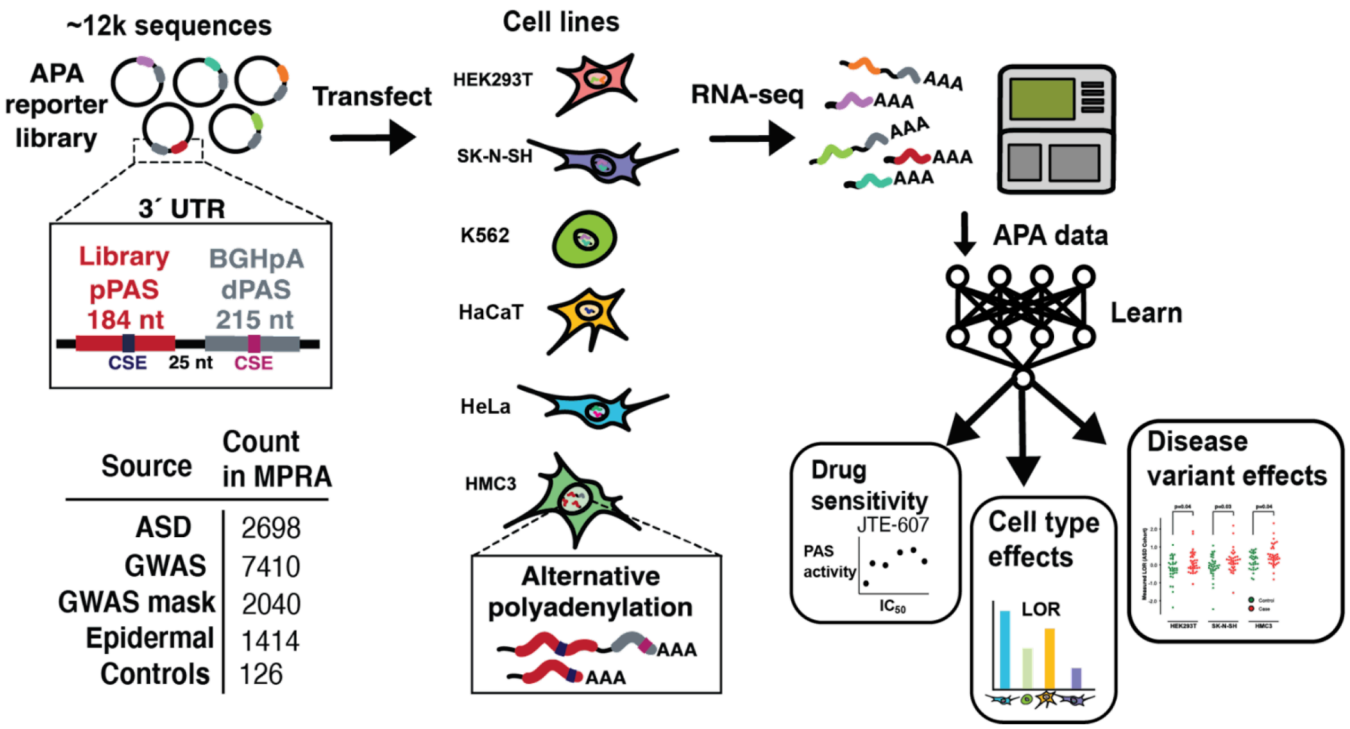


Figure 4.6 | A large-scale APA MPRA for systematic screening of GWAS variants
Diagram of the large-scale APA MPRA and its sequence sources. Libraries include disease-associated variants, GWAS hits, and synthetic controls. The assay has been implemented in HEK293 and K562 cells, with additional experiments planned in SK-N-SH, HaCaT, HeLa, and HMC3. Resulting mRNA isoforms are quantified by sequencing, and the data are used to train predictive models of APA. These models aim to capture drug sensitivity to the anticancer compound JTE-607, cell type-specific effects across diverse contexts, and variant-driven changes in APA linked to human disease.

4.6.3 Preliminary APA MPRA measurements in HEK293 and K562 reveal retraining opportunities for APARENT2

To date, we have performed experiments with this new MPRA in HEK293 and K562 cell lines, using paired variant and reference libraries. Each library was sequenced multiple times to maximize recovery of individual sequences, requiring at least 10 combined proximal and distal reads for inclusion. Out of 5,891 total possible library sequences in each library, we recovered 5,273 in HEK293 and 4,746 in K562 that had both of the paired variant and reference measurements. This yielded 10,546 measurements in HEK293 and 9,492 in K562, corresponding to variant and reference sequences, with the exception of a small subset of control sequences

from the 2019 APA MPRA that were intentionally identical across both reference and variant libraries.[226]

The replicate data of the logit scores in HEK293 is concordant for the variants (spearman r = 0.970) and references (spearman r = 0.981) (**Figure 4.7A**). We recovered 57 control sequences that were identical across both the variant and reference libraries to assess potential variability between libraries. For downstream analysis, reads were pooled across replicates for both variants and references. To confirm consistency, we compared the pooled logit values of these control sequences between the two libraries and found a very high correlation (Spearman r =0.996), indicating excellent consistency of measurements across libraries (**Figure 4.7B**). In K562, replicate concordance was similarly strong (variants: Spearman r =0.909 ; references : Spearman r = 0.945) (**Figure 4.7C**). Control sequences again demonstrated excellent reproducibility between the variant and reference libraries, with pooled logit values for these controls yielding a correlation of Spearman r = 0.970 between the two libraries (**Figure 4.7D**). Here, the logit value reflects the strength of a proximal polyadenylation site (PAS), as it represents the logit-transformed proportion of proximal isoform usage. For each variant, we estimated the log-odds ratio (LOR) of proximal usage relative to the corresponding reference sequence. Replicate concordance of measured LOR values was lower than in our previous MPRA, with Spearman r = 0.626 in HEK293 and r = 0.537 in K562 (**Figure 4.8A**), compared to r = 0.77 in the earlier assay of 94 variants.

Although the new libraries were sequenced to sufficient depth overall, individual read counts were generally lower than in the previous MPRA, suggesting that reduced coverage may introduce noise into replicate-level measurements. Nevertheless, when focusing on 93 of the same 94 sequences included in both MPRAs, the pooled LOR values from the current MPRA in

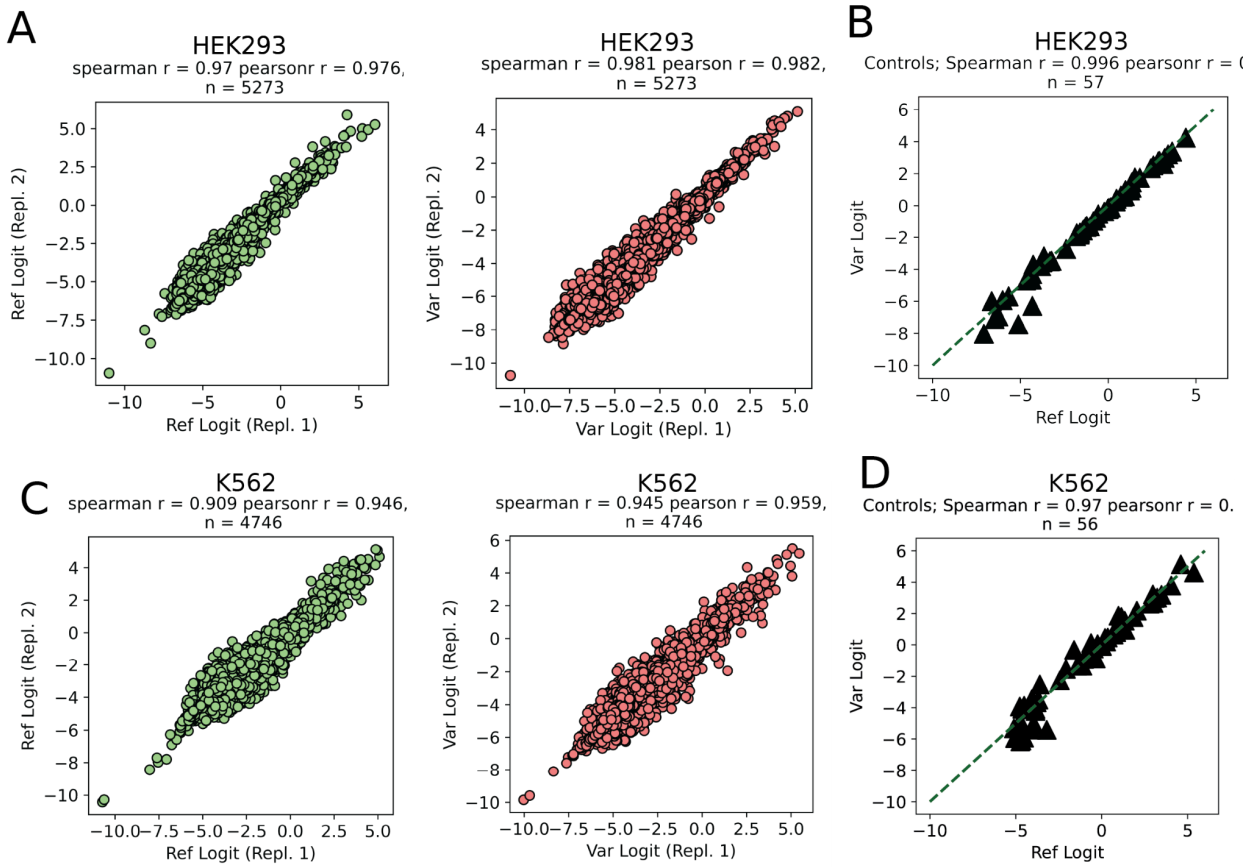


Figure 4.7 | Replicates and controls in large-scale APA MPRA measured in HEK293 and K562

(A) Replicate concordance of logit scores in HEK293 for variant sequences (Spearman r = 0.970, n = 5273) and reference sequences (Spearman r = 0.981, n = 5273). **(B)** Comparison of pooled logit values for 57 control sequences identical across variant and reference libraries in HEK293, showing excellent concordance (Spearman r = 0.996, n=57). **(C)** Replicate concordance of logit scores in K562 for variant sequences (Spearman r = 0.909, n = 4746) and reference sequences (Spearman r = 0.945, n = 4746). **(D)** Comparison of pooled logit values for 56 control sequences in K562, again showing strong agreement between variant and reference libraries (Spearman r = 0.970, n = 56).

HEK293 showed good agreement (Spearman r = 0.84, Pearson r =0.90) with the earlier measurements in HEK293 described in Section 4.4 (**Figure 4.8B**). This indicates that, despite reduced replicate concordance, many of the measurements could remain of sufficient quality for downstream analysis.

Next, we asked how well these measurements aligned with predictions from APARENT2 in order to begin benchmarking the model on our new library designs. We compared measured logit values (pooled across replicates) for both variants and references to their corresponding predicted logits. APARENT2 predicts logit scores for proximal and distal isoform usage, and these scores can be regressed against the experimentally measured isoform logits. This allows direct comparison of predicted versus observed logit values for both reference and variant sequences. In some cases, measured logit values were exactly 0, reflecting variants where all reads mapped to the proximal isoform and none to the distal isoform. As outlined in the **methods**, these represent boundary cases in which the proximal site is much stronger than the distal site, but the true magnitude of this effect cannot be quantified because of limited read depth. By contrast, APARENT2 predictions are continuous and rarely produce exact boundary values, instead estimating relative strength even when one site dominates. To avoid introducing artifacts into benchmarking, we excluded a handful of variants or references with measured logits equal to exactly 0, as these cases provide little quantitative resolution for comparing model predictions with experimental measurements. Future experiments with deeper sequencing may help resolve these boundary cases more reliably, enabling more complete benchmarking of model performance.

After filtering out these sequences, we observed strong agreement between experimental and predicted logit values. Both variant and reference logits correlated well with APARENT2 predictions, indicating that the model captures isoform usage patterns across these library designs with good accuracy (**Figure 4.8C,D**). However, the agreement between predicted and experimental LORs (variant effects) was weaker than the predictive correlations obtained for individual reference and variant logits.Across 4,042 sequences in HEK293 and 3,992 sequences

in K562, correlations between APARENT2-predicted and experimentally measured LORs were modest (Spearman r = 0.33 in HEK293 and r = 0.22 in K562), compared to our previously reported MPRA of 93 sequences, in which the weakest-performing cell line still achieved a Spearman correlation of 0.69 against APARENT2 prediction (HMC3, **Figure 4.4B**). We observed a substantial number of variants with low predicted LORs but with measurements showing large effects on APA, suggesting that some variant effects were not captured by the APARENT2 model. A logical next step will be to compare these sequences against existing functional assays to determine whether the experimental measurements are supported by orthogonal data, or whether these discrepancies reflect assay-specific artifacts.

To explain the weaker model performance of APARENT2 on this data, we explored potential sources, starting with one already accounted for in our analysis. First, at the level of absolute logit values, some reference or variant sequences produced measured logits of exactly zero, reflecting cases where all reads mapped to the proximal isoform. This outcome is expected when both PASs are strong and the proximal site is preferentially selected, consistent with a first-come, first-served mechanism reported in previously.226,244,245 This bias may contribute to the apparent artifact of measured logit values collapsing to zero when the proximal site dominates. Because these boundary cases mask the true strength of the proximal site, we filtered them out prior to subsequent comparisons with APARENT2 predictions.

To more fully characterize the intrinsic strength of these very strong PASs, a useful follow-up experiment would be to invert the reporter design: rather than perturbing the proximal site as the variable sequence of interest, the proximal PAS would be fixed and barcoded at the 5′ end of the 3’ UTR, while the variable region would sit as the distal site. This design would still quantify proximal versus distal PAS competition, but with their roles swapped, enabling

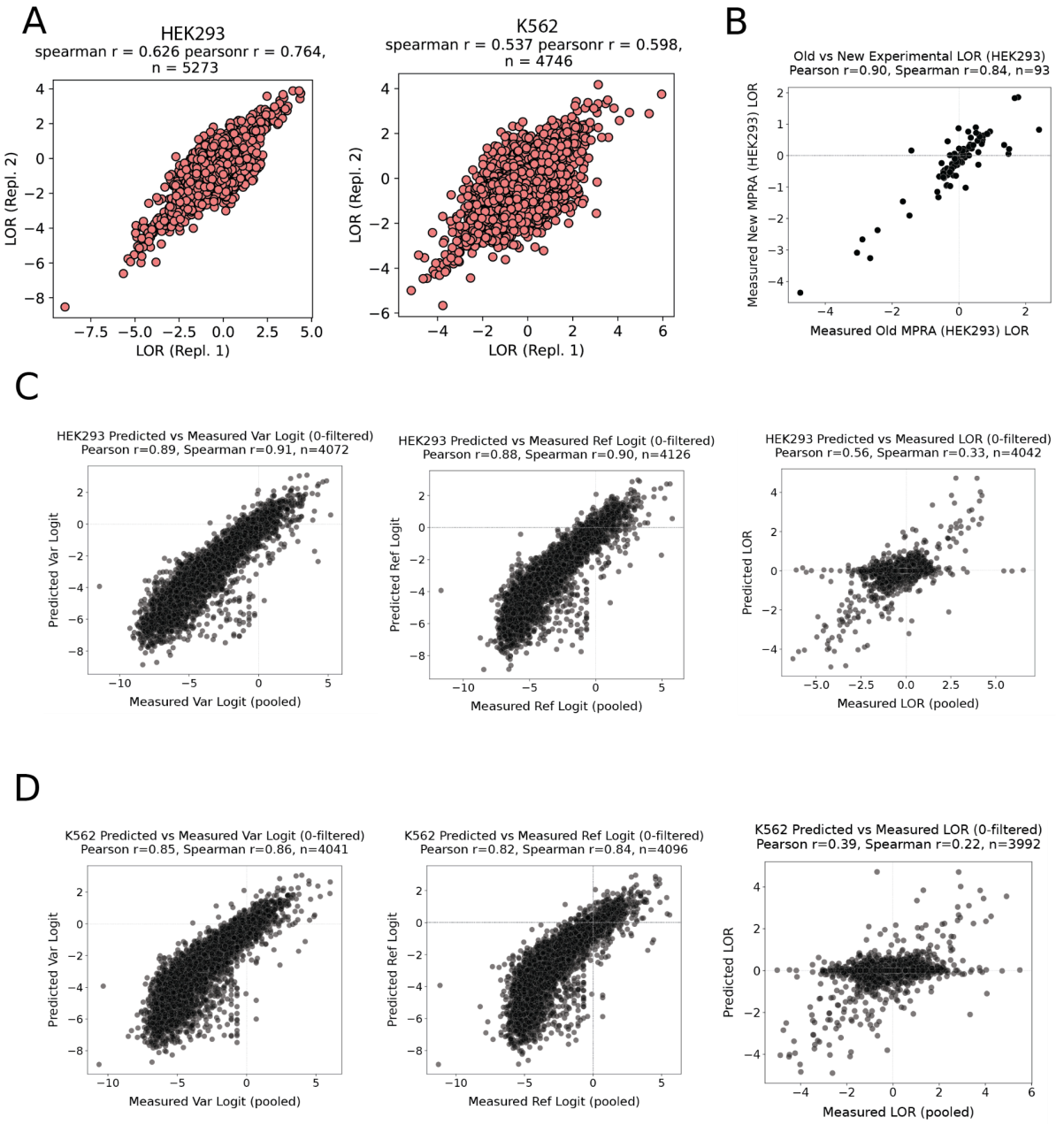


Figure 4.8 | Benchmarking APARENT2 predictions against experimental measurements
(A) Replicate concordance of LOR values in HEK293 and K562. **(B)** Comparison of measured HEK293 LORs for ASD variants between two independent MPRA libraries: the current study and the previously published *Genome Biology* assay (Spearman r = 0.84, n = 93) **(C)** HEK293: correlation between APARENT2-predicted and measured logit values for variants and references is shown on the left (variants: Spearman r = 0.91, n = 4072; references: Spearman r = 0.90, n = 4026), and correlation between predicted and measured LOR values is shown on the right (Spearman r = 0.33, n = 4042). **(D)** K562: same as in (C), with logit correlations shown on the left (variants: Spearman r = 0.86, n = 4041; references: Spearman r = 0.84, n = 4096) and LOR correlations on the right (Spearman r = 0.22, n = 3992).

measurement of strong distal PASs of interest relative to a fixed proximal PAS, which is preferentially used. Importantly, this strategy would likely require testing a range of proximal PAS strengths to identify a level that permits detectable competition from the distal site, thereby enabling more accurate quantification of strong PAS activity.

Second, at the level of LORs (Δlogit values between variant and reference), APARENT2 generally predicted only minor shifts in proximal PAS usage, whereas experimental measurements often revealed larger differences (**Figure 4.8C,D**). This suggests that the APARENT2 model underestimates the magnitude of variant-driven effects on PAS selection in the context of this specific MPRA. We hypothesize that this discrepancy arises because APARENT2 was trained on large randomized sequence libraries, which capture general sequence features but may underestimate the effects of single-base perturbations.

Because APARENT2 was trained on data from HEK293 cells, it may not fully capture the cell type–specific intricacies of APA in other contexts. For example, prior studies have shown that some cell types preferentially use proximal sites, whereas others favor distal sites, producing longer 3′ UTRs.[246] Performing similar APA MPRA measurements in additional cell types will therefore be important both for understanding cell type–specific effects and for refining models like APARENT2 to more accurately predict APA outcomes across diverse cell types.

Together, these observations point to a central limitation of APARENT2: its training data were derived largely from synthetic MPRA libraries in HEK293 cells. APARENT2 was trained on largely the same synthetic MPRA data as the original APARENT (Bogard et al., 2019), consisting primarily of randomized sequences.[226] For APARENT2, data were re-processed to standardize sequence length, recover missing cleavage distributions, and expand the training set

from ~2.4 to ~3.3 million sequences.[64] In parallel, the network architecture was upgraded to a deep residual convolutional design, enabling base-resolution prediction of cleavage profiles rather than a single proximal usage value. As a result, APARENT2 generalizes more effectively than APARENT across PAS contexts, although both models remain fundamentally limited by their reliance on data derived largely from randomized, degenerate libraries rather than human-derived MPRA sequences. Future efforts could benefit from re-training or fine-tuning APARENT2 on our measurements from this latest MPRA and on other datasets that report native human isoform usage, such as aQTLs, which directly measure allele-specific changes in proximal and distal PAS choice across tissues.[211]

Notably, a related approach was implemented in APARENT2, where tissue-resolved predictions were achieved by introducing a softmax layer and scaling baseline variant effect predictions against GTEx aQTLs. In practice, the model used tissue- specific APA datasets to adjust APARENT2's predictions through the linear framework, similar to how tissue specific splicing is predicted with MTSplice.[15,71,212] This strategy highlights the potential of leveraging native human isoform usage data from aQTLs to extend beyond plasmid reporter training contexts.

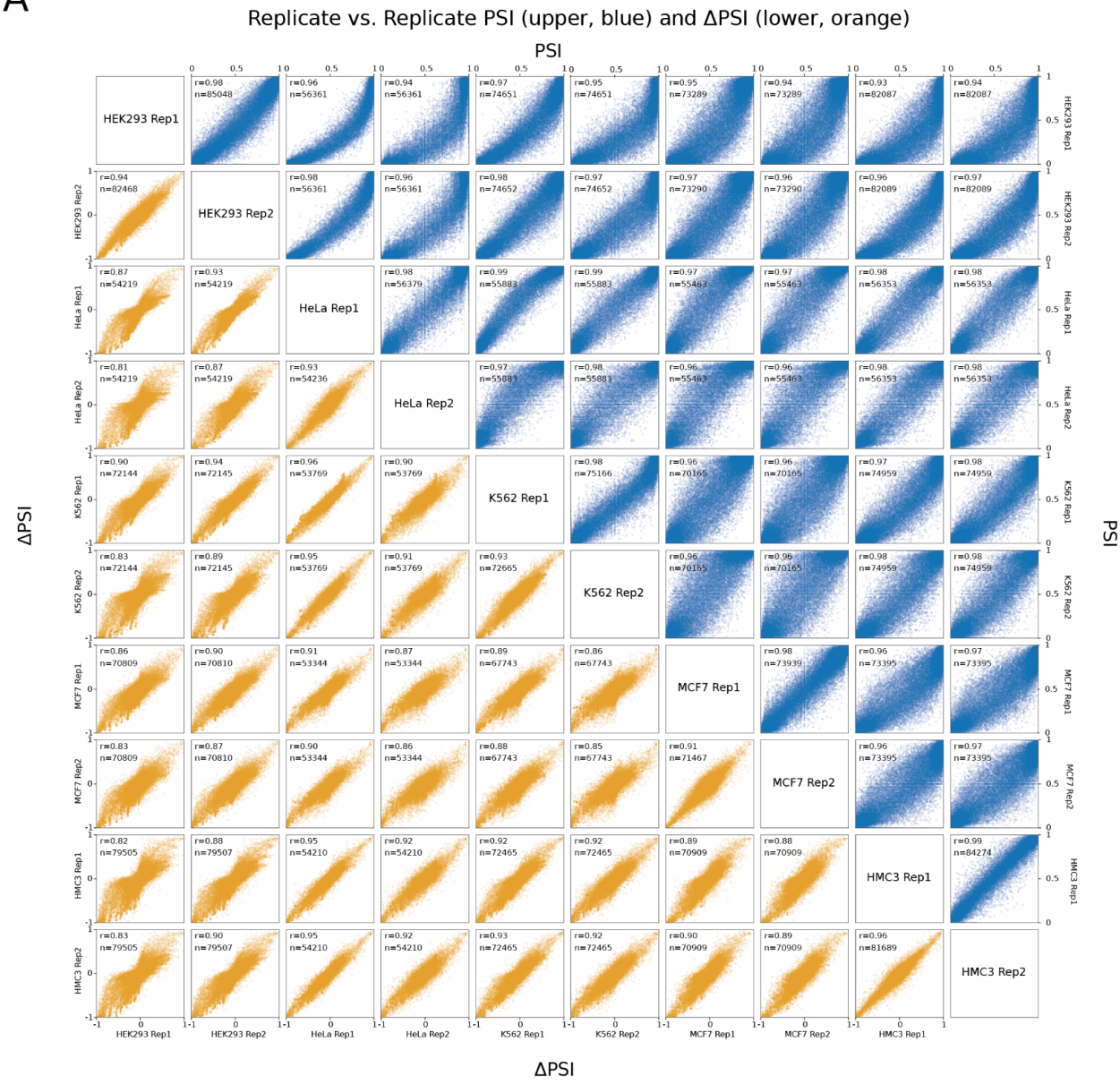
4.6.4 Future Plans for APA MPRAs

Besides efforts to retrain APARENT2 on the new data, I am also working to extend these experiments into additional cellular contexts. For example, HaCaT cells have been shown to be an attractive epidermal cell model and will be used to investigate the epidermal-specific APA variants included in our MPRA.[247] We also hope to include additional cell lines such as HeLa, SK-N-SH, and HMC3 cells. I have already begun some initial experiments in HeLa cells,

including drug perturbations with JTE-607. Predictions on our MPRA sequences using C3PO, an ML model developed by Angela Yu and collaborators, suggest that these PASs exhibit a wide range of drug responsiveness.[241] To validate these predictions, I aim to expand MPRA measurements in HeLa cells to also include JTE-607 drug perturbation experiments, following previously established protocols[241,248] Looking ahead, I plan to retrain APARENT2 on our existing HEK293 and K562 data, with the broader goal of developing models that more accurately capture variant effects on isoform usage in human-derived sequences across diverse cell types. In parallel, these efforts aim to inform therapeutic target discovery and enable systematic evaluation of compounds such as JTE-607 for correcting aberrant polyadenylation.

Supplementary Figures

A


Replicate vs. Replicate PSI (upper, blue) and ΔPSI (lower, orange)
PSI
ΔPSI
HEK293 Rep1
HEK293 Rep2
HeLa Rep1
HeLa Rep2
K562 Rep1
K562 Rep2
MCF7 Rep1
MCF7 Rep2
HMC3 Rep1
HMC3 Rep2

B

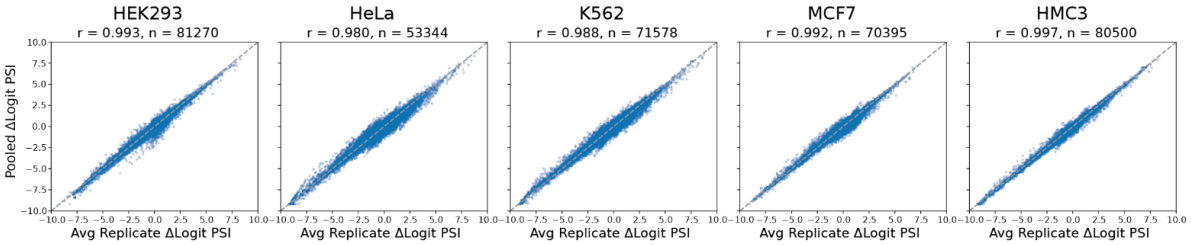
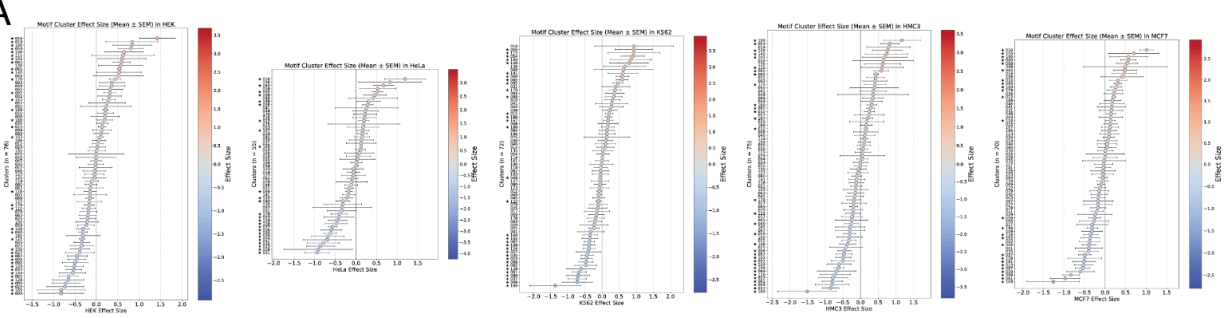

HEK293
r = 0.993, n = 81270
HeLa
r = 0.980, n = 53344
K562
r = 0.988, n = 71578
MCF7
r = 0.992, n = 70395
HMC3
r = 0.997, n = 80500
Pooled ΔLogit PSI
Avg Replicate ΔLogit PSI

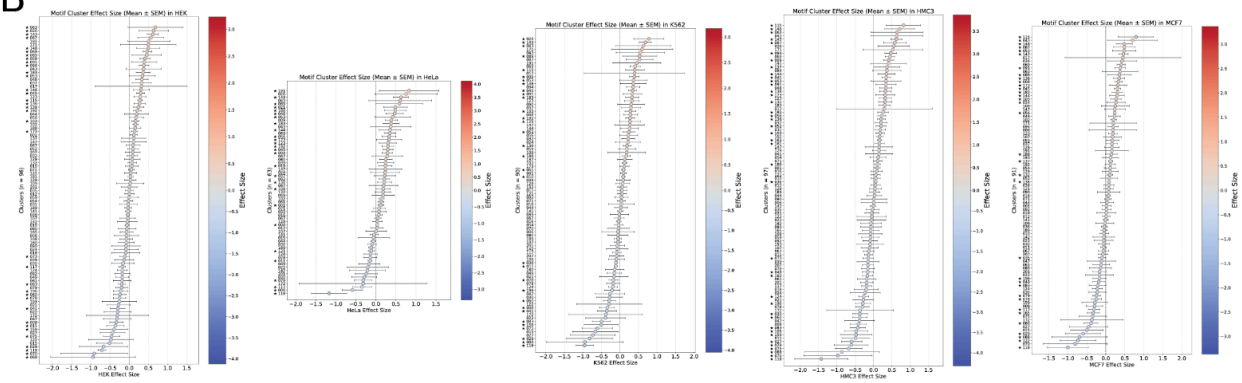
Figure S1 | Correlations of cell line replicates (PSI and ΔPSI) and replicate-averaged versus pooled Δlogit(PSI)

(A) PSIs are shown in the upper triangle in blue, covering two replicates each for HEK293, HeLa, K562, MCF7, and HMC3. ΔPSIs are shown in the lower triangle in orange, covering two replicates each for HEK293, HeLa, K562, MCF7, and HMC3. Each panel displays the number of sequences included and the Pearson r. **(B)** To maximize sequencing depth per measurement, replicates were pooled to compute a single Δlogit(PSI) value per variant sequence. Comparison with Δlogit(PSI) values obtained by averaging across individual replicates confirmed that pooling yields highly consistent results in all five cell lines: HEK293 ($r = 0.993$, $n = 81{,}270$), HeLa ($r = 0.980$, $n = 53{,}344$), K562 ($r = 0.988$, $n = 71{,}578$), MCF7 ($r = 0.992$, $n = 70{,}395$), and HMC3 ($r = 0.997$, $n = 80{,}500$).

A



B

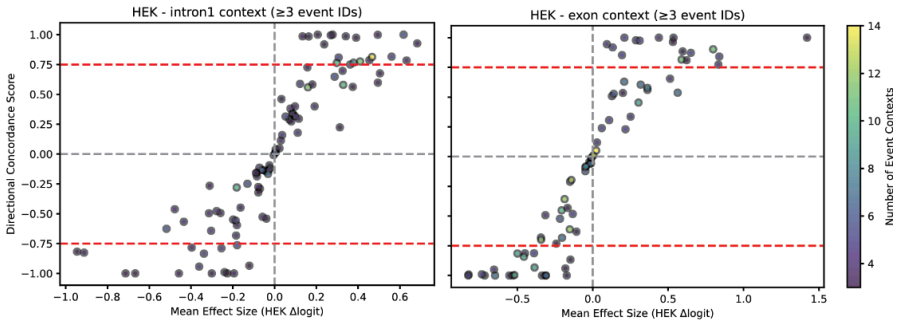


C

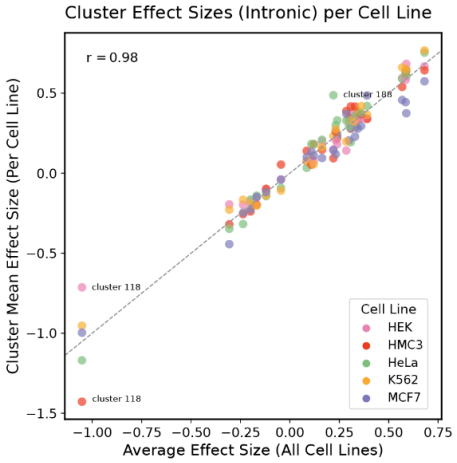
Directional Concordance =

$$\frac{\sum(\text{sign} \cdot |\text{effect}|)}{\sum|\text{effect}|}$$

Score ranges from -1 (all neg) to +1 (all pos); near 0 = mixed



D



E

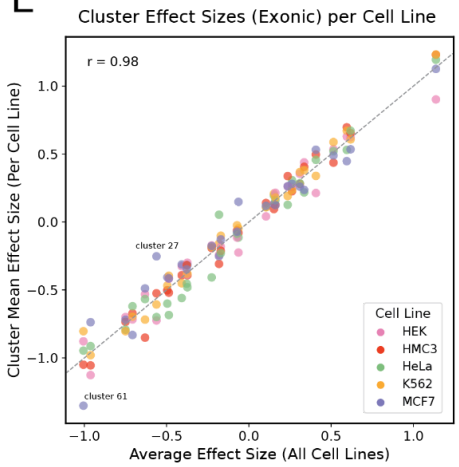


Figure S2 | Concordant RBP motif effect sizes in exonic and intronic regions
(A) Exonic region motif effect sizes across five human cell lines (HEK293, HeLa, K562, MCF7, and HMC3). Motif clusters with effect size estimates in at least three exon families are shown. Each point represents the mean Δlogit(PSI) across exon families in a given cell line, with error bars denoting the SEM. Clusters marked with a star pass the concordance filter (≥75% agreement in effect direction across exon families).
(B) Intronic regions. The same analysis as in (A) was applied to motif clusters located in intronic regions.
(C) Concordance filtering of motif effect sizes in HEK293 as an example cell line. This analysis was applied to all cell lines. Motif clusters were included in this analysis if effect sizes were measurable in at least three exon families within each region (intron 1 on the left, exon on the right) for that cell line. Each point represents a motif cluster, with the point color indicating the number of exon families contributing to the calculation. Directional concordance was quantified as the sum of signed effect sizes divided by the sum of absolute effect sizes across exon families. Clusters with strongly correlated effects across exon families approach values of -1 or 1, whereas clusters with weak or inconsistent effects approach 0. Here, we applied a filter of $|\text{concordance}| \geq 0.75$ to highlight motif clusters with strong and consistent effects across exon families.
(D) Exonic regions. To assess differences across cell types, the mean effect size of each motif cluster was calculated by averaging the effect sizes for a motif cluster across all cell lines (x-axis) and comparing this to the individual cluster means within each cell line (y-axis). Motif clusters included here were selected from the subsets that passed concordance filtering in (C) within each cell line. Points deviating from the diagonal indicate potential cell type-specific effects, such as in cluster 118.
(E) Intronic regions. The same analysis as in (D) applied to motif clusters in intronic regions. Across both contexts, effect sizes were highly correlated across cell lines, indicating that this analysis predominantly captures general *cis*-regulatory effects of motifs rather than cell type-specific activity. Points deviating from the diagonal indicate potential cell type-specific effects, such as in clusters 61 or 27.

Supplementary Tables

Gene & Exon	Full Sequence (hg38)	Exon Start-End (hg38)	Reference PSI	ΔLogit Min	ΔLogit Max	Number of Variants
ACADVL exon 3	chr17:7220389-7220549:+	7220464-7220529	0.2380	-3.2959	4.2370	568
ADA exon 7	chr20:44622987-44623147:-	44623056-44623127	0.6809	-5.6119	2.8908	711
ADAM15 exon 20	chr1:155061349-155061509:+	155061418-155061489	0.0036	0.0000	6.3635	2094
AFF2 exon 5	chrX:148837593-148837753:+	148837647-148837733	0.9515	-7.9371	1.5591	530
ALDH6A1 exon 3	chr14:74072517-74072677:-	74072583-74072657	0.9747	-9.1901	0.7181	458
ALDOA exon 7	chr16:30069249-30069409:+	30069306-30069389	0.9808	-8.4104	1.2073	787
ANK1 exon 41	chr8:41661856-41662016:-	41661931-41661996	0.9714	-8.5841	1.4020	1231
ANK3 exon 23	chr10:60166571-60166731:-	60166649-60166711	0.2345	-3.3940	5.2235	344
AP4M1 exon 3	chr7:100102606-100102766:+	100102675-100102746	0.0000	0.0000	0.0000	401
AP4S1 exon 5	chr14:31080504-31080664:+	31080573-31080644	0.9802	-8.6137	1.3463	412
ASL exon 13	chr7:66089195-66089355:+	66089276-66089335	0.1266	-3.4185	5.3007	646
ASS1 exon 5	chr9:130464027-130464187:+	130464111-130464167	0.9712	-8.4237	1.3523	711
AUH exon 9	chr9:91216039-91216199:-	91216132-91216179	0.6251	-5.3617	4.1595	384
BCKDHA exon 3	chr19:41410869-41411029:+	41410923-41411009	0.8692	-6.8148	3.4326	757
BEST1 exon 8	chr11:61959438-61959598:+	61959498-61959578	0.8397	-6.4202	1.9738	610
BIN1 exon 16	chr2:127051134-127051294:-	127051185-127051274	0.2405	-4.4150	4.9946	726
BLOC1S6 exon 3	chr15:45603047-45603207:+	45603104-45603187	0.1382	-2.8764	4.3883	333
CACNA1C exon 10	chr12:2547384-2547544:+	2547450-2547524	0.2930	-4.2775	3.8322	582
CACNA1C exon 30	chr12:2634240-2634400:+	2634297-2634380	0.0000	0.0000	4.5991	544
CACNA1C exon 31	chr12:2633572-2633732:+	2633629-2633712	0.8969	-7.0204	2.7124	518
CACNA1C exon 31	chr12:2648367-2648527:+	2648475-2648507	0.0017	0.0000	1.5853	474
CACNA2D4 exon 28	chr12:1810523-1810683:-	1810619-1810663	0.4584	-4.5835	4.1865	724
CASK exon 19	chrX:41557012-41557172:-	41557084-41557152	0.9231	-7.4050	1.8500	318
CLN3 exon 5	chr16:28488571-28488731:-	28488640-28488711	0.9355	-7.7806	2.4959	1062
COL11A2 exon 6	chr6:33185681-33185841:-	33185744-33185821	0.0014	0.0000	1.9548	947
COL11A2 exon 7	chr6:33184972-33185132:-	33185050-33185112	0.5634	-4.8372	4.3156	1838
COLQ exon 5	chr3:15478957-15479117:-	15479071-15479097	0.3083	-3.7620	3.2489	791
DEPDC5 exon 33	chr22:31861293-31861453:+	31861368-31861433	0.9128	-7.0807	2.0162	852
DEPDC5 exon 6	chr22:31766528-31766688:+	31766585-31766668	0.9287	-7.4663	1.7239	411
DMD exon 71	chrX:31177912-31178072:-	31178014-31178052	0.9712	-8.2383	1.0030	375
DNM1L exon 15	chr12:32737802-32737962:+	32737865-32737942	0.0828	-2.8745	5.4149	697
DNM1L exon 4	chr12:32708084-32708244:+	32708153-32708224	0.9452	-7.8214	1.5542	351
DNM1L exon 5	chr12:32710875-32711035:+	32710929-32711015	0.9708	-8.2392	1.8745	316
DRD2 exon 6	chr11:113414355-113414515:-	113414409-113414495	0.9836	-8.7065	1.4388	788
DYSF exon 17	chr2:71549251-71549411:+	71549350-71549391	0.1158	-2.4376	5.9357	860
ECEL1 exon 15	chr2:232481071-232481231:-	232481146-232481211	0.0008	0.0000	1.1710	722
ELN exon 13	chr7:74047576-74047736:+	74047675-74047716	0.3734	-4.3715	4.7629	541
ELN exon 24	chr7:74060044-74060204:+	74060140-74060184	0.0489	-1.9645	6.9501	806
ELN exon 5	chr7:74041111-74041271:+	74041216-74041251	0.4799	-4.6034	4.3761	810
EPB41L1 exon 13	chr20:36195224-36195384:+	36195329-36195364	0.3811	-4.0858	4.0838	688
EPB41L1 exon 18	chr20:36218822-36218982:+	36218876-36218962	0.9650	-8.1927	0.9764	756
ERCC2 exon 3	chr19:45369050-45369210:-	45369113-45369190	0.4489	-5.2168	4.5096	679
FANCA exon 38	chr16:89740784-89740944:-	89740862-89740924	0.9823	-8.6962	0.4906	302
FGF8 exon 3	chr10:101775110-101775270:-	101775164-101775250	0.0006	0.0000	2.4157	1548
GALC exon 2	chr14:87988435-87988595:-	87988507-87988575	0.9749	-8.4716	0.8123	359
GCDH exon 5	chr19:12892038-12892198:+	12892116-12892178	0.0717	-2.2080	6.1914	603
GDAP1 exon 2	chr8:74351181-74351341:+	74351274-74351321	0.9380	-8.1156	1.1216	347
GH1 exon 3	chr17:63917997-63918157:-	63918063-63918137	0.0078	-0.4091	2.4319	352
GNAS exon 3	chr20:58898845-58899005:+	58898941-58898985	0.1956	-3.3078	4.8470	1220
GNS exon 2	chr12:64752678-64752838:-	64752759-64752818	0.9831	-8.7494	0.4521	364
GPI exon 5	chr19:34377446-34377606:+	34377503-34377586	0.8374	-6.5666	2.4973	767
GUSB exon 3	chr7:65979814-65979974:-	65979877-65979954	0.0002	0.0000	3.7914	760
HARS2 exon 2	chr5:140693525-140693685:+	140693591-140693665	0.4898	-4.6065	3.7368	533
HMBS exon 10	chr11:119092023-119092183:+	119092125-119092163	0.0533	-2.2328	4.6302	803
HMBS exon 2	chr11:119088168-119088328:+	119088255-119088308	0.4738	-4.5249	2.8189	834
HMBS exon 6	chr11:119089620-119089780:+	119089683-119089760	0.3941	-4.1780	4.7774	421
IL2RA exon 5	chr10:6019850-6020010:-	6019919-6019990	0.9777	-8.5023	1.5411	400
IMPDH1 exon 4	chr7:128400812-128400972:-	128400893-128400952	0.0001	0.0000	2.0067	769
IMPDH1 exon 7	chr7:128400797-128400957:-	128400863-128400937	0.0250	-1.7823	4.7588	706

IVD exon 2	chr15:40407585-40407745:+	40407636-40407725	0.5111	-5.3611	3.8292	745
KCNQ2 exon 9	chr20:63431320-63431480:-	63431431-63431460	0.0003	0.0000	2.0343	440
L1CAM exon 3	chrX:153873208-153873368:-	153873334-153873348	0.0016	0.0000	1.6138	938
LARS1 exon 4	chr5:146171890-146172050:-	146171950-146172030	0.9835	-3.3054	0.6748	377
LMNA exon 10	chr1:156137603-156137763:+	156137654-156137743	0.9665	-8.2386	1.0171	839
MAPT exon 9	chr17:45993837-45993997:+	45993924-45993977	0.0010	0.0000	7.8363	573
MEF2C exon 5	chr5:88761249-88761409:-	88761330-88761389	0.0006	0.0000	7.5122	636
MEF2C exon 8	chr5:88730191-88730351:-	88730308-88730331	0.0150	-0.9271	3.8643	433
MEFV exon 8	chr16:3244234-3244394:-	3244342-3244374	0.6251	-5.1136	3.0179	730
MFAP5 exon 3	chr12:8660843-8661003:-	8660948-8660983	0.9529	-7.9262	1.5937	430
MLC1 exon 3	chr22:50083064-50083224:-	50083115-50083204	0.9727	-8.4516	1.1937	628
MLC1 exon 4	chr22:50080324-50080484:-	50080411-50080464	0.9206	-7.1217	1.9813	860
MUTYH exon 6	chr1:45332898-45333058:-	45332997-45333038	0.0034	0.0000	5.5971	774
MYBPC1 exon 4	chr12:101626770-101626930:+	101626872-101626910	0.9240	-7.2552	1.8918	748
MYH11 exon 42	chr16:15708783-15708943:-	15708885-15708923	0.6108	-5.4515	3.7273	698
MYH11 exon 6	chr16:15784678-15784838:-	15784798-15784818	0.0028	-0.0093	0.5977	430
NDUFAF5 exon 9	chr20:13816406-13816566:+	13816463-13816546	0.8672	-6.8340	2.3726	638
NEK1 exon 20	chr4:169508749-169508909:-	169508806-169508889	0.3742	-4.2250	4.3460	311
NF1 exon 31	chr17:31252860-31253020:+	31252938-31253000	0.9057	-6.9783	2.3792	445
NLRC3 exon 14	chr16:3548650-3548810:-	3548707-3548790	0.7144	-5.6814	2.7336	438
NPC2 exon 4	chr14:74480682-74480842:-	74480745-74480822	0.9855	-8.3815	1.2232	373
NRXN1 exon 8	chr2:50621185-50621345:-	50621281-50621325	0.0000	0.0000	2.9576	313
OCA2 exon 10	chr15:27990556-27990716:-	27990625-27990696	0.9203	-7.6589	1.8333	706
PAX5 exon 9	chr9:36846823-36846983:-	36846877-36846963	0.9819	-8.7112	1.4406	1724
PCCA exon 2	chr13:100102820-100102980:+	100102883-100102960	0.9797	-8.6153	1.3080	324
PDE6B exon 8	chr4:656152-656312:+	656245-656292	0.5490	-5.3385	3.9616	660
PIGH exon 3	chr14:67592615-67592775:-	67592675-67592755	0.9111	-7.1631	2.0262	507
PPT1 exon 8	chr1:40076822-40076982:-	40076891-40076962	0.9756	-8.4246	1.7714	402
RAB7A exon 4	chr3:128806309-128806469:+	128806372-128806449	0.9500	-8.8193	1.3547	745
RYR1 exon 94	chr19:38570553-38570713:+	38570607-38570693	0.1087	-3.0521	4.2817	572
SCN5A exon 24	chr3:38557211-38557371:-	38557298-38557351	0.8511	-6.4295	2.8010	577
SERPINB7 exon 3	chr18:63792303-63792463:+	63792393-63792443	0.9807	-8.6203	1.3284	424
SHOX2 exon 2	chr3:158105038-158105198:-	158105107-158105178	0.0477	-2.1311	7.6068	1010
SLC13A5 exon 2	chr17:6707008-6707168:-	6707131-6707148	0.0680	-1.6083	5.4041	304
SLC26A4 exon 11	chr7:107694340-107694500:+	107694403-107694480	0.6373	-5.1567	3.3187	557
SLC26A5 exon 12	chr7:103390409-103390569:-	103390472-103390549	0.7272	-5.6380	3.8558	555
SNCA exon 3	chr4:89828123-89828283:-	89828222-89828263	0.9843	-9.1902	0.6244	417
SOS1 exon 21	chr2:38989250-38989410:-	38989346-38989390	0.1462	-2.7697	5.6752	365
SPINT2 exon 3	chr19:38287795-38287955:+	38287876-38287935	0.9704	-8.4735	1.1078	548
STXBP2 exon 13	chr19:7643105-7643265:+	7643165-7643245	0.0017	0.0000	4.6999	946
SUMF1 exon 3	chr3:4449246-4449406:-	4449312-4449386	0.9637	-8.1277	1.6723	326
SUMF1 exon 8	chr3:4376310-4376470:-	4376391-4376450	0.9880	-7.4389	0.9993	354
SYNGAP1 exon 14	chr6:33442354-33442514:+	33442453-33442494	0.0016	0.0000	3.0971	762
TAC3 exon 5	chr12:57012802-57012962:-	57012889-57012942	0.9791	-8.9206	0.9203	641
TIMM44 exon 11	chr19:7928057-7928217:-	7928108-7928197	0.0547	-2.6541	4.7784	784
TNFRSF1A exon 3	chr12:6333825-6333985:-	6333945-6333965	0.0000	0.0000	0.0000	636
TNNT2 exon 4	chr1:201372007-201372167:-	201372133-201372147	0.8881	-6.6186	2.8294	512
TNNT2 exon 5	chr1:201369796-201369956:-	201369907-201369936	0.1645	-2.9302	4.5015	842
TSC2 exon 32	chr16:2082364-2082524:+	2082436-2082504	0.0059	-0.4247	2.9869	874
TTC8 exon 7	chr14:88843710-88843870:+	88843806-88843850	0.7465	-5.6842	3.3308	388
TTN-AS1 exon 4	chr2:178537307-178537467:+	178537361-178537447	0.0000	0.0000	0.0000	441
UROS exon 3	chr10:125816157-125816317:-	125816214-125816297	0.9598	-8.7013	1.1258	421
USB1 exon 4	chr16:58014186-58014346:+	58014273-58014326	0.9295	-7.8930	1.8662	351
USP28 exon 6	chr11:113834229-113834389:-	113834283-113834369	0.3645	-4.1615	4.2147	449
VPS33B exon 2	chr15:91017785-91017945:-	91017845-91017925	0.9798	-8.5880	0.8321	532
WDR35 exon 11	chr2:19962275-19962435:-	19962383-19962415	0.0025	0.0000	4.1151	332

Supplementary Table 1 | Summary of reference PSI and Δlogit(PSI) ranges for near-saturation exons

This table reports all exons with at least 300 assayed variants. For each exon we provide the gene and exon number, event identifier with the genomic coordinates (hg38) of the variable sequence, coordinates of the exon start and stop, the mean reference PSI calculated across all cell lines, and the minimum and maximum average Δlogit(PSI) across all cell lines. The minimum and maximum are determined across all variants for that exon. We also report the total number of variants observed for that exon. These data summarize the splicing landscape of highly covered exons and serve as a reference for interpreting mutational effects in this study.

Name	Sequence (5'-3')	Description
AMY_8	TAGAAGGCACAGTCGAACCGCTGTGTTCTGC	Custom R2 Primer RNA and DNA-Seq
AMY_9	GCAGAACACAGCGGTTCGACTGTGCCTTCTA	Index primer (IX7) RNA and DNA-seq
AMY_12	CAAGCAGAAGACGGCATACGAGATCGATNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HEK Rep 1
AMY_10	CAAGCAGAAGACGGCATACGAGATacgtNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HEK Rep 2, Round 1 DNA library PCR Reverse Primer
AMY_11	CAAGCAGAAGACGGCATACGAGATtacgNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HeLa Rep 1
skp_44_ESL_umi_RT_truseq701_rev	CAAGCAGAAGACGGCATACGAGATtcgcttaNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HeLa Rep 2
AMY_13	CAAGCAGAAGACGGCATACGAGATGTTANNNNNNNNTAGAAGGCACAGTCGAACC	RT primer K562 Rep 1
skp_47_ESL_umi_RT_truseq704_rev	CAAGCAGAAGACGGCATACGAGATGCTCAGGANNNNNNNNTAGAAGGCACAGTCGAACC	RT primer K562 Rep 2
skp_95_ESL_umi_RT_truseq711_rev	CAAGCAGAAGACGGCATACGAGATTGCCTCTTNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer MCF7 Rep 1
skp_96_ESL_umi_RT_truseq712_rev	CAAGCAGAAGACGGCATACGAGATTCCTCTACNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer MCF7 Rep 2
skp_97_ESL_umi_RT_truseq714_rev	CAAGCAGAAGACGGCATACGAGATTCATGAGCNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HMC3 Rep 1
skp_98_ESL_umi_RT_truseq715_rev	CAAGCAGAAGACGGCATACGAGATCCTGAGATNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer HMC3 Rep 2
AMY_2	TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNNNNNBNKBCAGAAGAACGGCATCAAAGTGAA	Round 1 RNA library PCR Forward Primer Mix - Primer 1
AMY_3	TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNAGAAGAACGGCATCAAAGTGAA	Round 1 RNA library PCR Forward Primer Mix - Primer 2
AMY_4	TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNAGAAGAACGGCATCAAAGTGAA	Round 1 RNA library PCR Forward Primer Mix - Primer 3
AMY_5	TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNBNNBNNHNNAGAAGAACGGCATCAAAGTGAA	Round 1 RNA library PCR Forward Primer Mix - Primer 4
AMY_7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Round 2 RNA library PCR Forward Primer
P7	CAAGCAGAAGACGGCATACGAGAT	Round 1/2/3 RNA library PCR/qPCR Reverse Primer - P7
P5	AATGATACGGCGACCACCGAGATCTACAC	Round 3 RNA library qPCR Forward Primer - P5
skp5_DMD_RT_primer_unique_index	CAAGCAGAAGACGGCATACGAGATgatgNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer DMD Control reporter
skp6_MAPT_RT_Unique_Index_primer	CAAGCAGAAGACGGCATACGAGATccagNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer MAPT Control reporter
skp7_CFTR_RT_Unique_Index_primer	CAAGCAGAAGACGGCATACGAGATcttcNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer CFTR Control reporter
skp8_SMN2_RT_Unique_Index_primer	CAAGCAGAAGACGGCATACGAGATgtctNNNNNNNNTAGAAGGCACAGTCGAACC	RT primer SMN2 Control reporter
AMY_26	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGTGAAACAAAATGCTTTTTAACATCCA	Round 1 DNA library PCR Forward Primer Mix - Primer 1
AMY_36	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNVMVCYGTGTGAAA CAAAATGCTTTTTAACAT	Round 1 DNA library PCR Forward Primer Mix - Primer 2
AMY_37	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNVMVTGTGAAACAA AATGCTTTTTAACAT	Round 1 DNA library PCR Forward Primer Mix - Primer 3
AMY_38	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNNTGTGAAACAAAAT GCTTTTTAACAT	Round 1 DNA library PCR Forward Primer Mix - Primer 4
Guide oligo part 1 top oligo V4	CACCGTTTATGTTGCAGCCAGCAGGTTT	Prime editing
Guide oligo part 1 bottom oligo V4	CTCTAAACCTGCTGGCTGCAACATAAAC	Prime editing
Guide oligo part 1 top oligo V5	CACCGCTTTGTCCCTGCTGTCTTAGTTT	Prime editing
Guide oligo part 1 bottom oligo V5	CTCTAAACTAAGACAGCAGGGACAAAGC	Prime editing
Guide oligo part 2 top oligo	/5Phos/AGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCG	Prime editing
Guide oligo part 2 bottom oligo	/5Phos/GCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAG	Prime editing
Guide oligo part 3 top oligo V4	GTGCTTGGGTCCCACCgGCCACCTCCGAGGCCTCTGCTGGCTGCAACATAA	Prime editing
Guide oligo part 3 bottom oligo V4	CGCGTTATGTTGCAGCCAGCAGAGGCCTCGGAGGTGGCcGGTGGGACCCAA	Prime editing
Guide oligo part 3 top oligo V5	GTGCGCAAGTGAAGCAGCCTCCaTAAGACAGCAGGG	Prime editing
Guide oligo part 3 bottom oligo V5	CGCGCCCTGCTGTCTTAtGGAGGCTGCTTCACTTGC	Prime editing
BIN1_DNA_Primer_Fwd	GGGCTTTGTCCCTGCTGTC	Prime editing
BIN1_DNA_Primer_Rev	GCTGTGTCCTGTCCTGTGTCTG	Prime editing
BIN1_RNA_Primer_Fwd_1	CACAGTTGCTGGGAAGGTCTC	Prime editing
BIN1_RNA_Primer_Rev_1	GAGATCAGAGTCAACCACGAGC	Prime editing
BIN1_RNA_Primer_Fwd_2	CACCACGACAGCAGGAAGA	Prime editing
BIN1_RNA_Primer_Rev_2	GAGTCAACCACGAGCCAGAG	Prime editing

Supplementary Table 2 | Primers used for library preparation, sequencing, and prime editing experiments

List of DNA and RNA sequencing primers as well as primers used for prime editing assays in Chapters 2 and 3.

Sequence Name	Sequence 5'-3'	Description
Control reporter SMN2 exon 7	TATAGCTATTTTTTTTAACTTCCTTTATTTTCCTTACAGGGTTTTAGACAAAATCAAA AAGAAGGAAGGTGCTCACATTCCTTAAATTAAGGAGTAAGTCTGCCAGCATTATG	SMN2 exon 7 + 5' and 3' intronic flanks
Control reporter MAPT exon 10	AGGCGGGTCCAGGGTGGCGTGTCACTCATCCTTTTTTCTGGCTACCAAAGGTGCAGAT AATTAATAAGAAGCTGGATCTTAGCAACGTCCAGTCCAAGTGTGGCTCAAAGGATAAT ATCAAACACGTCCCGGGAGGCGGCAGTGTGAGTACCTTCACACGTCC	MAPT exon 10 + 5' and 3' intronic flanks
Control reporter DMD exon 29	TGTATTTAGAAAAAAAAGGAGAAATAGTAATTATTGCAAATGTGTTTCAGTCACTTGA AAATTTGATGCGACATTCAGAGGATAACCCAAATCAGATTCGCATATTGGCACAGACC CTAACAGATGGCGGAGTCATGGATGAGCTAATCAATGAGGAACTTGAGACATTTAATT CTCGTTGGAGGGAACTACATGAAGAGGTATGAAGATAAGTGAAAAA	DMD exon 29 + 5' and 3' intronic flanks
Control reporter CFTR exon 12	TGACCAGGAAATAGAGAGGAAATGTAATTTAATTTCCATTTTCTTTTTAGAGCAGTAT ACAAAGATGCTGATTTGTATTTATTAGACTCTCCTTTTGGATACCTAGATGTTTTAAC AGAAAAAGAAATATTTGAAAGGTATGTTCTTTGAATACCTT	CFTR exon 12 + 5' and 3' intronic flanks
COLQ exon 5 variant chr3:15479074: G>C	CAGCCAAGCAACTGAGCCACATGGCCATGCCCAGTCATCAGCTGTACTAAGGCTTCTG CTTCCAGAACATCTCTTCCAAAGCAGCCTCTCCTTACCTGTCTTCTCCATCCATAGGG GGAGCTTGGCCGACCAGGAAGGAAGGTATGGTCTGCGCTGTTCTG	COLQ exon 5 variant used in low throughput RBP motif experiements
BIN1 exon 16 target variant 1 chr2: 127051220:C>G (hg38)	CTGGCTGTGTCCTGTCCTGTGTCTGACCCCAAGCCGGCATTTATGTTGCAGCCAGCAG AGGCCTCGGAGGTGGCCGGTGGGACCCAACCTGCGGCTGGAGCCCAGGAGCCAGGGGA GACGGCGGCAAGTGAAGCAGCCTCCGTAAGACAGCAGGGACAAAG	BIN1 exon 16 variant 1 prime editing
BIN1 exon 16 target variant 2 chr2: 127051153:C>T (hg38)	CTGGCTGTGTCCTGTCCTGTGTCTGACCCCAAGCCGGCATTTATGTTGCAGCCAGCAG AGGCCTCGGAGGTGGCGGGTGGGACCCAACCTGCGGCTGGAGCCCAGGAGCCAGGGGA GACGGCGGCAAGTGAAGCAGCCTCCATAAGACAGCAGGGACAAAG	BIN1 exon 16 variant 2 prime editing

Supplementary Table 3 | Sequences used for low-throughput splicing experiments
List of sequences used in prime editing, COLQ variant assays, and control reporter experiments.

Primer	Sequence (5'-3')
AMY_7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCT
P5_forward	AATGATACGGCGACCACCGAGATCTACAC
P7_reverse	CAAGCAGAAGACGGCATACGAGAT
skp_22_PolyT_RT_rev_UMI_PE1_index2_P7	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTTT
skp_23_PolyT_RT_rev_UMI_PE1_index1_redo_P7	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_37_PolyT_RT_rev_UMI_PE1_index8_P7	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_24_PolyT_RT_rev_UMI_PE1_index3_P7	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_25_PolyT_RT_rev_UMI_PE1_index4_P7	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_35_PolyT_RT_rev_UMI_PE1_index6_P7	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_34_PolyT_RT_rev_UMI_PE1_index5_P7	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT
skp_36_PolyT_RT_rev_UMI_PE1_index7_P7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCNNNNNNNNVTTTTTTTTTTTTTTTTT

Supplementary Table 4 | Primers used for APA MPRA experiments
List of DNA and RNA sequencing primers, as well as primers used for all APA MPRA library preparations described in Chapter 4.

Methods

AS MPRA methods (Chapters 2 and 3)

Experimental Overview

The exon skipping plasmid library was amplified for DNA-seq in order to construct a map of SNVs and their associated unique barcode sequences. Following transfection of this library in HEK293 cells, RNA-seq was performed to quantify the PSI associated with each unique barcoded reporter.

DNA-seq Library Preparation

Because each variant contains the same backbone region, there is low nucleotide diversity in each sequence. To minimize the amount of PhiX needed for Illumina sequencing, "offset" nucleotides between the Illumina sequencing adapter region and a plasmid specific primer were used in order to create diversity of bases in generated amplicons for increased percent of clusters passed filtering. To prepare the plasmid library for DNA-sequencing, 1ng of plasmid was initially amplified for 5 cycles with primers that bind the plasmid, including those “offset” sequences to generate library diversity. Following this PCR, qPCR with P5/P7 Illumina primers was used to selectively amplify the products containing sequencing adaptors and to prevent over-amplification. Amplification was conducted and monitored with qPCR and stopped early to minimize PCR biases. Samples were size-selected to keep fragments ≥ 100nt using KAPA Pure Beads (Roche). Sample size distributions were analyzed using Tapestation High Sensitivity D1000 (Agilent).

DNA-seq

Library concentrations were quantified using qPCR quantification with the NEBNext Library Quant Kit for Illumina (NEB) as well as Qubit 1x dsDNA HS (Thermo Fisher). Concentrations between the two methods were averaged to determine optimal library loading concentrations for sequencing. Sequencing of DNA libraries, with 2% spiked-in PhiX, was iteratively performed on 300 cycle MiSeq Reagent V2 kits (Illumina) until the desired sequencing coverage and depth was obtained. Custom primers were used with 251 cycles on read 1, 45 cycles on read 2, and 12 cycles on Index 1. Reads from multiple DNA-seq interactions were concatenated and processed together to build a comprehensive map of DNA barcodes.

Cell Culture and Transfection

HEK-293 cells (ATCC, CRL-1573), MCF7 cells (ATCC, HTB22), and HeLa cells (ATCC, CCL-2.2) were cultured in DMEM (Gibco). K-562 cells (ATCC, CCL-243) were cultured in RPMI (Gibco). HCM3 cells (ATCC, CRL-3304) were cultured in EMEM (ATCC). All media was supplemented with 10% fetal bovine serum (Cytiva) and 1% penicillin/streptomycin (Gibco), and cells were cultured at 37°C and 5% CO2. For transfection, cells were seeded to a density of 200,000 cells/mL in a 10 cm plate, and transfected with 15 μg of plasmid library using Lipofectamine 3000 (Thermo Fisher) with at least 2 biological replicates for each library. Cells were harvested for RNA extraction 36–48 h after transfection. Less than 5% of the cells were used for flow cytometry to confirm transfection efficiency ,and the remaining cells were used for RNA extraction.

RNA-seq Library Preparation

RNA extraction was performed with the Monarch® Total RNA Miniprep Kit (NEB). mRNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) with 5μg of total RNA input per sample. The resulting mRNAs were reverse transcribed with SuperScript IV Reverse Transcriptase (Thermo Fisher) to generate cDNA using an RT primer specific to the 3′ end of the transcripts (downstream of the barcode sequence). Library cDNA was then amplified with KAPA HiFi HotStart ReadyMix (Roche) for 5 cycles with a forward primer that binds upstream of the Citrine exon 1 splice junction and a reverse primer that binds downstream of the barcode region. As with the DNA library preparation, additional "offset" sequences to generate library diversity were added to the 5′ end of the amplicon. A second 5-cycle PCR step was done using a forward primer that binds PE1 and P5 and a reverse primer that binds P7. Lastly, qPCR with P5/P7 Illumina primers were used to selectively amplify the products containing sequencing adaptors and to prevent over-amplification. Amplification was conducted and monitored with qPCR and stopped early to minimize PCR biases. Samples were size-selected to keep fragments ≥ 100nt using KAPA Pure Beads (Roche). Sample size distributions were analyzed using Tapestation High Sensitivity D1000 (Agilent).

RNA-seq

Library concentrations were quantified using qPCR quantification with the NEBNext Library Quant Kit for Illumina (NEB) as well as Qubit 1x dsDNA HS (Thermo Fisher). Concentrations between the two methods were averaged to determine optimal library loading concentrations for sequencing. Libraries were sequenced on a NextSeq 500/550 Mid Output (300 cycle) kit using custom primers, with 212 cycles on read 1 to span the library sequence into both splice junctions,

86 cycles on read 2 for barcode sequences, and 12 cycles on Index 1 for UMI and indices. Libraries were pooled with 10% PhiX prior to sequencing.

Bioinformatic Workflow

DNA-seq reads were used to extract 20-nt barcodes from R2, anchored by a required 20-bp forward shared region (allowing up to 5 mismatches). Barcodes were clustered with starcode (distance = 1), and consensus barcode+insert sequences supported by ≥5 reads were assembled into a FASTA reference for STAR indexing. Reference FASTA entries were assembled from each unique 20-nt barcode cluster. For each barcode cluster and library member combination, the reference sequence consisted of the last 25 nt of Citrine exon 1, SMN2 intron 6, the variable exon and intronic flanks, SMN2 intron 7, Citrine exon 2, and the 20-nt barcode. The matched barcode+insert references were used to build STAR genome indexes for RNA-seq mapping. Resulting RNA-seq reads are passed through a bioinformatic pipeline to count exon skipping events. Here, we take advantage of STAR's capability to detect and count expected and novel exon-exon spanning reads to determine splice isoform abundance. Accurately counting PSI for single reporter constructs will be pivotal for the future development of ML models to predict isoform abundance based on sequence. Thus, a mini MPRA library spanning 13 sequences sampled from the full DNA library was used to optimize a computational pipeline for producing accurate PSI calculations (**Figure 2.3B**). The optimized pipeline 1) clusters DNA barcodes using starcode, 2) collapses unique molecular identifiers (UMIs) to remove duplicate RNA reads caused by PCR amplification, and 3) uses STAR for alignment, mapping, and subsequent PSI calculation (17-19). Once this pipeline was optimized, the mini MPRA library resulted in PSI

values that correlated well with PSIs for the same sequences found in the full MPRA (**Figure 2.3B**). Following bioinformatic processing, PSI values were computed for each library variant using reads supporting inclusion or exclusion at the expected splice junctions. ΔPSI values were calculated relative to the corresponding wild-type reference sequence. For global analysis of splicing effects across all cell types, reads from two replicates per cell line were first pooled. PSI and ΔPSI values were calculated from the pooled reads for each of the five cell lines, and the resulting PSIs and ΔPSIs were averaged across cell types.

Validation of MPRA plasmid backbone performance

To confirm that the MPRA plasmid backbone did not introduce artificial splicing effects independent of the inserted variable sequence, we performed control experiments with reference exons that have well-characterized exon skipping behavior. Four exons were selected based on their association with human disease and extensive prior study: *SMN2* exon 7 (54 nt), *MAPT* exon 10 (93 nt), *CFTR* exon 12 (87 nt), and *DMD* exon 29 (150 nt). [4,13,249,250] Each exon was cloned with intronic flanks into the same plasmid backbone used for the construction of the MPRA library.

Splicing outcomes were assayed by transfecting plasmids into cells and measuring exon inclusion by RT-PCR with primers in Citrine flanking the splice junctions. Isoform ratios were quantified from gel band intensities in ImageJ, which correlated with published values ($r = 0.83$). PCR products were also sequenced by RNA-seq to provide a higher-resolution benchmark, which yielded similar results ($r = 0.85$ with literature values). PSI values determined by gel and RNA-seq were highly correlated with each other ($r = 1.0$) (**Figure 2.3A**). These results validate

that the plasmid backbone itself does not bias exon skipping and is suitable for high-throughput MPRA assays.

We further benchmarked our MPRA against previously published AS MPRAs. Beyond the ParSE-seq comparison (Figure 3G), we evaluated overlap with the MFASS dataset, another large-scale minigene assay.[40] PSI values were compared for more than 350 overlapping variants between our MPRA and the MFASS dataset, revealing a strong correlation (r = 0.74) (**Figure 2.3 C**). Importantly, these results were obtained despite the use of different plasmid backbones in the two designs (*SMN1* intronic sequence for MFASS versus *SMN2* intronic sequence for our MPRA), demonstrating that splicing MPRAs are generalizable and robust across sequence contexts.

Logit and ΔLogit PSI Transformations

PSI values of 0 and 1, which are undefined on the logit scale, PSI values were clipped to the interval [0.01,0.99] prior to logit transformation. This clipping threshold is consistent with that employed by FRASER, a splicing-focused method that similarly addresses numerical instability on the logit scale at the PSI extreme.[84] We selected 0.01 and 0.99 as empirically justified clipping thresholds to match the resolution imposed by the average read depth of our library, with mean per-sequence coverage across pooled replicates ranging from 112 (MCF7) to 242 (HEK293). This approach also ensures compatibility on the logit scale between variants and wild-type references, which differ in average read depth. To maximize sequencing depth per measurement, we pooled replicates to compute a single Δlogit(PSI) value per variant sequence. We confirmed that this pooling strategy yields results highly consistent with averaging

Δlogit(PSI) values calculated from individual replicates, with r^2 values ranging from 0.96 to 1.0 across comparisons (**Figure S1B**).

Model predictions with SpliceAI

SpliceAI predictions were generated using the original ensemble of five trained models.[74] For prediction, we used the 161nt sequence in our reporter construct (Citrine exon 1, *SMN2* intron 6, variable exon, *SMN2* intron 7, Citrine exon 2), resulting in a 1,563-nt sequence. To satisfy SpliceAI input requirements, sequences were further padded with 10,000 Ns (5,000 upstream and 5,000 downstream). Predictions were run on both single and double variants.

Raw SpliceAI outputs provided acceptor and donor probabilities for all sequence positions. For each sequence, the predicted PSI was calculated as the product of the donor probability at the expected SD and the acceptor probability at the expected SA on the variable exon.[251] These PSI values were then clipped to the interval [0.01, 0.99] and transformed to logit space. Δlogit PSI was then calculated as the difference between variant and reference logits for direct comparison with experimental measurements and predictions from other models. Sequences with reference PSI values of exactly 0 or 1 were excluded when assessing model performance, as such cases are unlikely to show measurable variant effects but can still be scored by models.

Model predictions with Pangolin

Pangolin predictions were obtained using the probability output model, which estimates the likelihood that a site is a splice site. Sequences were prepared identically to those used for SpliceAI: each 161 nt variable sequence in the reporter context (Citrine exon 1, *SMN2* intron 6, the variable exon, *SMN2* intron 7, Citrine exon 2) and padded with 5,000 nucleotides of Ns on

both sides. Predictions were run with each of the four tissue-specific models (heart, liver, brain, and testis). For every sequence, probabilities were recorded at the expected SA and SD for each tissue. Probabilities were clipped to the range [0.01, 0.99] and transformed to logit space. For each tissue, we calculated ΔlogitSD and ΔlogitSA as the difference between alternate and reference logits at the SD and SA, respectively. To obtain a single PSI value and summarize across tissues, we used the absolute maximum of ΔlogitSA and the absolute maximum of ΔlogitSD values observed. We then averaged the ΔlogitSA and ΔlogitSD maxima to use as a representative Δlogit(PSI) value. PSI values were defined following the Pangolin methods as the maximum difference in splice site probabilities across tissues, using the strongest SD and SA values (which could come from different tissues) and averaging these maxima to yield a single PSI estimate per sequence. In our approach, this maximum-difference criterion was applied in logit space rather than probability space.[74]As with SpliceAI, sequences with reference PSI values of exactly 0 or 1 were excluded when assessing model performance.

Model predictions with MMSplice

Variant effects were predicted using MMSplice with the VCFDataloader.[32] To enable this, we generated custom FASTA, GTF, and VCF files for our synthetic sequences. For predictions, we used the 161nt sequence in our reporter construct (Citrine exon 1, *SMN2* intron 6, variable exon, *SMN2* intron 7, Citrine exon 2), resulting in a 1,563 nt sequence. The FASTA was composed of these full-length sequences (1563 nt) for the references only. The GTF file defined the reference transcript length as 1,563 nt and annotated all exon boundaries, including those of the variable exon and both Citrine exons in the reporter. The VCF contained all variant positions and base substitutions. For double mutants, the VCF sequence was defined over the interval

spanning both substitutions, such that the reference and alternate alleles encompassed the entire region. Predictions were limited to variants within 100 nt of the variable exon due to the default overhangs implemented in the MMSplice model; thus, intronic variants located further than 100 nt from the exon were not scored. As done with all other model predictions, sequences with reference PSI values of exactly 0 or 1 were excluded when assessing model performance.

Model predictions with HAL

HAL predictions were obtained using the HAL web server (http://splicing.cs.washington.edu/SE).[61] Reference PSI values were averaged across replicates within each cell line, then provided to HAL on the percent scale after clipping to 0.01-99.99 as required by the tool. HAL predicts ΔPSI for substitutions within the exon or the first 6 nucleotides of the downstream intron; single- or double-mutant sequences with variants outside this window were not used for predictions. For comparison across models and with experimental data, predicted reference and variant PSIs were converted from percent to fraction, clipped to [0.01, 0.99], transformed to logit space, and used to compute Δlogit PSI. As done with all other model predictions, sequences with reference PSI values of exactly 0 or 1 were excluded when assessing model performance.

Prime Editing

Prime editing experiments were performed following the workflow described in Doman et al[252] PEmax (cas9-RT fusion protein, Addgene ID: 174820), pU6-tevopreq1-GG-acceptor (cloning vector, Addgene ID: 174038), and hMLH1dn (mismatch repair suppressor, Addgene ID: 178114) were ordered from Addgene. Guide pegRNAs were designed using Predict and cloned

into the pU5 backbone using Golden Gate assembly[253]. After being sequence verified, individual colonies of all plasmids were grown in LB with carbenicillin, and plasmids were extracted using Qiagen Midi prep kits. HEK293 cells were seeded in 24 well-plates with a density of 100,000 cells/well in HyClone DMEM with High Glucose at 36 C and 5% CO2. Cells were transfected with 1.94 ug of DNA in ratio 3:1 Prime editing vector: epegRNA vector using Lipofectamine 3000 (Thermo Fisher) with at least two replicates of each condition. Cells were harvested for genome extraction 48 hours after transfection or reseeded for subsequent transfections. Up to 5 subsequent transfections were performed for maximum editing efficiency. Cells were detached using TrypLE (Gibco) and resuspended in PBS with Proteinase K Between 1 and 5% of cells were used for flow cytometry to confirm transfection. gDNA and total RNA were extracted from the remaining cells using the NEB Monarch Total RNA Kit, and mRNA extraction was performed using the NEBNext Poly(A) mRNA Magnetic Isolation Module. The resulting mRNAs were reverse transcribed with SuperScript IV Reverse Transcriptase (Thermo Fisher) to generate cDNA using an anchored polyT primer. Target PCR using KAPA HiFi HotStart ReadyMix (Roche) is used to amplify the region surrounding the edit (~300 bp). Genomic PCR amplification was performed with a single set of primers, while cDNA amplicons were generated using nested PCRs with a total of 25 cycles for both conditions. Genomic and cDNA fragments were sequenced using Plasmidsaurus Premium PCR (performed with Oxford Nanopore MinION R10.4.1). From the genomic DNA samples, reads were filtered for those containing exact matches for both the expected primer sequence as well as the gRNA sequence. Editing efficiency was calculated as the ratio of reads containing the expected edit at the expected location over the total number of reads correctly assigned to the reference sequence. For RNA read alignment, reference sequences were generated for the exon-included and exon-skipped isoforms. UCSC

Genome Browser sequences for BIN1 exon 16 and its surrounding exons were used to define these references. Each read was compared against these two reference sequences with Bio.pairwise2 (up to 3 mismatches allowed) and assigned to the matching isoform. Reads were classified as included or excluded and PSI was calculated of the ratio of included reads over the total number of reads.

Motif Effect Size Analysis

We assembled a library of RBP motifs by aggregating position weight matrices (PWMs) from the ATtRACT, oRNAment, CISBP-RNA databases[4-6], retaining only those annotated for *Homo sapiens*. Redundant PWMs across databases were merged, and the resulting set was clustered with the RSAT matrix-clustering tool[75], which we modified to disable reverse-complement comparisons. Clustering was performed using normalized correlation (Ncor) as the similarity metric, complete linkage, a correlation threshold of 0.50, Ncor threshold of 0.40, width threshold of 3, and reverse-complement comparisons disabled. This procedure produced a non-redundant set of motif clusters representing distinct RBP binding preferences.

All sequences from our MPRA library (including both reference and variant constructs) were scanned against the unclustered PWM set using FIMO from the MEME suite v5.5.0, with reverse complement scanning disabled. Only significant motif matches were retained based on FIMO q-value[110]. Because the MPRA library includes systematic single- and double-nucleotide mutations of each reference sequence, we grouped each reference and its associated variants into an "exon family." Within each family, we identified which sequences contained at least one significant hit to a motif in each cluster and which did not.

We computed Δlogit(PSI) values for each exon family in each cell line by comparing the logit-transformed PSI values of the hit-containing vs. non-hit sequences. PSI values were calculated from pooled replicate read counts and clipped to the interval [0.01, 0.99], and transformed with the logit function:

$$\mathrm{logit}(\mathrm{PSI}) = \ln\left(\frac{\mathrm{PSI}}{1-\mathrm{PSI}}\right)$$

To avoid artifacts from splice site disruption, variants overlapping the core splice donor or acceptor dinucleotides (defined as ±4 nt from either junction) were excluded from the analysis. Effect size analysis was performed separately for each cell line. RBP motif effect sizes were calculated for exon families that contained at least 10 variants in both groups (≥10 hits and ≥10 non-hits). For each exon family in each cell line, the effect size was defined as the difference between the average logit(PSI) values of hit-containing and non-hit sequences:

$$\mathrm{Effect\ Size}_{\mathrm{motif}} = \Delta\mathrm{logit}(\mathrm{PSI} \mid \mathrm{motif}) = \frac{1}{N_+}\sum_{i\in+}\mathrm{logit}(\mathrm{PSI}_i) - \frac{1}{N_-}\sum_{j\in-}\mathrm{logit}(\mathrm{PSI}_j)$$

where N_+ and N_- denote the number of hit-containing and non-hit sequences, respectively. Calculations were considered further only if the criterion required to calculate effect sizes was satisfied in at least three exon families. For each RBP cluster, we then calculated a concordance score to quantify the consistency of motif effects across exon families. For each RBP cluster, concordance across exon families was quantified as the weighted sign of effect sizes:

$$\mathrm{Concordance} = \frac{\sum_k \Delta\mathrm{logit}(\mathrm{PSI})_k}{\sum_k |\Delta\mathrm{logit}(\mathrm{PSI})_k|}$$

This metric ranges from -1 (consistently negative effects across exon families) to +1 (consistently positive effects across exon families), with values near 0 indicating inconsistent

directionality across exon families. For downstream interpretation, we retained RBP clusters with absolute values of concordance scores ≥ 0.75 in at least three exon families and observed in at least two cell lines. For these clusters, effect sizes were averaged across cell lines to estimate the overall regulatory impact of each RBP motif cluster. We then averaged the effect sizes across those cell lines to estimate each cluster's overall regulatory impact.

Mingap analysis

To quantify cell type-specific splicing, we defined the ΔPSI mingap metric as the separation between the most extreme PSI value for a sequence and its next closest value across cell lines. Specifically, ΔPSI mingap high was calculated as the absolute difference between the highest PSI and the next highest PSI, reflecting cell type-specific exon inclusion. Conversely, ΔPSI mingap low was calculated as the absolute difference between the lowest PSI and the next lowest PSI, reflecting cell type-specific exon skipping. The larger of the two was taken as the overall ΔPSI mingap value for a sequence with measurements present in all 5 cell lines. Variants with ΔPSI mingap ≥ 0.25 were classified as cell type-specific.

APA MPRA methods (Chapter 4)

MPRA Design (small APA MPRA)

A total of 76 predicted outlier variants from the filtered autism data were chosen for experimental validation (38 case variants and 38 matched controls). Additionally, we tested 9 GWAS SNPs, 3 hand-picked examples from the F2, SCAF8, and MECP2 genes, 2 apaQTLs, 4 matched control SNPs from gnomAD, and 6 control PASs that had previously been measured in the MPRA of Bogard et al.[226]

MPRA Design (large APA MPRA)

More than half of the sequences were derived from GWAS studies, allowing us to examine APA variants linked to a broader spectrum of diseases. In addition, the libraries incorporated an expanded set of ASD whole-genome sequences from our previous MPRA, as well as GWAS mask variants generated with the Scrambler architecture[242]. In this framework, masks stochastically perturb the input sequence while preferentially retaining the most important nucleotides for APARENT2's predictions, thereby revealing regulatory elements that drive polyadenylation isoform choice. Mask-derived variants were included in our MPRA to extend discovery beyond naturally occurring variants. We have also included epidermal-specific polyadenylation site sources from previous studies in human keratinocytes.[243] We also included 126 control PASs that had previously been measured in the MPRA of Bogard et al.[226]

Cloning

For the variant and reference libraries for all APA MPRAs, a vector was constructed by cloning two separate 250-nt oligo pool libraries (Twist) 25nt upstream of the bGH polyadenylation signal in a mCherry reporter plasmid. All libraries contained homology to the vector and were constructed using In-Fusion assembly (Takara). Library sizes were estimated by plating serial dilutions of library transformation and extrapolating CFUs based on colony counts. The remaining transformants were grown in 100-mL LB overnight culture, and libraries were prepared using a HiSpeed Plasmid Midi Kit (Qiagen). Selected individual clones from each library were Sanger sequenced to confirm library assembly. Cloning was performed at a scale such that the number of colonies was at least 100x greater than the number of perturbations in the library.

Cell culture and transfection

HEK 293T cells (ATCC, CRL-3216) were cultured in Dulbecco's modified Eagle medium (Gibco), SK-N-SH cells (ATCC HTB-11) were cultured in MEM (Gibco), HCM3 cells (ATCC, CRL-3304) were cultured in EMEM (ATCC), and K562 cells (ATCC CCL-243) were cultured in RPMI (Gibco), all supplemented with 10% fetal bovine serum (Cytiva) and 1% penicillin/streptomycin (Gibco). For small library transfections, 300,000 (HEK293T) or 500,000 (SK-N-SH, HMC3) cells were plated (2mL) in tissue culture-treated 6-well plates (Fisher Scientific) 24h prior such that they would be 50-70% confluent on the day of transfection. Variant and reference libraries were transfected into cells using Lipofectamine 3000 (Thermo Fisher) with 2 biological replicates for each library. Media were changed 5h post-transfection.

For the large library transfections in HEK293, transfections were scaled to accommodate 2.2 million cells in a 10cm plate. For K562 cells, Neon nucleofection (Thermo) was performed with 1 million total cells.

RNA extraction

Cells were harvested for RNA extraction 36-48 h after transfection. Cells were detached by incubatingfor 2-5 min at room temperature with 1mL of TrypLE (Gibco). Once cells were detached, they were added to 3mL of media with 10% FBS. Cells were rinsed twice with 1x PBS (Gibco). One to 5% of the cells were used for flow cytometry to confirm transfection efficiency, and the remaining cells were lysed and passed through a QIAshredder (Qiagen) to homogenize cell lysates. Total RNA was extracted from lysates using the RNeasy kit (Qiagen) with additional on-column DNaseI digestion performed following the manufacturer's protocol. mRNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) with 5µg of total RNA input per sample.

Sequencing library construction

The resulting mRNAs were reverse transcribed with SuperScript IV Reverse Transcriptase (Thermo Fisher) to generate cDNA. Polyadenylated mRNA was reverse transcribed with an anchored polyT primer containing Illumina adaptor sequences, a unique molecular identifier (UMI), and unique index sequences for each sample (P7-index-PE2-UMI-T18VN). RNA hydrolysis was performed on cDNA, and samples were purified using DNA Clean & Concentrator-5 (Zymo). Library cDNA was then amplified with KAPA HiFi HotStart ReadyMix (Roche) using a library-specific forward primer containing

additional Illumina adaptor sequences (P5-PE1) and reverse primer matching the adaptor sequence added during RT. Amplification was conducted and monitored with qPCR and stopped early to minimize PCR biases. Samples were size-selected to keep fragments 100 nt using KAPA Pure Beads (Roche). Sample size distributions were analyzed using Tapestation High Sensitivity D1000 (Agilent).

Library sequencing

Library concentrations were quantified using qPCR quantification with the NEBNext Library Quant Kit for Illumina (NEB) as well as Qubit 1x dsDNA HS (Thermo Fisher). Concentrations between the two methods were averaged to determine optimal library loading concentrations for sequencing. Libraries were pooled with 10% PhiX and sequenced on MiSeq (Illumina) with a MiSeq Reagent Nano or Micro v2 (300 cycles) kit. For the small MPRA libraries, paired-end sequencing was performed with read 1 (292 nt) covering the library sequences, read 2 (8 nt) covering the UMIs, and the index read (6 nt) for demultiplexing pooled samples. For large MPRA libraries, paired-end sequencing was performed on NextSeq 500/550 Mid Output kits (300 cycles) for RNA-seq with read 1 (300 cycles) covering the library sequences, and the index 1 read (16 cycles) for the UMIs and index reads for demultiplexing pooled samples.

Barcoding and mapping

The open-source adaptor trimming software package cutadapt v1.15 was used to trim adapters off read 1. Read 1 was then aligned and mapped to the known respective library sequences. The sequence upstream of the proximal PAS CSE was used to map cleaved mRNA

back to the full UTR sequence (allowing for substitution errors). Reads were searched 5′ to 3′ across for the site of polyadenylation, as sequencing reads were long enough to precisely locate cut sites for all proximal isoforms. The site of polyadenylation was identified by searching for a consecutive run of 20 A's (allowing for substitution errors). A read was considered distally cleaved if no polyadenylation site was found in read 1. Mapped reads were collapsed over UMIs in read 2. The final counts were pooled across the 2 replicates.

Calculating Logits

To quantify isoform usage, we estimated the true isoform logit for each pair of APA sites as

$$\ln\left(\frac{c_p + c_{\text{pseudo}}}{c_p + c_d + c_{\text{pseudo}}}\right)$$

Where c_p is the number of proximal isoform reads, c_d is the number of distal isoform reads, and c_{pseudo} is a pseudo-count added to both numerator and denominator to prevent division by zero.

Case 1: Distal absent (c_d=0 and c_p>0)
The logit simplifies to

$$\ln\left(\frac{c_p + c_{\text{pseudo}}}{c_p + c_{\text{pseudo}}}\right) = \ln(1) = 0$$

Case 2: Proximal absent (c_p=0 and c_d>0)
The logit becomes

$$\ln\left(\frac{c_{\text{pseudo}}}{c_d + c_{\text{pseudo}}}\right) < 0$$

Case 3: Both absent (c_p=0 and c_d=0)

In this boundary case, the logit evaluates to

$$\ln\left(\frac{c_{\text{pseudo}}}{c_{\text{pseudo}}}\right) = \ln(1) = 0$$

However, this scenario does not occur in experimental data, as at least one isoform always receives reads.

Calculating Variant Effects with LOR

To quantify the effect of a variant on isoform usage, we calculated the log-odds ratio (LOR) as the difference in logit values between the variant and its corresponding reference sequence. For a variant with proximal and distal counts $(c_p^{\text{var}}, c_d^{\text{var}})$ and a reference with counts $(c_p^{\text{ref}}, c_d^{\text{ref}})$

The logit values are defined as

Variant:

$$\ln\left(\frac{c_p^{\text{var}} + c_{\text{pseudo}}}{c_p^{\text{var}} + c_d^{\text{var}} + c_{\text{pseudo}}}\right)$$

Reference:

$$\ln\left(\frac{c_p^{\text{ref}} + c_{\text{pseudo}}}{c_p^{\text{ref}} + c_d^{\text{ref}} + c_{\text{pseudo}}}\right)$$

The effect of the variant is given by the difference in logit values between the variant and reference,

$$\Delta\text{logit} = \text{logit}_{\text{var}} - \text{logit}_{\text{ref}}$$

which is equivalent to the log-odds ratio (LOR),

$$\text{LOR} = \ln\left(\frac{\frac{c_p^{\text{var}}+c_{\text{pseudo}}}{c_p^{\text{var}}+c_d^{\text{var}}+c_{\text{pseudo}}}}{\frac{c_p^{\text{ref}}+c_{\text{pseudo}}}{c_p^{\text{ref}}+c_d^{\text{ref}}+c_{\text{pseudo}}}}\right)$$

Positive LOR values correspond to variants that increase proximal isoform usage relative to the reference, while negative values correspond to variants that increase distal usage.

Bibliography

1. Suzuki, H., Aoki, Y., Kameyama, T., Saito, T., Masuda, S., Tanihata, J., Nagata, T., Mayeda, A., Takeda, S., and Tsukahara, T. (2016). Endogenous multiple exon skipping and back-splicing at the DMD mutation hotspot. Int. J. Mol. Sci. *17*, 1722.

2. Bougé, A.-L., Murauer, E., Beyne, E., Miro, J., Varilh, J., Taulan, M., Koenig, M., Claustres, M., and Tuffery-Giraud, S. (2017). Targeted RNA-Seq profiling of splicing pattern in the DMD gene: exons are mostly constitutively spliced in human skeletal muscle. Sci. Rep. *7*, 39094. https://doi.org/10.1038/srep39094.

3. Aartsma-Rus, A. (2010). Antisense-mediated modulation of splicing: therapeutic implications for Duchenne muscular dystrophy. RNA Biol. *7*, 453–461.

4. Ruggiu, M., McGovern, V.L., Lotti, F., Saieva, L., Li, D.K., Kariya, S., Monani, U.R., Burghes, A.H.M., and Pellizzoni, L. (2012). A Role for SMN Exon 7 Splicing in the Selective Vulnerability of Motor Neurons in Spinal Muscular Atrophy. Mol. Cell. Biol. *32*, 126–138. https://doi.org/10.1128/MCB.06077-11.

5. Ottesen, E.W., Singh, N.N., Luo, D., Kaas, B., Gillette, B.J., Seo, J., Jorgensen, H.J., and Singh, R.N. (2023). Diverse targets of SMN2-directed splicing-modulating small molecule therapeutics for spinal muscular atrophy. Nucleic Acids Res. *51*, 5948–5980. https://doi.org/10.1093/nar/gkad259.

6. Singh, R.N., and Singh, N.N. (2018). Mechanism of Splicing Regulation of Spinal Muscular Atrophy Genes. Adv. Neurobiol. *20*, 31–61. https://doi.org/10.1007/978-3-319-89689-2_2.

7. Shibayama, A., Cook, E.H., Feng, J., Glanzmann, C., Yan, J., Craddock, N., Jones, I.R., Goldman, D., Heston, L.L., and Sommer, S.S. (2004). MECP2 structural and 3′‐UTR variants in schizophrenia, autism and other psychiatric diseases: A possible association with autism. Am. J. Med. Genet. B Neuropsychiatr. Genet. *128B*, 50–53. https://doi.org/10.1002/ajmg.b.30016.

8. Yuen, R.K., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., Tong, X., Sun, Y., Cao, D., Zhang, T., et al. (2016). Genome-wide characteristics of de novo mutations in autism. Npj Genomic Med. *1*, 16027. https://doi.org/10.1038/npjgenmed.2016.27.

9. Szkop, K.J., Cooke, P.I.C., Humphries, J.A., Kalna, V., Moss, D.S., Schuster, E.F., and Nobeli, I. (2017). Dysregulation of Alternative Poly-adenylation as a Potential Player in Autism Spectrum Disorder. Front. Mol. Neurosci. *10*, 279. https://doi.org/10.3389/fnmol.2017.00279.

10. Parras, A., Anta, H., Santos-Galindo, M., Swarup, V., Elorza, A., Nieto-González, J.L., Picó, S., Hernández, I.H., Díaz-Hernández, J.I., Belloc, E., et al. (2018). Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. Nature *560*,

441–446. https://doi.org/10.1038/s41586-018-0423-5.

11. Rhine, C.L., Neil, C., Wang, J., Maguire, S., Buerer, L., Salomon, M., Meremikwu, I.C., Kim, J., Strande, N.T., and Fairbrother, W.G. (2022). Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. PLOS Genet. *18*, e1009884. https://doi.org/10.1371/journal.pgen.1009884.

12. Love, J.E., Hayden, E.J., and Rohn, T.T. (2015). Alternative splicing in Alzheimer's disease. J. Park. Dis. Alzheimers Dis. *2*, 6.

13. Momeni, P., Pittman, A., Lashley, T., Vandrovcova, J., Malzer, E., Luk, C., Hulette, C., Lees, A., Revesz, T., Hardy, J., et al. (2009). Clinical and pathological features of an Alzheimer's disease patient with the MAPT ΔK280 mutation. Neurobiol. Aging *30*, 388–393. https://doi.org/10.1016/j.neurobiolaging.2007.07.013.

14. Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., and Haroutunian, V. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. Nat. Genet. *50*, 1584–1592.

15. Linder, J., Koplik, S.E., Kundaje, A., and Seelig, G. (2022). Deciphering the impact of genetic variation on human polyadenylation using APARENT2. Genome Biol. *23*, 232. https://doi.org/10.1186/s13059-022-02799-4.

16. Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu. Rev. Biochem. *72*, 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720.

17. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. *102*, 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002.

18. Jiang, W., and Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. Comput. Struct. Biotechnol. J. *19*, 183–195. https://doi.org/10.1016/j.csbj.2020.12.009.

19. Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. *6*, 386–398. https://doi.org/10.1038/nrm1645.

20. Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu. Rev. Biochem. *72*, 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720.

21. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476. https://doi.org/10.1038/nature07509.

22. Ding, F., and Elowitz, M.B. (2019). Constitutive splicing and economies of scale in gene expression. Nat. Struct. Mol. Biol. *26*, 424–432. https://doi.org/10.1038/s41594-019-0226-x.

23. Sterne-Weiler, T., and Sanford, J.R. (2014). Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome Biol. *15*, 201. https://doi.org/10.1186/gb4150.

24. Rogalska, M.E., Vivori, C., and Valcárcel, J. (2023). Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat. Rev. Genet. *24*, 251–269. https://doi.org/10.1038/s41576-022-00556-8.

25. The Geuvadis Consortium, Lappalainen, T., Sammeth, M., Friedländer, M.R., 'T Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511. https://doi.org/10.1038/nature12531.

26. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., and Mei, R. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.

27. Tian, J., Wang, Z., Mei, S., Yang, N., Yang, Y., Ke, J., Zhu, Y., Gong, Y., Zou, D., and Peng, X. (2019). CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. Nucleic Acids Res. *47*, D909–D916.

28. Consortium, Gte. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

29. Capitanchik, C., Wilkins, O.G., Wagner, N., Gagneur, J., and Ule, J. (2025). From computational models of the splicing code to regulatory mechanisms and therapeutic implications. Nat. Rev. Genet. *26*, 171–190. https://doi.org/10.1038/s41576-024-00774-2.

30. Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711. https://doi.org/10.1016/j.cell.2015.09.054.

31. Zeng, T., and Li, Y.I. (2022). Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. *23*, 103. https://doi.org/10.1186/s13059-022-02664-4.

32. Cheng, J., Nguyen, T.Y.D., Cygan, K.J., Çelik, M.H., Fairbrother, W.G., Avsec, Ž., and Gagneur, J. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. *20*, 48. https://doi.org/10.1186/s13059-019-1653-z.

33. Cheng, J., Çelik, M.H., Kundaje, A., and Gagneur, J. (2021). MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. *22*, 94. https://doi.org/10.1186/s13059-021-02273-7.

34. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015.

35. Leung, M.K.K., Xiong, H.Y., Lee, L.J., and Frey, B.J. (2014). Deep learning of the tissue-regulated splicing code. Bioinformatics *30*, i121–i129. https://doi.org/10.1093/bioinformatics/btu277.

36. Liao, S.E., Sudarshan, M., and Regev, O. (2023). Deciphering RNA splicing logic with interpretable machine learning. Proc. Natl. Acad. Sci. *120*, e2221165120. https://doi.org/10.1073/pnas.2221165120.

37. Gupta, K., Yang, C., McCue, K., Bastani, O., Sharp, P.A., Burge, C.B., and Solar-Lezama, A. (2024). Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing. Genome Biol. *25*, 23. https://doi.org/10.1186/s13059-023-03162-x.

38. Chiang, H.-L., Chen, Y.-T., Su, J.-Y., Lin, H.-N., Yu, C.-H.A., Hung, Y.-J., Wang, Y.-L., Huang, Y.-T., and Lin, C.-L. (2022). Mechanism and modeling of human disease-associated near-exon intronic variants that perturb RNA splicing. Nat. Struct. Mol. Biol., 1–13. https://doi.org/10.1038/s41594-022-00844-1.

39. Adamson, S.I., Zhan, L., and Graveley, B.R. (2018). Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol. *19*, 71. https://doi.org/10.1186/s13059-018-1437-x.

40. Cheung, R., Insigne, K.D., Yao, D., Burghard, C.P., Wang, J., Hsiao, Y.-H.E., Jones, E.M., Goodman, D.B., Xiao, X., and Kosuri, S. (2019). A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. Mol. Cell *73*, 183-194.e8. https://doi.org/10.1016/j.molcel.2018.10.037.

41. Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374. https://doi.org/10.1101/gr.119628.110.

42. Culler, S.J., Hoff, K.G., Voelker, R.B., Berglund, J.A., and Smolke, C.D. (2010). Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. Nucleic Acids Res. *38*, 5152–5165. https://doi.org/10.1093/nar/gkq248.

43. Yu, Y., Maroney, P.A., Denker, J.A., Zhang, X.H.-F., Dybkov, O., Lührmann, R., Jankowsky, E., Chasin, L.A., and Nilsen, T.W. (2008). Dynamic Regulation of Alternative Splicing by Silencers that Modulate 5′ Splice Site Competition. Cell *135*, 1224–1236. https://doi.org/10.1016/j.cell.2008.10.046.

44. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic Identification and Analysis of Exonic Splicing Silencers. Cell *119*, 831–845. https://doi.org/10.1016/j.cell.2004.11.010.

45. La Fleur, A., Shi, Y., and Seelig, G. (2024). Decoding biology with massively parallel reporter assays and machine learning. Genes Dev. *38*, 843–865. https://doi.org/10.1101/gad.351800.124.

46. Wong, M.S., Kinney, J.B., and Krainer, A.R. (2018). Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites. Mol. Cell *71*, 1012-1026.e3. https://doi.org/10.1016/j.molcel.2018.07.033.

47. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123. https://doi.org/10.1038/nature13695.

48. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype–phenotype landscape for the alternative splicing of a human exon. Nat. Commun. *7*, 11558. https://doi.org/10.1038/ncomms11558.

49. Ke, S., Anquetil, V., Zamalloa, J.R., Maity, A., Yang, A., Arias, M.A., Kalachikov, S., Russo, J.J., Ju, J., and Chasin, L.A. (2018). Saturation mutagenesis reveals manifold determinants of exon definition. Genome Res. *28*, 11–24. https://doi.org/10.1101/gr.219683.116.

50. Braun, S., Enculescu, M., Setty, S.T., Cortés-López, M., De Almeida, B.P., Sutandy, F.X.R., Schulz, L., Busch, A., Seiler, M., Ebersberger, S., et al. (2018). Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. Nat. Commun. *9*, 3315. https://doi.org/10.1038/s41467-018-05748-7.

51. Cortés-López, M., Schulz, L., Enculescu, M., Paret, C., Spiekermann, B., Quesnel-Vallières, M., Torres-Diz, M., Unic, S., Busch, A., Orekhova, A., et al. (2022). High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. Nat. Commun. *13*, 5570. https://doi.org/10.1038/s41467-022-31818-y.

52. Gergics, P., Smith, C., Bando, H., Jorge, A.A.L., Rockstroh-Lippold, D., Vishnopolska, S.A., Castinetti, F., Maksutova, M., Carvalho, L.R.S., Hoppmann, J., et al. (2021). High-throughput splicing assays identify missense and silent splice-disruptive POU1F1 variants underlying pituitary hormone deficiency. Am. J. Hum. Genet. *108*, 1526–1539. https://doi.org/10.1016/j.ajhg.2021.06.013.

53. Smith, C., Burugula, B.B., Dunn, I., Aradhya, S., Kitzman, J.O., and Yee, J.L. (2023). High-Throughput Splicing Assays Identify Known and Novel WT1 Exon 9 Variants in Nephrotic Syndrome. Kidney Int. Rep. *8*, 2117–2125. https://doi.org/10.1016/j.ekir.2023.07.033.

54. Abell, N.S., DeGorter, M.K., Gloudemans, M.J., Greenwald, E., Smith, K.S., He, Z., and Montgomery, S.B. (2022). Multiple causal variants underlie genetic associations in humans. Science *375*, 1247–1254. https://doi.org/10.1126/science.abj5117.

55. Maricque, B.B., Dougherty, J.D., and Cohen, B.A. (2016). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of *cis* -regulatory activity in neural cells. Nucleic Acids Res., gkw942. https://doi.org/10.1093/nar/gkw942.

56. Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and

Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods *17*, 1083–1091. https://doi.org/10.1038/s41592-020-0965-y.

57. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R.J., Costello, J.F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. *10*, 3583.

58. Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., and Schaffner, S.F. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell *165*, 1519–1529.

59. Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855. https://doi.org/10.1038/ng.3837.

60. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277. https://doi.org/10.1038/nbt.2137.

61. Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. Cell *163*, 698–711. https://doi.org/10.1016/j.cell.2015.09.054.

62. Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270. https://doi.org/10.1038/nbt.2136.

63. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175. https://doi.org/10.1038/nbt.1589.

64. Zhao, W., Pollack, J.L., Blagev, D.P., Zaitlen, N., McManus, M.T., and Erle, D.J. (2014). Massively parallel functional annotation of 3′ untranslated regions. Nat. Biotechnol. *32*, 387–391. https://doi.org/10.1038/nbt.2851.

65. Griesemer, D., Xue, J.R., Reilly, S.K., Ulirsch, J.C., Kukreja, K., Davis, J.R., Kanai, M., Yang, D.K., Butts, J.C., Guney, M.H., et al. (2021). Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. Cell *184*, 5247-5260.e19. https://doi.org/10.1016/j.cell.2021.08.025.

66. Sample, P.J., Wang, B., Reid, D.W., Presnyak, V., McFadyen, I.J., Morris, D.R., and Seelig, G. (2019). Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat. Biotechnol. *37*, 803–809. https://doi.org/10.1038/s41587-019-0164-5.

67. Castillo-Hair, S., Fedak, S., Wang, B., Linder, J., Havens, K., Certo, M., and Seelig, G. (2024). Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning. Nat. Commun. *15*, 5284.

68. Rabani, M., Pieper, L., Chew, G.-L., and Schier, A.F. (2017). A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. Mol. Cell *68*, 1083-1094.e5. https://doi.org/10.1016/j.molcel.2017.11.014.

69. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806. https://doi.org/10.1126/science.1254806.

70. Ling, J.P., Wilks, C., Charles, R., Leavey, P.J., Ghosh, D., Jiang, L., Santiago, C.P., Pang, B., Venkataraman, A., Clark, B.S., et al. (2020). ASCOT identifies key regulators of neuronal subtype-specific splicing. Nat. Commun. *11*, 137. https://doi.org/10.1038/s41467-019-14020-5.

71. Cheng, J., Çelik, M.H., Kundaje, A., and Gagneur, J. (2021). MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. *22*, 94. https://doi.org/10.1186/s13059-021-02273-7.

72. The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*, e1001046. https://doi.org/10.1371/journal.pbio.1001046.

73. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

74. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015.

75. Smith, C., and Kitzman, J.O. (2023). Benchmarking splice variant prediction algorithms using massively parallel splicing assays. Preprint, https://doi.org/10.1101/2023.05.04.539398 https://doi.org/10.1101/2023.05.04.539398.

76. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158. https://doi.org/10.1038/s41588-017-0004-9.

77. Lorson, C.L. (2000). An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. Hum. Mol. Genet. *9*, 259–265. https://doi.org/10.1093/hmg/9.2.259.

78. Lorson, C.L., Hahnen, E., Androphy, E.J., and Wirth, B. (1999). A single nucleotide in the *SMN* gene regulates splicing and is responsible for spinal muscular atrophy. Proc. Natl.

Acad. Sci. *96*, 6307–6311. https://doi.org/10.1073/pnas.96.11.6307.

79. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985. https://doi.org/10.1093/nar/gkt1113.

80. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45*, D840–D845. https://doi.org/10.1093/nar/gkw971.

81. The Geuvadis Consortium, Lappalainen, T., Sammeth, M., Friedländer, M.R., 'T Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511. https://doi.org/10.1038/nature12531.

82. Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855. https://doi.org/10.1038/ng.3837.

83. Baeza-Centurion, P., Miñana, B., Schmiedel, J.M., Valcárcel, J., and Lehner, B. (2019). Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell *176*, 549-563.e23. https://doi.org/10.1016/j.cell.2018.12.010.

84. Mertes, C., Scheller, I.F., Yépez, V.A., Çelik, M.H., Liang, Y., Kremer, L.S., Gusic, M., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. Nat. Commun. *12*, 529. https://doi.org/10.1038/s41467-020-20573-7.

85. Baeza-Centurion, P., Miñana, B., Valcárcel, J., and Lehner, B. (2020). Mutations primarily alter the inclusion of alternatively spliced exons. eLife *9*, e59959. https://doi.org/10.7554/eLife.59959.

86. Louie, W., Shen, M.W., Tahiry, Z., Zhang, S., Worstell, D., Cassa, C.A., Sherwood, R.I., and Gifford, D.K. (2021). Machine learning based CRISPR gRNA design for therapeutic exon skipping. PLOS Comput. Biol. *17*, e1008605. https://doi.org/10.1371/journal.pcbi.1008605.

87. Hervoso, J.L., Amoah, K., Dodson, J., Choudhury, M., Bhattacharya, A., Quinones-Valdez, G., Pasaniuc, B., and Xiao, X. (2024). Splicing-specific transcriptome-wide association uncovers genetic mechanisms for schizophrenia. Am. J. Hum. Genet. *111*, 1573–1587. https://doi.org/10.1016/j.ajhg.2024.06.001.

88. Dao, K., Jungers, C.F., Djuranovic, S., and Mustoe, A.M. (2025). U-rich elements drive pervasive cryptic splicing in 3' UTR massively parallel reporter assays. Nat. Commun. *16*, 6844. https://doi.org/10.1038/s41467-025-62000-9.

89. Cheng, J., Nguyen, T.Y.D., Cygan, K.J., Çelik, M.H., Fairbrother, W.G., Avsec, žiga, and

Gagneur, J. (2019). MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. *20*, 48. https://doi.org/10.1186/s13059-019-1653-z.

90. Arnaud, P., Cadenet, M., Mougin, Z., Le Goff, C., Perbet, S., Francois, M., Dupuis-Girod, S., Boileau, C., and Hanna, N. (2023). Early-Onset Aortic Dissection: Characterization of a New Pathogenic Splicing Variation in the MYH11 Gene with Several In-Frame Abnormal Transcripts. Hum. Mutat. *2023*, 1–7. https://doi.org/10.1155/2023/1410230.

91. Chesneau, B., Plancke, A., Rolland, G., Marcheix, B., Dulac, Y., Edouard, T., Plaisancié, J., Aubert‐Mucca, M., Julia, S., Langeois, M., et al. (2021). A +3 variant at a donor splice site leads to a skipping of the *MYH11* exon 32, a recurrent RNA defect causing Heritable Thoracic Aortic Aneurysm and Dissection and/or Patent Ductus Arteriosus. Mol. Genet. Genomic Med. *9*, e1814. https://doi.org/10.1002/mgg3.1814.

92. Eddinger, T.J., and Meer, D.P. (2007). Myosin II isoforms in smooth muscle: heterogeneity and function. Am. J. Physiol.-Cell Physiol. *293*, C493–C508. https://doi.org/10.1152/ajpcell.00131.2007.

93. Ohyabu, L., Takasaki, T., Akiba, S., Nomum, S., Enokizono, N., Sagara, Y., Hirol, J., Nagai, R., and Yoshida, H. (1998). Immunohistochemical studies on expression of human vascular smooth muscle myosin heavy chain isoforms in normal mammary glands, benign mammary disorders and mammary carcinomas. Pathol. Int. *48*, 433–439. https://doi.org/10.1111/j.1440-1827.1998.tb03929.x.

94. Colbert, M.C., Kirby, M.L., and Robbins, J. (1996). Endogenous Retinoic Acid Signaling Colocalizes With Advanced Expression of the Adult Smooth Muscle Myosin Heavy Chain Isoform During Development of the Ductus Arteriosus. Circ. Res. *78*, 790–798. https://doi.org/10.1161/01.RES.78.5.790.

95. Arnaud, P., Hanna, N., Benarroch, L., Aubart, M., Bal, L., Bouvagnet, P., Busa, T., Dulac, Y., Dupuis-Girod, S., Edouard, T., et al. (2019). Genetic diversity and pathogenic variants as possible predictors of severity in a French sample of nonsyndromic heritable thoracic aortic aneurysms and dissections (nshTAAD). Genet. Med. *21*, 2015–2024. https://doi.org/10.1038/s41436-019-0444-y.

96. Mariscalco, G., Debiec, R., Elefteriades, J.A., Samani, N.J., and Murphy, G.J. (2018). Systematic Review of Studies That Have Evaluated Screening Tests in Relatives of Patients Affected by Nonsyndromic Thoracic Aortic Disease. J. Am. Heart Assoc. *7*, e009302. https://doi.org/10.1161/JAHA.118.009302.

97. Alhopuro, P., Phichith, D., Tuupanen, S., Sammalkorpi, H., Nybondas, M., Saharinen, J., Robinson, J.P., Yang, Z., Chen, L.-Q., Orntoft, T., et al. (2008). Unregulated smooth-muscle myosin in human intestinal neoplasia. Proc. Natl. Acad. Sci. *105*, 5513–5518. https://doi.org/10.1073/pnas.0801213105.

98. Luo, Y.-B., Mastaglia, F.L., and Wilton, S.D. (2014). Normal and aberrant splicing of *LMNA*. J. Med. Genet. *51*, 215–223. https://doi.org/10.1136/jmedgenet-2013-102119.

99. O'Neill, M.J., Yang, T., Laudeman, J., Calandranis, M.E., Harvey, M.L., Solus, J.F., Roden, D.M., and Glazer, A.M. (2024). ParSE-seq: a calibrated multiplexed assay to facilitate the clinical classification of putative splice-altering variants. Nat. Commun. *15*, 8320. https://doi.org/10.1038/s41467-024-52474-4.

100. Peters, S., Thompson, B.A., Perrin, M., James, P., Zentner, D., Kalman, J.M., Vandenberg, J.I., and Fatkin, D. (2022). Arrhythmic Phenotypes Are a Defining Feature of Dilated Cardiomyopathy-Associated *SCN5A* Variants: A Systematic Review. Circ. Genomic Precis. Med. *15*. https://doi.org/10.1161/CIRCGEN.121.003432.

101. Barc, J., Tadros, R., Glinge, C., Chiang, D.Y., Jouni, M., Simonet, F., Jurgens, S.J., Baudic, M., Nicastro, M., Potet, F., et al. (2022). Genome-wide association analyses identify new Brugada syndrome risk loci and highlight a new mechanism of sodium channel regulation in disease susceptibility. Nat. Genet. *54*, 232–239. https://doi.org/10.1038/s41588-021-01007-6.

102. O'Neill, M.J., Wada, Y., Hall, L.D., Mitchell, D.W., Glazer, A.M., and Roden, D.M. (2022). Functional Assays Reclassify Suspected Splice-Altering Variants of Uncertain Significance in Mendelian Channelopathies. Circ. Genomic Precis. Med. *15*. https://doi.org/10.1161/CIRCGEN.122.003782.

103. Bardai, A., Amin, A.S., Blom, M.T., Bezzina, C.R., Berdowski, J., Langendijk, P.N.J., Beekman, L., Klemens, C.A., Souverein, P.C., Koster, R.W., et al. (2013). Sudden cardiac arrest associated with use of a non-cardiac drug that reduces cardiac excitability: evidence from bench, bedside, and community. Eur. Heart J. *34*, 1506–1516. https://doi.org/10.1093/eurheartj/eht054.

104. Hong, K., Guerchicoff, A., Pollevick, G., Oliva, A., Dumaine, R., Dezutter, M., Burashnikov, E., Wu, Y., Brugada, J., and Brugada, P. (2005). Cryptic 5? splice site activation in SCN5A associated with Brugada syndrome. J. Mol. Cell. Cardiol. *38*, 555–560. https://doi.org/10.1016/j.yjmcc.2004.10.015.

105. Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. Dis. Model. Mech. *3*, 133–135. https://doi.org/10.1242/dmm.005439.

106. Miller, D.T., Lee, K., Abul-Husn, N.S., Amendola, L.M., Brothers, K., Chung, W.K., Gollob, M.H., Gordon, A.S., Harrison, S.M., Hershberger, R.E., et al. (2023). ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *25*, 100866. https://doi.org/10.1016/j.gim.2023.100866.

107. Sondka, Z., Dhir, N.B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M., et al. (2024). COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic Acids Res. *52*, D1210–D1217. https://doi.org/10.1093/nar/gkad986.

108. Bhattacharya, S., Dunn, P., Thomas, C.G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K.A., Shankar, R.D., Shen-Orr, S.S., et al. (2018). ImmPort, toward repurposing

of open access immunological assay data for translational and clinical research. Sci. Data *5*, 180015. https://doi.org/10.1038/sdata.2018.15.

109. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

110. Glidden, D.T., Buerer, J.L., Saueressig, C.F., and Fairbrother, W.G. (2021). Hotspot exons are common targets of splicing perturbations. Nat. Commun. *12*, 2756. https://doi.org/10.1038/s41467-021-22780-2.

111. Dourlen, P., Kilinc, D., Landrieu, I., Chapuis, J., and Lambert, J.-C. (2025). BIN1 and Alzheimer's disease: the tau connection. Trends Neurosci. *48*, 349–361. https://doi.org/10.1016/j.tins.2025.03.004.

112. Wechsler-Reya, R., Sakamuro, D., Zhang, J., Duhadaway, J., and Prendergast, G.C. (1997). Structural Analysis of the Human BIN1 Gene. J. Biol. Chem. *272*, 31453–31458. https://doi.org/10.1074/jbc.272.50.31453.

113. Spooner, H.C., and Dixon, R.E. (2025). Bend it like BIN1: how a membrane-curving adaptor protein shapes cardiac physiology. Am. J. Physiol.-Heart Circ. Physiol. *329*, H94–H108. https://doi.org/10.1152/ajpheart.00198.2025.

114. Perdreau-Dahl, H., Lipsett, D.B., Frisk, M., Kermani, F., Carlson, C.R., Brech, A., Shen, X., Bergan-Dahl, A., Hou, Y., Tuomainen, T., et al. (2023). BIN1, Myotubularin, and Dynamin-2 Coordinate T-Tubule Growth in Cardiomyocytes. Circ. Res. *132*. https://doi.org/10.1161/CIRCRESAHA.122.321732.

115. Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome. WIREs RNA *4*, 61–76. https://doi.org/10.1002/wrna.1141.

116. Singh, R.K., and Cooper, T.A. (2012). Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med. *18*, 472–482. https://doi.org/10.1016/j.molmed.2012.06.006.

117. Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nat. Rev. Genet. *17*, 19–32. https://doi.org/10.1038/nrg.2015.3.

118. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222. https://doi.org/10.1038/s41586-018-0461-z.

119. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424. https://doi.org/10.1038/gim.2015.30.

120. Abou Tayoun, A.N., Pesaran, T., DiStefano, M.T., Oza, A., Rehm, H.L., Biesecker, L.G., Harrison, S.M., and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum. Mutat. *39*, 1517–1524. https://doi.org/10.1002/humu.23626.

121. Anna, A., and Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. J. Appl. Genet. *59*, 253–268. https://doi.org/10.1007/s13353-018-0444-7.

122. Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M.J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat. Rev. Mol. Cell Biol. *14*, 153–165. https://doi.org/10.1038/nrm3525.

123. Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. *10*, 741–754. https://doi.org/10.1038/nrm2777.

124. Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718. https://doi.org/10.1016/j.cell.2009.02.009.

125. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415. https://doi.org/10.1038/ng.259.

126. Eom, T., Zhang, C., Wang, H., Lay, K., Fak, J., Noebels, J.L., and Darnell, R.B. (2013). NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. eLife *2*, e00178. https://doi.org/10.7554/eLife.00178.

127. Piton, A. (2025). NOVA1/2 genes and alternative splicing in neurodevelopment. Curr. Opin. Genet. Dev. *93*, 102373. https://doi.org/10.1016/j.gde.2025.102373.

128. Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580–586. https://doi.org/10.1038/nature05304.

129. Zhang, X., Chen, M.H., Wu, X., Kodani, A., Fan, J., Doan, R., Ozawa, M., Ma, J., Yoshida, N., Reiter, J.F., et al. (2016). Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell *166*, 1147-1162.e15. https://doi.org/10.1016/j.cell.2016.07.025.

130. Freytag, M., Kluth, M., Bady, E., Hube-Magg, C., Makrypidi-Fraune, G., Heinzer, H., Höflmayer, D., Weidemann, S., Uhlig, R., Huland, H., et al. (2020). Epithelial splicing regulatory protein 1 and 2 (ESRP1 and ESRP2) upregulation predicts poor prognosis in prostate cancer. BMC Cancer *20*, 1220. https://doi.org/10.1186/s12885-020-07682-8.

131. Ishii, H., Saitoh, M., Sakamoto, K., Kondo, T., Katoh, R., Tanaka, S., Motizuki, M., Masuyama, K., and Miyazawa, K. (2014). Epithelial Splicing Regulatory Proteins 1 (ESRP1) and 2 (ESRP2) Suppress Cancer Cell Motility via Different Mechanisms. J. Biol.

Chem. *289*, 27386–27399. https://doi.org/10.1074/jbc.M114.589432.

132. Advani, R., Luzzi, S., Scott, E., Dalgliesh, C., Weischenfeldt, J., Munkley, J., and Elliott, D.J. (2023). Epithelial specific splicing regulator proteins as emerging oncogenes in aggressive prostate cancer. Oncogene *42*, 3161–3168. https://doi.org/10.1038/s41388-023-02838-9.

133. Taylor, K., Sznajder, Ł.J., Cywoniuk, P., Thomas, J.D., Swanson, M.S., and Sobczak, K. (2018). MBNL splicing activity depends on RNA binding site structural context. Nucleic Acids Res. *46*, 9119–9133. https://doi.org/10.1093/nar/gky565.

134. Lin, X., Miller, J.W., Mankodi, A., Kanadia, R.N., Yuan, Y., Moxley, R.T., Swanson, M.S., and Thornton, C.A. (2006). Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. Hum. Mol. Genet. *15*, 2087–2097. https://doi.org/10.1093/hmg/ddl132.

135. Zhou, H., Xu, J., and Pan, L. (2025). Functions of the Muscleblind-like protein family and their role in disease. Cell Commun. Signal. *23*, 97. https://doi.org/10.1186/s12964-025-02102-5.

136. Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J. Neurosci. *34*, 11929–11947. https://doi.org/10.1523/JNEUROSCI.1860-14.2014.

137. Song, Y., Botvinnik, O.B., Lovci, M.T., Kakaradov, B., Liu, P., Xu, J.L., and Yeo, G.W. (2017). Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. Mol. Cell *67*, 148-161.e5. https://doi.org/10.1016/j.molcel.2017.06.003.

138. Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59. https://doi.org/10.1038/nature09000.

139. Furlanis, E., Traunmüller, L., Fucile, G., and Scheiffele, P. (2019). Landscape of ribosome-engaged transcript isoforms reveals extensive neuronal-cell-class-specific alternative splicing programs. Nat. Neurosci. *22*, 1709–1717. https://doi.org/10.1038/s41593-019-0465-5.

140. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans (2015). Science *348*, 648–660. https://doi.org/10.1126/science.1262110.

141. Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. *11*, 75–87. https://doi.org/10.1038/nrg2673.

142. Weyn-Vanhentenryck, S.M., Feng, H., Ustianenko, D., Duffié, R., Yan, Q., Jacko, M., Martinez, J.C., Goodwin, M., Zhang, X., Hengst, U., et al. (2018). Precise temporal

regulation of alternative splicing during neural development. Nat. Commun. *9*, 2189. https://doi.org/10.1038/s41467-018-04559-0.

143. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177. https://doi.org/10.1038/nature12311.

144. Benoit Bouvrette, L.P., Bovaird, S., Blanchette, M., and Lécuyer, E. (2019). oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. Nucleic Acids Res., gkz986. https://doi.org/10.1093/nar/gkz986.

145. Giudice, G., Sánchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). ATtRACT—a database of RNA-binding proteins and associated motifs. Database *2016*, baw035. https://doi.org/10.1093/database/baw035.

146. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

147. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. *45*, e119–e119. https://doi.org/10.1093/nar/gkx314.

148. Reimão-Pinto, M.M., Castillo-Hair, S.M., Seelig, G., and Schier, A.F. (2025). The regulatory landscape of 5′ UTRs in translational control during zebrafish embryogenesis. Dev. Cell *60*, 1498-1515.e8. https://doi.org/10.1016/j.devcel.2024.12.038.

149. Gajos, M., Jasnovidova, O., van Bömmel, A., Freier, S., Vingron, M., and Mayer, A. (2021). Conserved DNA sequence features underlie pervasive RNA polymerase pausing. Nucleic Acids Res. *49*, 4402–4420. https://doi.org/10.1093/nar/gkab208.

150. Reber, S., Stettler, J., Filosa, G., Colombo, M., Jutzi, D., Lenzken, S.C., Schweingruber, C., Bruggmann, R., Bachi, A., Barabino, S.M., et al. (2016). Minor intron splicing is regulated by FUS and affected by ALS‐associated FUS mutants. EMBO J. *35*, 1504–1521. https://doi.org/10.15252/embj.201593791.

151. Horn, T., Gosliga, A., Li, C., Enculescu, M., and Legewie, S. (2023). Position-dependent effects of RNA-binding proteins in the context of co-transcriptional splicing. Npj Syst. Biol. Appl. *9*, 1–22. https://doi.org/10.1038/s41540-022-00264-3.

152. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719. https://doi.org/10.1038/s41586-020-2077-3.

153. Witten, J.T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. Trends Genet. *27*, 89–97. https://doi.org/10.1016/j.tig.2010.12.001.

154. Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., et al. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell *70*, 854-867.e9. https://doi.org/10.1016/j.molcel.2018.05.001.

155. Wang, J. (2005). Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. Nucleic Acids Res. *33*, 5053–5062. https://doi.org/10.1093/nar/gki810.

156. Liu, H.-X., Zhang, M., and Krainer, A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. *12*, 1998–2012. https://doi.org/10.1101/gad.12.13.1998.

157. Jobbins, A.M., Reichenbach, L.F., Lucas, C.M., Hudson, A.J., Burley, G.A., and Eperon, I.C. (2018). The mechanisms of a mammalian splicing enhancer. Nucleic Acids Res. *46*, 2145–2158. https://doi.org/10.1093/nar/gky056.

158. Schaal, T.D., and Maniatis, T. (1999). Multiple Distinct Splicing Enhancers in the Protein-Coding Sequences of a Constitutively Spliced Pre-mRNA. Mol. Cell. Biol. *19*, 261–273. https://doi.org/10.1128/MCB.19.1.261.

159. Xiao, X., Wang, Z., Jang, M., Nutiu, R., Wang, E.T., and Burge, C.B. (2009). Splice site strength–dependent activity and genetic buffering by poly-G runs. Nat. Struct. Mol. Biol. *16*, 1094–1100. https://doi.org/10.1038/nsmb.1661.

160. Caputi, M., and Zahler, A.M. (2001). Determination of the RNA Binding Specificity of the Heterogeneous Nuclear Ribonucleoprotein (hnRNP) H/H′/F/2H9 Family. J. Biol. Chem. *276*, 43850–43859. https://doi.org/10.1074/jbc.M102861200.

161. Schaub, M.C., Lopez, S.R., and Caputi, M. (2007). Members of the Heterogeneous Nuclear Ribonucleoprotein H Family Activate Splicing of an HIV-1 Splicing Substrate by Promoting Formation of ATP-dependent Spliceosomal Complexes. J. Biol. Chem. *282*, 13617–13626. https://doi.org/10.1074/jbc.M700774200.

162. Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B., and Carstens, R.P. (2009). ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. Mol. Cell *33*, 591–601. https://doi.org/10.1016/j.molcel.2009.01.025.

163. Derham, J.M., and Kalsotra, A. (2023). The discovery, function, and regulation of epithelial splicing regulatory proteins (ESRP) 1 and 2. Biochem. Soc. Trans. *51*, 1097–1109. https://doi.org/10.1042/BST20221124.

164. Tan, A.Y., and Manley, J.L. (2009). The TET Family of Proteins: Functions and Roles in Disease. J. Mol. Cell Biol. *1*, 82–92. https://doi.org/10.1093/jmcb/mjp025.

165. Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Mol. Cell

36, 996–1006. https://doi.org/10.1016/j.molcel.2009.12.003.

166. Ye, R., Hu, N., Cao, C., Su, R., Xu, S., Yang, C., Zhou, X., and Xue, Y. (2023). Capture RIC-seq reveals positional rules of PTBP1-associated RNA loops in splicing regulation. Mol. Cell *83*, 1311-1327.e7. https://doi.org/10.1016/j.molcel.2023.03.001.

167. Boutz, P.L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., and Black, D.L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev. *21*, 1636–1652.

168. Liu, H.-L., Lu, X.-M., Wang, H.-Y., Hu, K.-B., Wu, Q.-Y., Liao, P., Li, S., Long, Z.-Y., and Wang, Y.-T. (2023). The role of RNA splicing factor PTBP1 in neuronal development. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1870*, 119506. https://doi.org/10.1016/j.bbamcr.2023.119506.

169. Spellman, R., Rideau, A., Matlin, A., Gooding, C., Robinson, F., McGlincy, N., Grellscheid, S.N., Southby, J., Wollerton, M., and Smith, C.W.J. (2005). Regulation of alternative splicing by PTB and associated factors. Biochem. Soc. Trans. *33*, 457–460.

170. Keppetipola, N., Sharma, S., Li, Q., and Black, D.L. (2012). Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Crit. Rev. Biochem. Mol. Biol. *47*, 360–378.

171. Kwiatek, L., Landry-Voyer, A.-M., Latour, M., Yague-Sanz, C., and Bachand, F. (2023). PABPN1 prevents the nuclear export of an unspliced RNA with a constitutive transport element and controls human gene expression via intron retention. RNA *29*, 644–662. https://doi.org/10.1261/rna.079294.122.

172. Bergeron, D., Pal, G., Beaulieu, Y.B., Chabot, B., and Bachand, F. (2015). Regulated Intron Retention and Nuclear Pre-mRNA Decay Contribute to *PABPN1* Autoregulation. Mol. Cell. Biol. *35*, 2503–2517. https://doi.org/10.1128/MCB.00070-15.

173. Huang, L., Li, G., Du, C., Jia, Y., Yang, J., Fan, W., Xu, Y., Cheng, H., and Zhou, Y. (2023). The polyA tail facilitates splicing of last introns with weak 3′ splice sites via PABPN1. EMBO Rep. *24*, e57128. https://doi.org/10.15252/embr.202357128.

174. Caetano, S., Garcia, A.R., Figueira, I., and Brito, M.A. (2023). MEF2C and miR-194-5p: New Players in Triple Negative Breast Cancer Tumorigenesis. Int. J. Mol. Sci. *24*, 14297. https://doi.org/10.3390/ijms241814297.

175. Hakim, N.H.A., Kounishi, T., Alam, A.H.M.K., Tsukahara, T., and Suzuki, H. (2010). Alternative splicing of *Mef2c* promoted by Fox‐1 during neural differentiation in P19 cells. Genes Cells *15*, 255–267. https://doi.org/10.1111/j.1365-2443.2009.01378.x.

176. Sekiyama, Y., Suzuki, H., and Tsukahara, T. (2012). Functional Gene Expression Analysis of Tissue-Specific Isoforms of Mef2c. Cell. Mol. Neurobiol. *32*, 129–139. https://doi.org/10.1007/s10571-011-9743-9.

177. Di Giorgio, E., Hancock, W.W., and Brancolini, C. (2018). MEF2 and the tumorigenic process, hic sunt leones. Biochim. Biophys. Acta BBA - Rev. Cancer *1870*, 261–273. https://doi.org/10.1016/j.bbcan.2018.05.007.

178. Park, S., Brugiolo, M., Akerman, M., Das, S., Urbanski, L., Geier, A., Kesarwani, A.K., Fan, M., Leclair, N., Lin, K.-T., et al. (2019). Differential Functions of Splicing Factors in Mammary Transformation and Breast Cancer Metastasis. Cell Rep. *29*, 2672-2688.e7. https://doi.org/10.1016/j.celrep.2019.10.110.

179. Bei, M., and Xu, J. (2024). SR proteins in cancer: function, regulation, and small inhibitor. Cell. Mol. Biol. Lett. *29*, 78. https://doi.org/10.1186/s11658-024-00594-6.

180. Pon, J.R., and Marra, M.A. (2016). MEF2 transcription factors: developmental regulators and emerging cancer genes. Oncotarget *7*, 2297–2312. https://doi.org/10.18632/oncotarget.6223.

181. Yang, M., Ke, Y., Kim, P., and Zhou, X. (2021). ExonSkipAD provides the functional genomic landscape of exon skipping events in Alzheimer's disease. Brief. Bioinform. *22*, bbaa438. https://doi.org/10.1093/bib/bbaa438.

182. Bosè, F., Renna, L.V., Fossati, B., Arpa, G., Labate, V., Milani, V., Botta, A., Micaglio, E., Meola, G., and Cardani, R. (2019). TNNT2 Missplicing in Skeletal Muscle as a Cardiac Biomarker in Myotonic Dystrophy Type 1 but Not in Myotonic Dystrophy Type 2. Front. Neurol. *10*, 992. https://doi.org/10.3389/fneur.2019.00992.

183. Valero, R., Bayés, M., Francisca Sánchez-Font, M., González-Angulo, O., Gonzàlez-Duarte, R., and Marfany, G. (2001). Characterization of alternatively spliced products and tissue-specific isoforms of USP28 and USP25. Genome Biol. *2*, research0043.1. https://doi.org/10.1186/gb-2001-2-10-research0043.

184. Oh, J., Pradella, D., Kim, Y., Shao, C., Li, H., Choi, N., Ha, J., Di Matteo, A., Fu, X.-D., Zheng, X., et al. (2021). Global Alternative Splicing Defects in Human Breast Cancer Cells. Cancers *13*, 3071. https://doi.org/10.3390/cancers13123071.

185. Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. *25*, 1–13. https://doi.org/10.1101/gr.181990.114.

186. Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell *159*, 1511–1523. https://doi.org/10.1016/j.cell.2014.11.035.

187. Miro, J., Bougé, A.-L., Murauer, E., Beyne, E., Da Cunha, D., Claustres, M., Koenig, M., and Tuffery-Giraud, S. (2020). First Identification of RNA-Binding Proteins That Regulate Alternative Exons in the Dystrophin Gene. Int. J. Mol. Sci. *21*, 7803. https://doi.org/10.3390/ijms21207803.

188. Jin, H., Zhang, C., Zwahlen, M., Von Feilitzen, K., Karlsson, M., Shi, M., Yuan, M., Song, X., Li, X., Yang, H., et al. (2023). Systematic transcriptional analysis of human cell lines for gene expression landscape and tumor representation. Nat. Commun. *14*, 5417. https://doi.org/10.1038/s41467-023-41132-w.

189. Griesemer, D., Xue, J.R., Reilly, S.K., Ulirsch, J.C., Kukreja, K., Davis, J.R., Kanai, M., Yang, D.K., Butts, J.C., Guney, M.H., et al. (2021). Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. Cell *184*, 5247-5260.e19. https://doi.org/10.1016/j.cell.2021.08.025.

190. Linder, J., La Fleur, A., Chen, Z., Ljubeti, A., Baker, D., Kannan, S., and Seelig, G. (2022). Interpreting Neural Networks for Biological Sequences by Learning Stochastic Masks. Nat. Mach. Intell. *4*, 41–54. https://doi.org/10.1038/s42256-021-00428-6.

191. Chan, K.Y., Jang, M.J., Yoo, B.B., Greenbaum, A., Ravi, N., Wu, W.-L., Sánchez-Guardado, L., Lois, C., Mazmanian, S.K., Deverman, B.E., et al. (2017). Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. Nat. Neurosci. *20*, 1172–1179. https://doi.org/10.1038/nn.4593.

192. Naldini, L. (2015). Gene therapy returns to centre stage. Nature *526*, 351–360. https://doi.org/10.1038/nature15818.

193. Ling, J.P., Bygrave, A.M., Santiago, C.P., Carmen-Orozco, R.P., Trinh, V.T., Yu, M., Li, Y., Liu, Y., Bowden, K.D., Duncan, L.H., et al. (2022). Cell-specific regulation of gene expression using splicing-dependent frameshifting. Nat. Commun. *13*, 5773. https://doi.org/10.1038/s41467-022-33523-2.

194. Havens, M.A., and Hastings, M.L. (2016). Splice-switching antisense oligonucleotides as therapeutic drugs. Nucleic Acids Res. *44*, 6549–6563. https://doi.org/10.1093/nar/gkw533.

195. DeNicola, A.B., and Tang, Y. (2019). Therapeutic approaches to treat human spliceosomal diseases. Curr. Opin. Biotechnol. *60*, 72–81. https://doi.org/10.1016/j.copbio.2019.01.003.

196. Du, M., Jillette, N., Zhu, J.J., Li, S., and Cheng, A.W. (2020). CRISPR artificial splicing factors. Nat. Commun. *11*, 2973. https://doi.org/10.1038/s41467-020-16806-4.

197. Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. Nat. Rev. Genet. *14*, 496–506. https://doi.org/10.1038/nrg3482.

198. MacDonald, C.C., and Redondo, J.-L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol. Cell. Endocrinol. *190*, 1–8. https://doi.org/10.1016/S0303-7207(02)00044-8.

199. Tian, B., and Graber, J.H. (2012). Signals for pre‐mRNA cleavage and polyadenylation. Wiley Interdiscip. Rev. RNA *3*, 385–396.

200. Grozdanov, P.N., Masoumzadeh, E., Latham, M.P., and MacDonald, C.C. (2018). The

structural basis of CstF-77 modulation of cleavage and polyadenylation through stimulation of CstF-64 activity. Nucleic Acids Res. *46*, 12022–12039.

201. Shi, Y. (2012). Alternative polyadenylation: new insights from global analyses. Rna *18*, 2105–2117.

202. Derti, A., Garrett-Engele, P., MacIsaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

203. Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell *43*, 853–866.

204. Davis, R., and Shi, Y. (2014). The polyadenylation code: a unified model for the regulation of mRNA alternative poly adenylation. J. Zhejiang Univ. Sci. B *15*, 429–437. https://doi.org/10.1631/jzus.B1400076.

205. Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D., and Chance, P.F. (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. Immunogenetics *53*, 435–439. https://doi.org/10.1007/s002510100358.

206. Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G., et al. (2007). Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood *109*, 4599–4606. https://doi.org/10.1182/blood-2006-08-039859.

207. Stacey, S.N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D.F., Magnusson, O.T., Gudjonsson, S.A., Sigurgeirsson, B., Thorisdottir, K., et al. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat. Genet. *43*, 1098–1103. https://doi.org/10.1038/ng.926.

208. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., Dines, J.N., Hause, R.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat. Genet. *50*, 874–882. https://doi.org/10.1038/s41588-018-0122-z.

209. Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530. https://doi.org/10.1038/nbt.2205.

210. White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. *110*, 11952–11957.

211. Cui, Y., Peng, F., Wang, D., Li, Y., Li, J.S., Li, L., and Li, W. (2022). 3′ aQTL-atlas: an atlas

of 3′ UTR alternative polyadenylation quantitative trait loci across human normal tissues. Nucleic Acids Res *50*, 39–45.

212. Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N.D., Li, Y., Chen, Y.E., Ji, P., Peng, F., et al. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat. Genet. *53*, 994–1005. https://doi.org/10.1038/s41588-021-00864-5.

213. Mittleman, B.E., Pott, S., Warland, S., Zeng, T., Mu, Z., Kaur, M., Gilad, Y., and Li, Y. (2020). Alternative polyadenylation mediates genetic regulation of gene expression. eLife *9*, e57492. https://doi.org/10.7554/eLife.57492.

214. Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant Methods *9*, 29. https://doi.org/10.1186/1746-4811-9-29.

215. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.

216. Vainberg Slutskin, I., Weinberger, A., and Segal, E. (2019). Sequence determinants of polyadenylation-mediated regulation. Genome Res. *29*, 1635–1647. https://doi.org/10.1101/gr.247312.118.

217. Leung, M.K.K., Delong, A., and Frey, B.J. (2018). Inference of the human polyadenylation code. Bioinformatics *34*, 2889–2898. https://doi.org/10.1093/bioinformatics/bty211.

218. Arefeen, A., Xiao, X., and Jiang, T. (2019). DeepPASTA: deep neural network based polyadenylation site analysis. Bioinformatics *35*, 4577–4585. https://doi.org/10.1093/bioinformatics/btz283.

219. Danckwardt, S., Hentze, M.W., and Kulozik, A.E. (2008). 3’ end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J. *27*, 482–498. https://doi.org/10.1038/sj.emboj.7601932.

220. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. https://doi.org/10.1038/nbt.3300.

221. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods *12*, 931–934. https://doi.org/10.1038/nmeth.3547.

222. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18. https://doi.org/10.1038/s41588-018-0295-5.

223. Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

https://doi.org/10.1101/gr.200535.115.

224. Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403. https://doi.org/10.1038/s41576-019-0122-6.

225. Li, Z., Li, Y., Zhang, B., Li, Y., Long, Y., Zhou, J., Zou, X., Zhang, M., Hu, Y., Chen, W., et al. (2020). DeeReCT-APA: Prediction of Alternative Polyadenylation Site Usage Through Deep Learning. Preprint, https://doi.org/10.1101/2020.03.26.009373 https://doi.org/10.1101/2020.03.26.009373.

226. Bogard, N., Linder, J., Rosenberg, A.B., and Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell *178*, 91-106.e23. https://doi.org/10.1016/j.cell.2019.04.046.

227. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6.

228. Wang, R., Nambiar, R., Zheng, D., and Tian, B. (2018). PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. Nucleic Acids Res. *46*, 315–319.

229. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., and Ganna, A. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

230. Wang, J., Huang, D., Zhou, Y., Yao, H., Liu, H., Zhai, S., Wu, C., Zheng, Z., and Zhao, K. (2020). CAUSALdb: a database for disease/ trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Res *48*, 807–816.

231. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., and Delaneau, O. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

232. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., and Mountjoy, E. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res *47*, 1005–1012.

233. Kanai, M., Ulirsch, J.C., Karjalainen, J., Kurki, M., Karczewski, K.J., Fauman, E., Wang, Q.S., Jacobs, H., and Aguet, F. (2021). Insights from complex trait fine-mapping across diverse populations. medRxiv, 09 03 21262 975.

234. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576.

https://doi.org/10.1126/science.aat6576.

235. Newnham, C.M., Hall-Pogar, T., Liang, S., Wu, J., Tian, B., Hu, J., and Lutz, C.S. (2010). Alternative polyadenylation of MeCP2: influence of cis-acting elements and trans-acting factors. RNA Biol. *7*, 361–372. https://doi.org/10.4161/rna.7.3.11564.

236. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron *68*, 192–195. https://doi.org/10.1016/j.neuron.2010.10.006.

237. Litterman, A.J., Kageyama, R., Le Tonqueze, O., Zhao, W., Gagnon, J.D., Goodarzi, H., Erle, D.J., and Ansel, K.M. (2019). A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. Genome Res. *29*, 896–906. https://doi.org/10.1101/gr.242552.118.

238. Alsheikh, A.J., Wollenhaupt, S., King, E.A., Reeb, J., Ghosh, S., Stolzenburg, L.R., Tamim, S., Lazar, J., Davis, J.W., and Jacob, H.J. (2022). The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. BMC Med. Genomics *15*, 74. https://doi.org/10.1186/s12920-022-01216-w.

239. Yin, C., Castillo-Hair, S., Byeon, G.W., Bromley, P., Meuleman, W., and Seelig, G. (2025). Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. Cell Syst. *16*, 101302. https://doi.org/10.1016/j.cels.2025.101302.

240. Friedman, R.Z., Ramu, A., Lichtarge, S., Wu, Y., Tripp, L., Lyon, D., Myers, C.A., Granas, D.M., Gause, M., Corbo, J.C., et al. (2025). Active learning of enhancers and silencers in the developing neural retina. Cell Syst. *0*. https://doi.org/10.1016/j.cels.2024.12.004.

241. Liu, L., Yu, A.M., Wang, X., Soles, L.V., Chen, Y., Yoon, Y., Sarkan, K.S.K., Valdez, M.C., Linder, J., Marazzi, I., et al. (2023). The anti-cancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3′ processing machinery. Preprint, https://doi.org/10.1101/2023.04.11.536453 https://doi.org/10.1101/2023.04.11.536453.

242. Linder, J., La Fleur, A., Chen, Z., Ljubeti, A., Baker, D., Kannan, S., and Seelig, G. (2022). Interpreting Neural Networks for Biological Sequences by Learning Stochastic Masks. Nat. Mach. Intell. *4*, 41–54. https://doi.org/10.1038/s42256-021-00428-6.

243. Chen, X., Lloyd, S.M., Kweon, J., Gamalong, G.M., and Bao, X. (2021). Epidermal progenitors suppress GRHL3-mediated differentiation through intronic polyadenylation promoted by CPSF-HNRNPA3 collaboration. Nat. Commun. *12*, 448. https://doi.org/10.1038/s41467-020-20674-3.

244. Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. *15*, 163–175.

245. Dezazzo, J.D., and Imperiale, M.J. (1989). Sequences upstream of AAUAAA influence poly (A) site selection in a complex transcription unit. Mol. Cell. Biol. *9*, 4951–4961.

246. Shulman, E.D., and Elkon, R. (2019). Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. Nucleic Acids Res. *47*, 10027–10039. https://doi.org/10.1093/nar/gkz781.

247. Colombo, I., Sangiovanni, E., Maggio, R., Mattozzi, C., Zava, S., Corbett, Y., Fumagalli, M., Carlino, C., Corsetto, P.A., Scaccabarozzi, D., et al. (2017). HaCaT Cells as a Reliable In Vitro Differentiation Model to Dissect the Inflammatory/Repair Response of Human Keratinocytes. Mediators Inflamm. *2017*, 1–12. https://doi.org/10.1155/2017/7435621.

248. Cui, Y., Wang, L., Ding, Q., Shin, J., Cassel, J., Liu, Q., Salvino, J.M., and Tian, B. (2023). Elevated pre-mRNA 3′ end processing activity in cancer cells renders vulnerability to inhibition of cleavage and polyadenylation. Nat. Commun. *14*, 4480. https://doi.org/10.1038/s41467-023-39793-8.

249. Pagani, F. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum. Mol. Genet. *12*, 1111–1120. https://doi.org/10.1093/hmg/ddg131.

250. Pagani, F., Raponi, M., and Baralle, F.E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc. Natl. Acad. Sci. *102*, 6368–6372. https://doi.org/10.1073/pnas.0502288102.

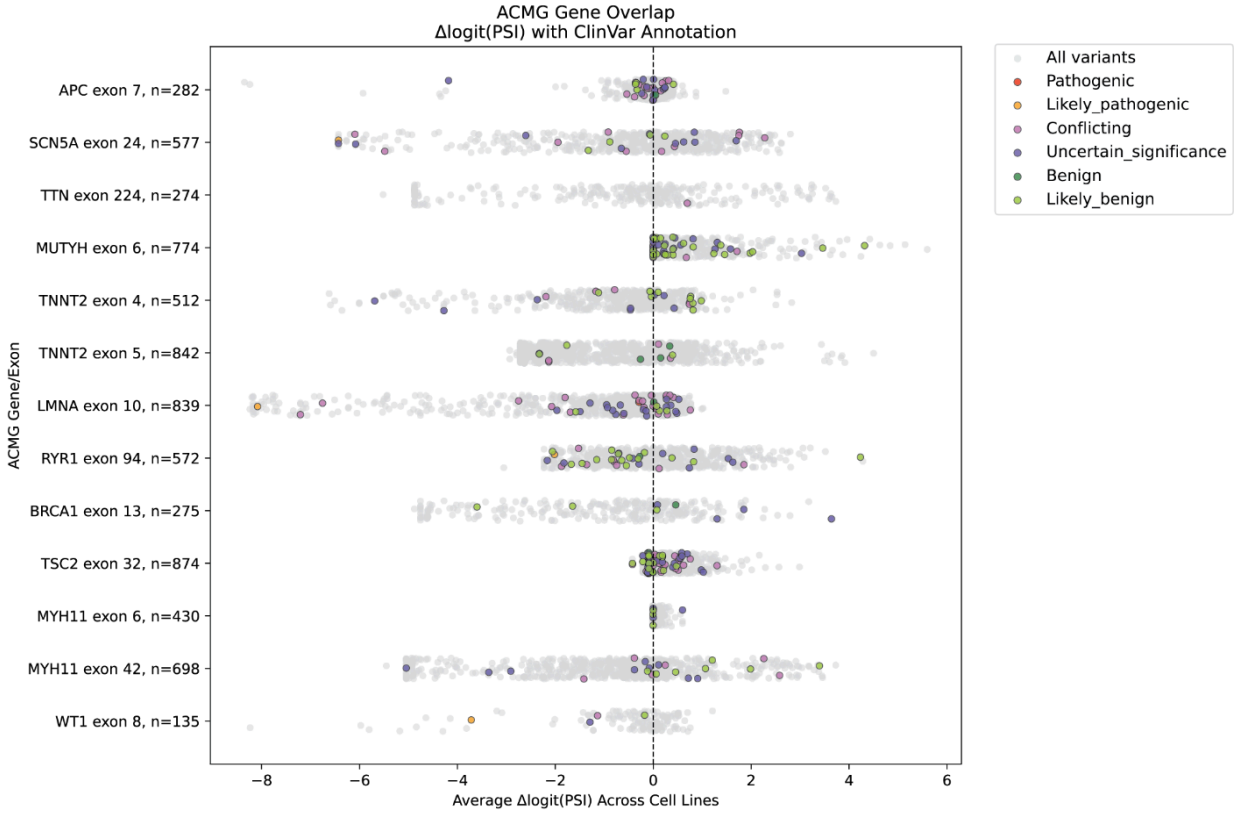
251. Dao, K., Jungers, C.F., Djuranovic, S., and Mustoe, A.M. (2024). U-rich elements drive pervasive cryptic splicing in 3’ UTR massively parallel reporter assays. Preprint, https://doi.org/10.1101/2024.08.05.606557 https://doi.org/10.1101/2024.08.05.606557.

252. Doman, J.L., Raguram, A., Newby, G.A., and Liu, D.R. (2020). Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. Nat. Biotechnol. *38*, 620–628. https://doi.org/10.1038/s41587-020-0414-6.

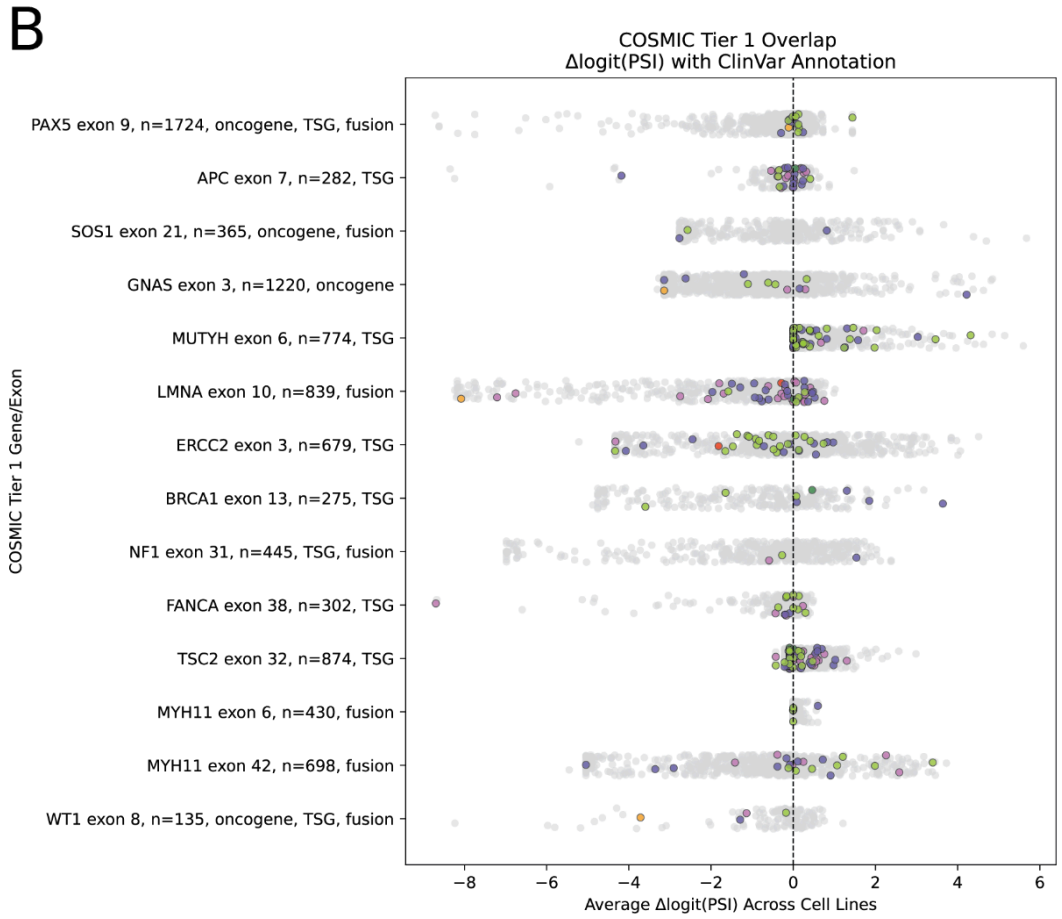
253. Mathis, N., Allam, A., Kissling, L., Marquart, K.F., Schmidheini, L., Solari, C., Balázs, Z., Krauthammer, M., and Schwank, G. (2023). Predicting prime editing efficiency and product purity by deep learning. Nat. Biotechnol. *41*, 1151–1159. https://doi.org/10.1038/s41587-022-01613-7.

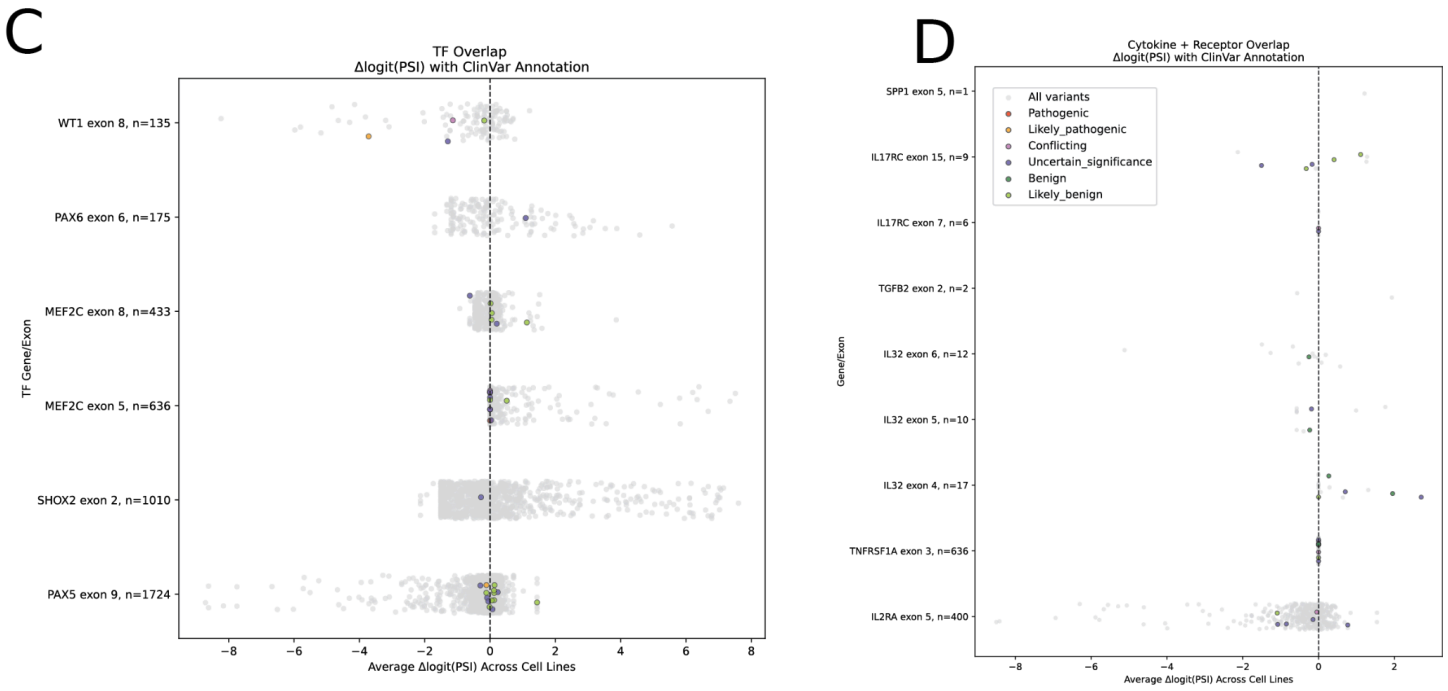
Appendix 1

A


ACMG Gene Overlap
Δlogit(PSI) with ClinVar Annotation
All variants
Pathogenic
Likely_pathogenic
Conflicting
Uncertain_significance
Benign
Likely_benign
APC exon 7, n=282
SCN5A exon 24, n=577
TTN exon 224, n=274
MUTYH exon 6, n=774
TNNT2 exon 4, n=512
TNNT2 exon 5, n=842
LMNA exon 10, n=839
RYR1 exon 94, n=572
BRCA1 exon 13, n=275
TSC2 exon 32, n=874
MYH11 exon 6, n=430
MYH11 exon 42, n=698
WT1 exon 8, n=135
ACMG Gene/Exon
−8
−6
−4
−2
0
2
4
6
Average Δlogit(PSI) Across Cell Lines

B


COSMIC Tier 1 Overlap
Δlogit(PSI) with ClinVar Annotation
PAX5 exon 9, n=1724, oncogene, TSG, fusion
APC exon 7, n=282, TSG
SOS1 exon 21, n=365, oncogene, fusion
GNAS exon 3, n=1220, oncogene
MUTYH exon 6, n=774, TSG
LMNA exon 10, n=839, fusion
ERCC2 exon 3, n=679, TSG
BRCA1 exon 13, n=275, TSG
NF1 exon 31, n=445, TSG, fusion
FANCA exon 38, n=302, TSG
TSC2 exon 32, n=874, TSG
MYH11 exon 6, n=430, fusion
MYH11 exon 42, n=698, fusion
WT1 exon 8, n=135, oncogene, TSG, fusion
COSMIC Tier 1 Gene/Exon
−8
−6
−4
−2
0
2
4
6
Average Δlogit(PSI) Across Cell Lines



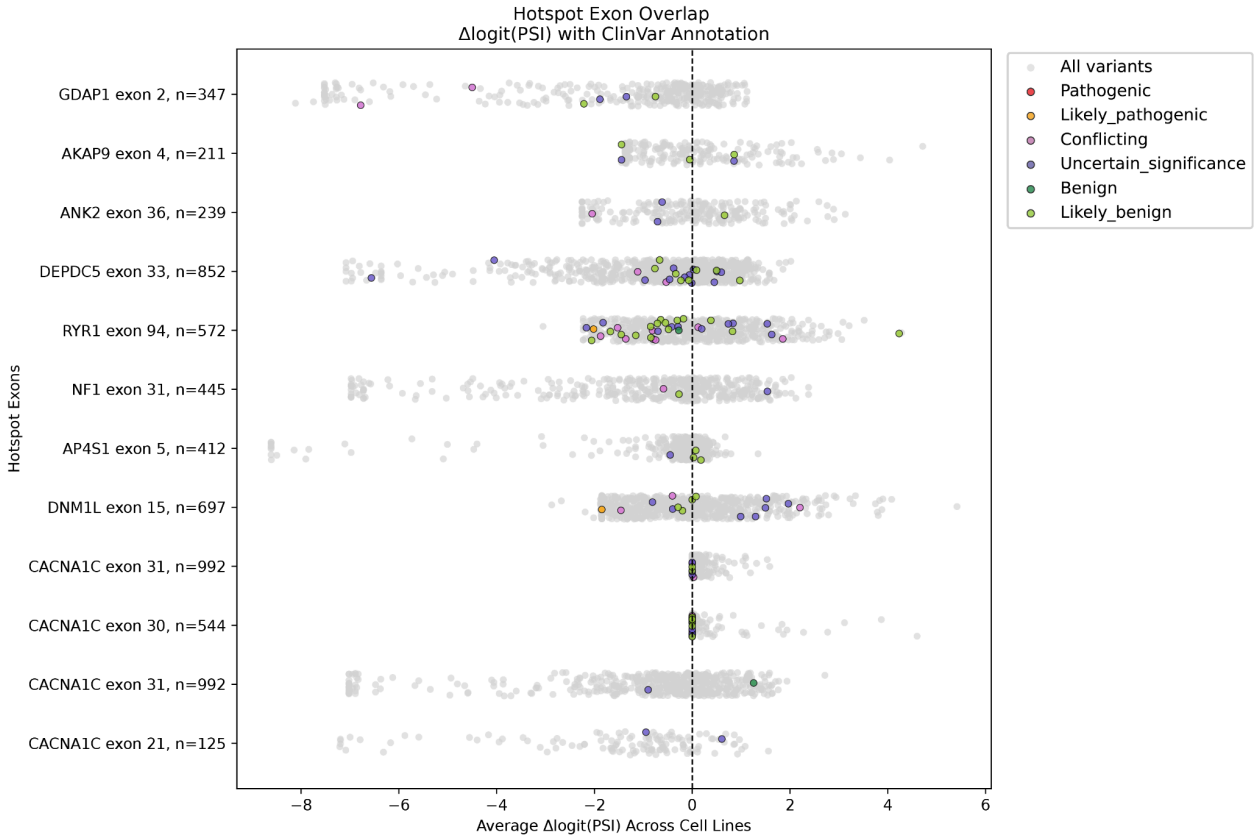
Appendix 1.1 | Δlogit(PSI) distributions for variants in clinically and biologically important gene sets.

(A) Variants from genes in the ACMG SF v3.2 list.[106] Of the 81 genes in this set, 13 exons from 11 genes were represented in our dataset with at least 100 variants measured in our MPRA. Distributions of variant Δlogit(PSI) are shown, with ClinVar-annotated SNVs highlighted.

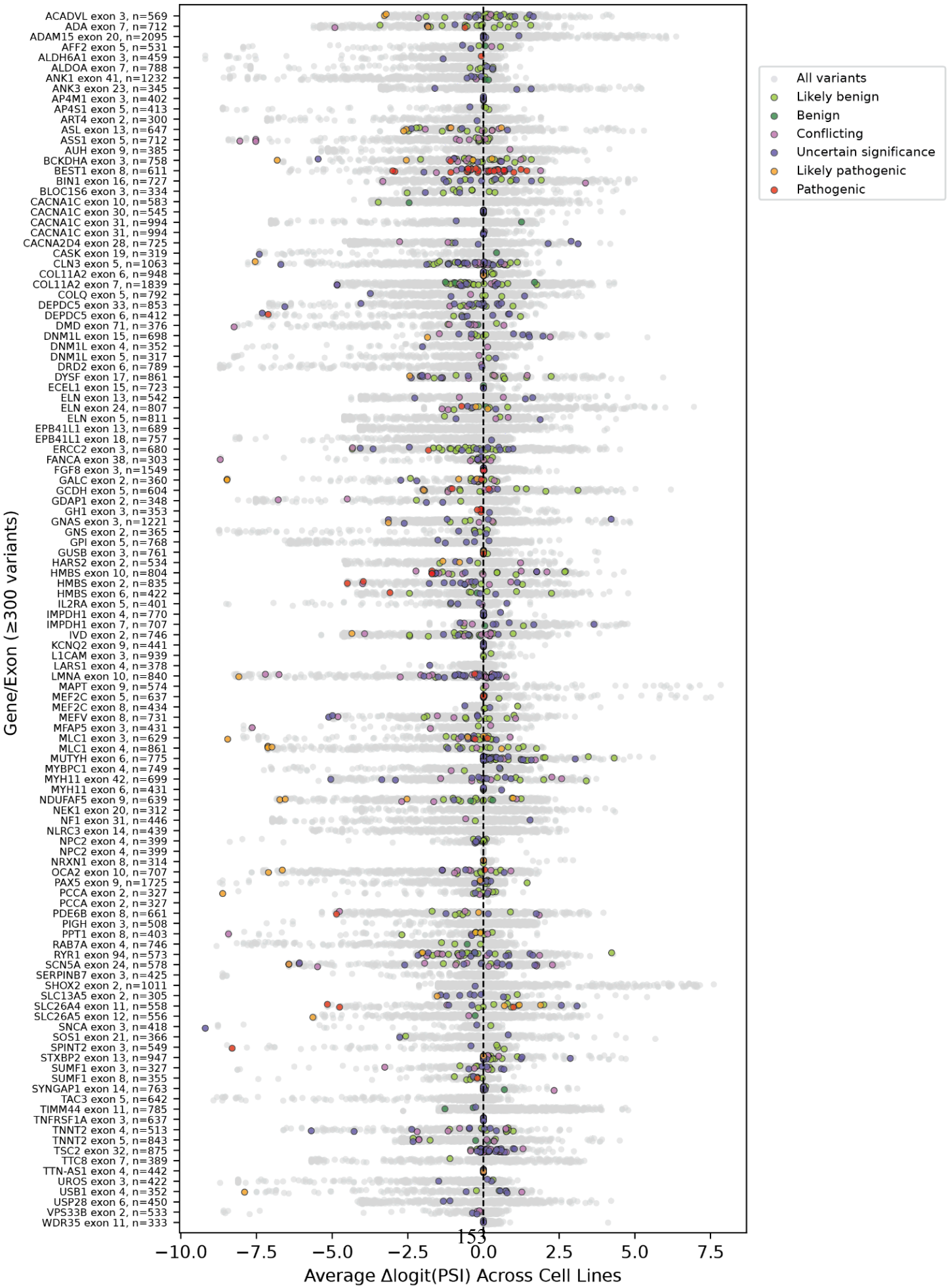
(B) Exons from COSMIC Cancer Gene Census (CGC) Tier 1 genes that were represented in our MPRA with at least 100 measured variants are shown.[107] Δlogit(PSI) distributions are shown with ClinVar annotations indicated, and genes further classified by functional category (tumor suppressor genes, oncogenes, or gene fusions). Tier 1 genes have strong evidence of cancer relevance from both mutational and functional studies.

(C) Exons from six transcription factor (TF) genes, overlapping a catalog of 1,639 known and likely human TFs and their motifs, were represented in our MPRA with at least 100 measured variants. Distributions of variant Δlogit(PSI) values are shown.[109]

(D) Variants from 9 exons belonging to cytokines and cytokine receptors, as defined in the ImmPort Cytokine registry. Δlogit(PSI) distributions are shown. [108]

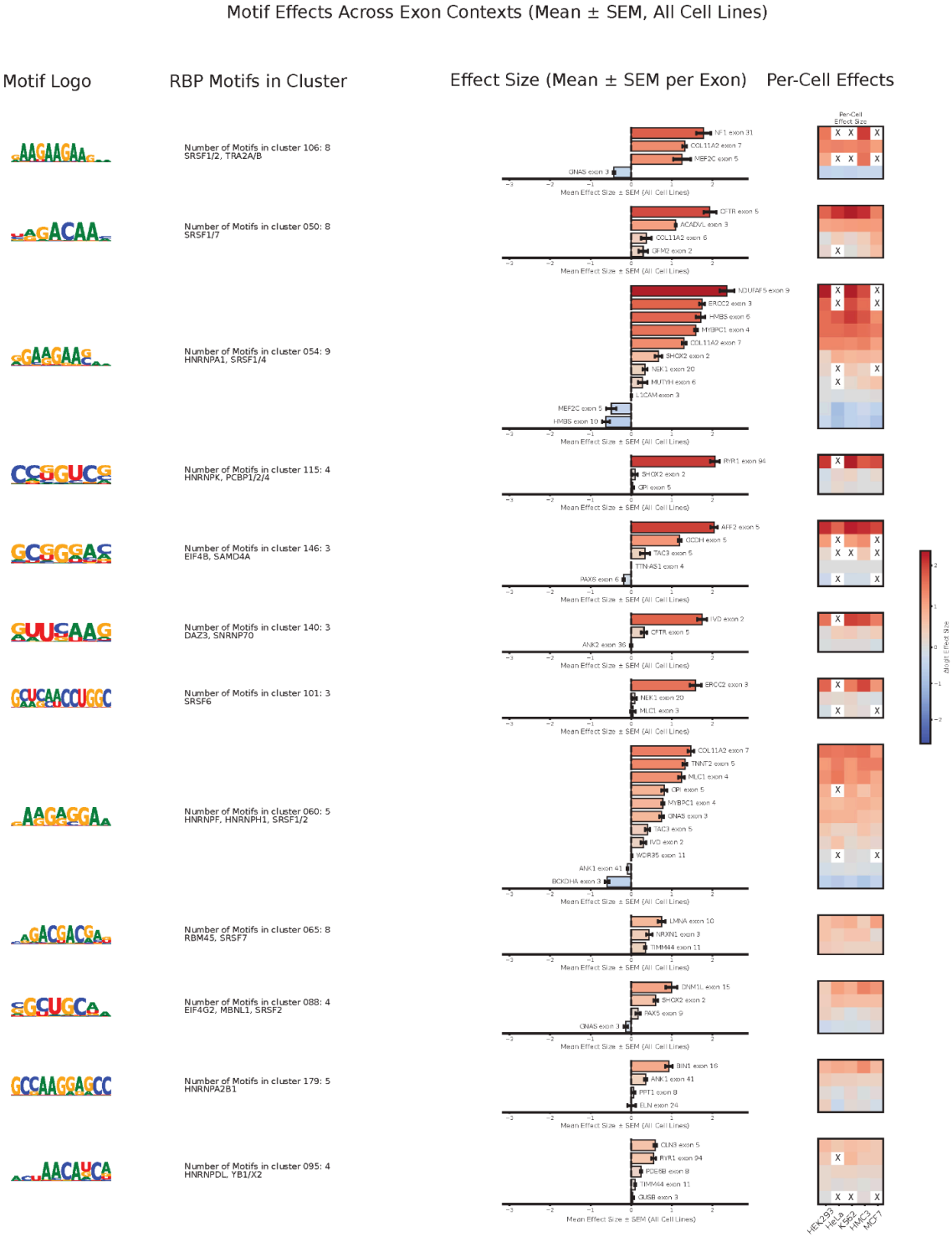


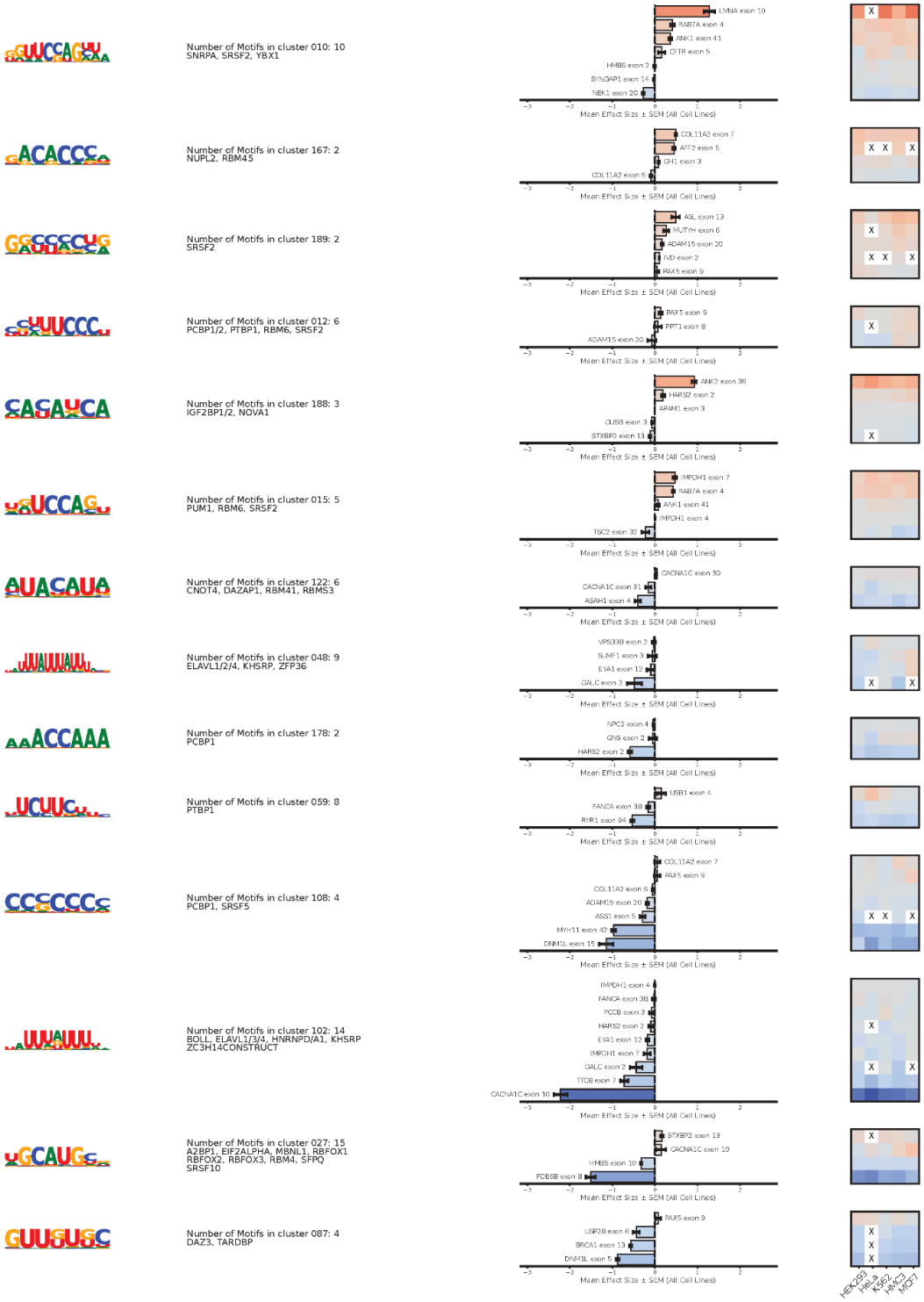
Appendix 1.2 |Δlogit(PSI) distributions for variants in mutational hotspot exons.
Of the exons classified as hotspot exons in the HEK293T, 45 of these exons from 42 genes were represented in our dataset.[110] Distributions of variant Δlogit(PSI) are shown for 12 of these hotspot exons that have at least 100 variants measured in our MPRA. ClinVar-annotated SNVs are highlighted. Hotspot exons are common targets of splicing perturbations, consistent with their enrichment for SDVs in our MPRA.

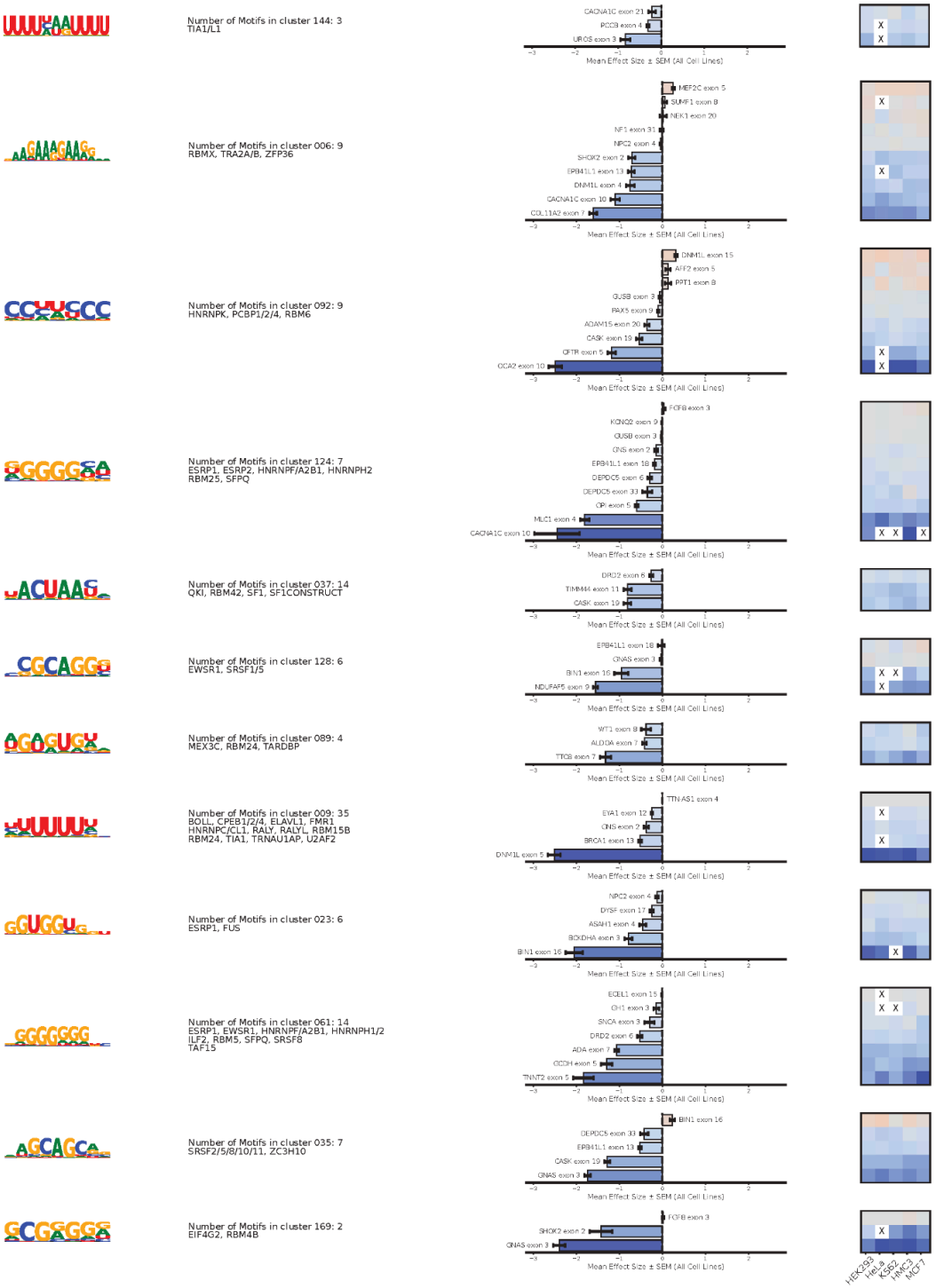

ACADVL exon 3, n=569
ADA exon 7, n=712
ADAM15 exon 20, n=2095
AFF2 exon 5, n=531
ALDH6A1 exon 3, n=459
ALDOA exon 7, n=788
ANK1 exon 41, n=1232
ANK3 exon 23, n=345
AP4M1 exon 3, n=402
AP4S1 exon 5, n=413
ART4 exon 2, n=300
ASL exon 13, n=647
ASS1 exon 5, n=712
AUH exon 9, n=385
BCKDHA exon 3, n=758
BEST1 exon 8, n=611
BIN1 exon 16, n=727
BLOC1S6 exon 3, n=334
CACNA1C exon 10, n=583
CACNA1C exon 30, n=545
CACNA1C exon 31, n=994
CACNA1C exon 31, n=994
CACNA2D4 exon 28, n=725
CASK exon 19, n=319
CLN3 exon 5, n=1063
COL11A2 exon 6, n=948
COL11A2 exon 7, n=1839
COLQ exon 5, n=792
DEPDC5 exon 33, n=853
DEPDC5 exon 6, n=412
DMD exon 71, n=376
DNM1L exon 15, n=698
DNM1L exon 4, n=352
DNM1L exon 5, n=317
DRD2 exon 6, n=789
DYSF exon 17, n=861
ECEL1 exon 15, n=723
ELN exon 13, n=542
ELN exon 24, n=807
ELN exon 5, n=811
EPB41L1 exon 13, n=689
EPB41L1 exon 18, n=757
ERCC2 exon 3, n=680
FANCA exon 38, n=303
FGF8 exon 3, n=1549
GALC exon 2, n=360
GCDH exon 5, n=604
GDAP1 exon 2, n=348
GH1 exon 3, n=353
GNAS exon 3, n=1221
GNS exon 2, n=365
GPI exon 5, n=768
GUSB exon 3, n=761
HARS2 exon 2, n=534
HMBS exon 10, n=804
HMBS exon 2, n=835
HMBS exon 6, n=422
IL2RA exon 5, n=401
IMPDH1 exon 4, n=770
IMPDH1 exon 7, n=707
IVD exon 2, n=746
KCNQ2 exon 9, n=441
L1CAM exon 3, n=939
LARS1 exon 4, n=378
LMNA exon 10, n=840
MAPT exon 9, n=574
MEF2C exon 5, n=637
MEF2C exon 8, n=434
MEFV exon 8, n=731
MFAP5 exon 3, n=431
MLC1 exon 3, n=629
MLC1 exon 4, n=861
MUTYH exon 6, n=775
MYBPC1 exon 4, n=749
MYH11 exon 42, n=699
MYH11 exon 6, n=431
NDUFAF5 exon 9, n=639
NEK1 exon 20, n=312
NF1 exon 31, n=446
NLRC3 exon 14, n=439
NPC2 exon 4, n=399
NPC2 exon 4, n=399
NRXN1 exon 8, n=314
OCA2 exon 10, n=707
PAX5 exon 9, n=1725
PCCA exon 2, n=327
PCCA exon 2, n=327
PDE6B exon 8, n=661
PIGH exon 3, n=508
PPT1 exon 8, n=403
RAB7A exon 4, n=746
RYR1 exon 94, n=573
SCN5A exon 24, n=578
SERPINB7 exon 3, n=425
SHOX2 exon 2, n=1011
SLC13A5 exon 2, n=305
SLC26A4 exon 11, n=558
SLC26A5 exon 12, n=556
SNCA exon 3, n=418
SOS1 exon 21, n=366
SPINT2 exon 3, n=549
STXBP2 exon 13, n=947
SUMF1 exon 3, n=327
SUMF1 exon 8, n=355
SYNGAP1 exon 14, n=763
TAC3 exon 5, n=642
TIMM44 exon 11, n=785
TNFRSF1A exon 3, n=637
TNNT2 exon 4, n=513
TNNT2 exon 5, n=843
TSC2 exon 32, n=875
TTC8 exon 7, n=389
TTN-AS1 exon 4, n=442
UROS exon 3, n=422
USB1 exon 4, n=352
USP28 exon 6, n=450
VPS33B exon 2, n=533
WDR35 exon 11, n=333
Gene/Exon (≥300 variants)
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
Average Δlogit(PSI) Across Cell Lines
All variants
Likely benign
Benign
Conflicting
Uncertain significance
Likely pathogenic
Pathogenic

Appendix 1.3 | Δlogit(PSI) distributions for highly saturated exons.
A subset of 115 exons from the MPRA with high mutational saturation, each represented by at least 300 variants (single and double nucleotide substitutions). This subset illustrates the dense mutational coverage achieved, approaching saturation and providing context into splicing effects when perturbing nearly all of the cis-regulatory elements in each exon.

Appendix 2

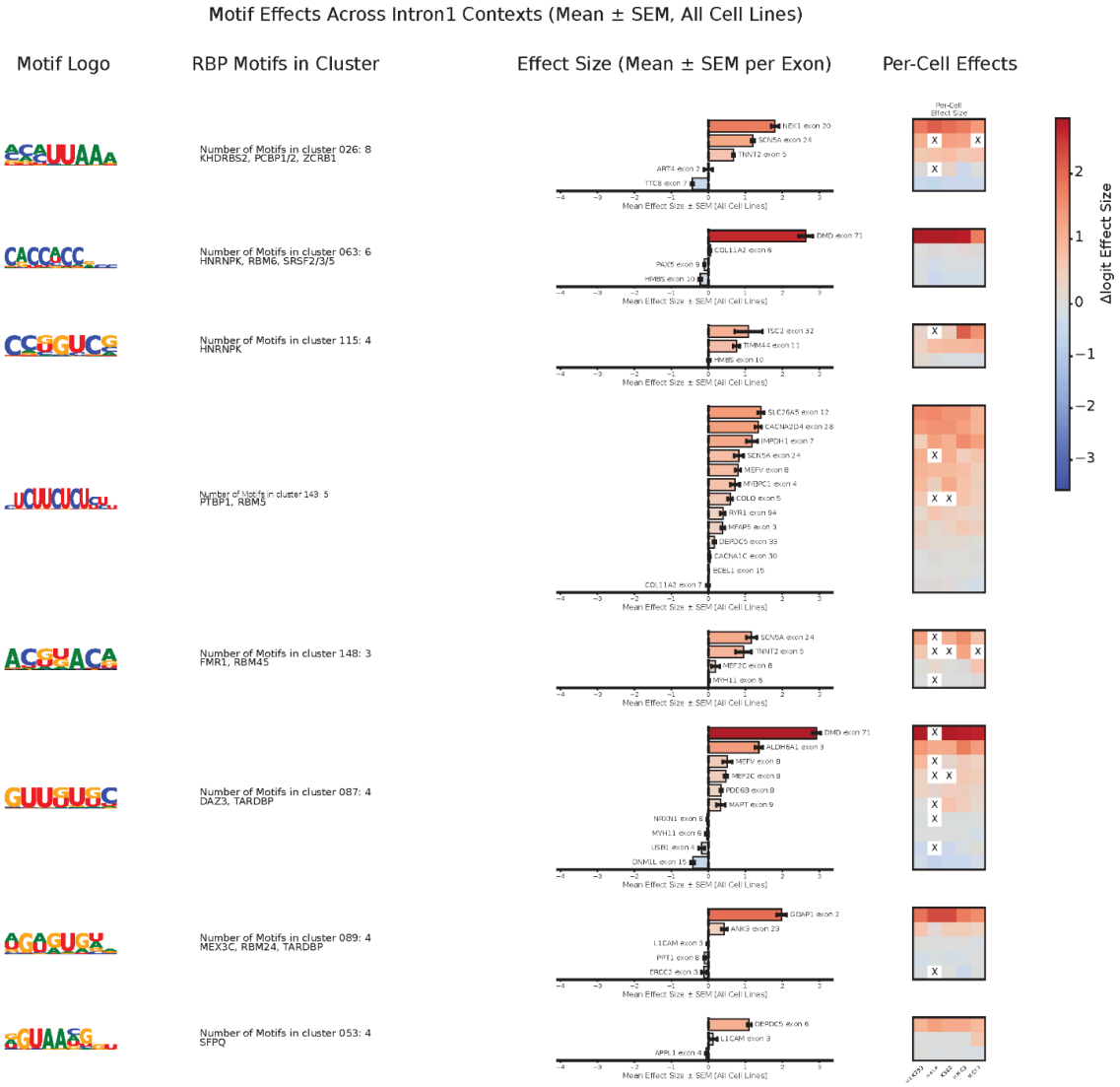

Motif Effects Across Exon Contexts (Mean ± SEM, All Cell Lines)
Motif Logo
RBP Motifs in Cluster
Effect Size (Mean ± SEM per Exon)
Per-Cell Effects
Number of Motifs in cluster 106: 8
SRSF1/2, TRA2A/B
Number of Motifs in cluster 050: 8
SRSF1/7
Number of Motifs in cluster 054: 9
HNRNPA1, SRSF1/4
Number of Motifs in cluster 115: 4
HNRNPK, PCBP1/2/4
Number of Motifs in cluster 146: 3
EIF4B, SAMD4A
Number of Motifs in cluster 140: 3
DAZ3, SNRNP70
Number of Motifs in cluster 101: 3
SRSF6
Number of Motifs in cluster 060: 5
HNRNPF, HNRNPH1, SRSF1/2
Number of Motifs in cluster 065: 8
RBM45, SRSF7
Number of Motifs in cluster 088: 4
EIF4G2, MBNL1, SRSF2
Number of Motifs in cluster 179: 5
HNRNPA2B1
Number of Motifs in cluster 095: 4
HNRNPDL, YB1/X2
Mean Effect Size ± SEM (All Cell Lines)
Per-Cell Effect Size
HEK293
HeLa
K562
HMC3
MCF7
Weight Effect Size


Number of Motifs in cluster 010: 10
SNRPA, SRSF2, YBX1
LMNA exon 10
RAB7A exon 4
ANK1 exon 41
CFTR exon 5
HMBS exon 2
SYNGAP1 exon 14
NEK1 exon 20
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 167: 2
NUPL2, RBM45
COL11A2 exon 7
AFF2 exon 5
GH1 exon 3
COL11A2 exon 6
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 189: 2
SRSF2
ASL exon 13
MUTYH exon 6
ADAM15 exon 20
IVD exon 2
PAX5 exon 9
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 012: 6
PCBP1/2, PTBP1, RBM6, SRSF2
PAX5 exon 9
PPT1 exon 8
ADAM15 exon 20
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 188: 3
IGF2BP1/2, NOVA1
ANK2 exon 36
HARS2 exon 2
AP4M1 exon 3
GUSB exon 3
STXBP2 exon 13
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 015: 5
PUM1, RBM6, SRSF2
IMPDH1 exon 7
RAB7A exon 4
ANK1 exon 41
IMPDH1 exon 4
TSC2 exon 32
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 122: 6
CNOT4, DAZAP1, RBM41, RBMS3
CACNA1C exon 30
CACNA1C exon 31
ASAH1 exon 4
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 048: 9
ELAVL1/2/4, KHSRP, ZFP36
VPS33B exon 2
SUMF1 exon 3
EYA1 exon 12
GALC exon 2
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 178: 2
PCBP1
NPC2 exon 4
GNS exon 2
HARS2 exon 2
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 059: 8
PTBP1
USB1 exon 4
FANCA exon 38
RYR1 exon 94
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 108: 4
PCBP1, SRSF5
COL11A2 exon 7
PAX5 exon 9
COL11A2 exon 6
ADAM15 exon 20
ASS1 exon 5
MYH11 exon 42
DNM1L exon 15
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 102: 14
BOLL, ELAVL1/3/4, HNRNPD/A1, KHSRP
ZC3H14CONSTRUCT
IMPDH1 exon 4
FANCA exon 38
PCCB exon 3
HARS2 exon 2
EYA1 exon 12
IMPDH1 exon 7
GALC exon 2
TTC8 exon 7
CACNA1C exon 10
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 027: 15
A2BP1, EIF2ALPHA, MBNL1, RBFOX1
RBFOX2, RBFOX3, RBM4, SFPQ
SRSF10
STXBP2 exon 13
CACNA1C exon 10
HMBS exon 10
PDE6B exon 8
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 087: 4
DAZ3, TARDBP
PAX5 exon 9
USP28 exon 6
BRCA1 exon 13
DNM1L exon 5
Mean Effect Size ± SEM (All Cell Lines)
HEK293
HeLa
K562
HMC3
MCF7

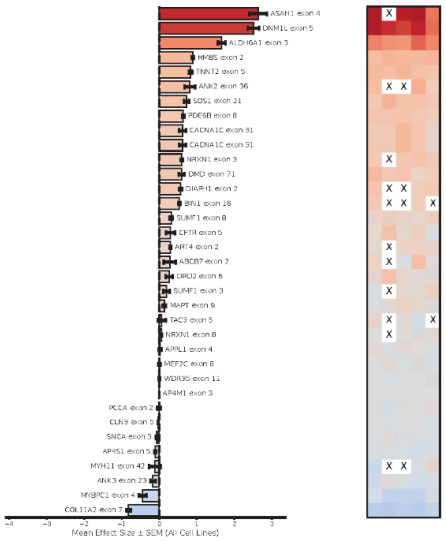

Number of Motifs in cluster 144: 3
TIA1/L1
CACNA1C exon 21
PCCB exon 4
UROS exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 006: 9
RBMX, TRA2A/B, ZFP36
MEF2C exon 5
SUMF1 exon 8
NEK1 exon 20
NF1 exon 35
NPC2 exon 4
SHOK2 exon 2
EPB41L1 exon 13
DNM1L exon 4
CACNA1C exon 10
COL11A2 exon 7
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 092: 9
HNRNPK, PCBP1/2/4, RBM6
DNM1L exon 15
AFF2 exon 5
PPT1 exon 8
GUSB exon 3
RAX5 exon 9
ADAM15 exon 20
CASK exon 19
CFTR exon 5
OCA2 exon 10
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 124: 7
ESRP1, ESRP2, HNRNPF/A2B1, HNRNPH2
RBM25, SFPQ
FOPB exon 3
KCNQ2 exon 9
GUSB exon 3
GNS exon 2
EPB41L1 exon 18
DEPDC5 exon 6
DEPDC5 exon 33
CPI exon 5
MLC1 exon 4
CACNA1C exon 10
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 037: 14
QKI, RBM42, SF1, SF1CONSTRUCT
DRD2 exon 6
TIMM44 exon 11
CASK exon 19
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 128: 6
EWSR1, SRSF1/5
EPB41L1 exon 18
GNAS exon 3
BIN1 exon 16
NDUFAF5 exon 9
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 089: 4
MEX3C, RBM24, TARDBP
WT1 exon 8
ALDOA exon 7
TTC8 exon 7
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 009: 35
BOLL, CPEB1/2/4, ELAVL1, FMR1
HNRNPC/CL1, RALY, RALYL, RBM15B
RBM24, TIA1, TRNAU1AP, U2AF2
TTN-AS1 exon 4
EYA1 exon 12
GNS exon 2
BRCA1 exon 13
DNM1L exon 5
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 023: 6
ESRP1, FUS
NPC2 exon 4
DYSF exon 17
ASAH1 exon 4
BOKDHA exon 3
BIN1 exon 16
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 061: 14
ESRP1, EWSR1, HNRNPF/A2B1, HNRNPH1/2
ILF2, RBM5, SFPQ, SRSF8
TAF15
EGEL1 exon 15
GH1 exon 3
SNCA exon 3
DRD2 exon 6
ADA exon 7
CCDH exon 5
TNNT2 exon 5
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 035: 7
SRSF2/5/8/10/11, ZC3H10
BIN1 exon 16
DEPDC5 exon 33
EPB41L1 exon 13
CASK exon 19
GNAS exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 169: 2
EIF4G2, RBM4B
FOPB exon 3
SHOK2 exon 2
GNAS exon 3
Mean Effect Size ± SEM (All Cell Lines)
HEK293
HeLa
K562
HMC3
MCF7

Appendix 2.1 | Concordant RBP motifs effects in exonic regions.
Motif clusters with consistent effects across exon families are shown. Exonic splicing enhancers (ESEs) promote exon inclusion (red), whereas exonic splicing silencers (ESSs) promote exon skipping (blue). The heatmap displays effect sizes for each exon family across individual cell lines, and the accompanying bar plots show mean effect sizes across all cell lines with error bars representing the SEM.

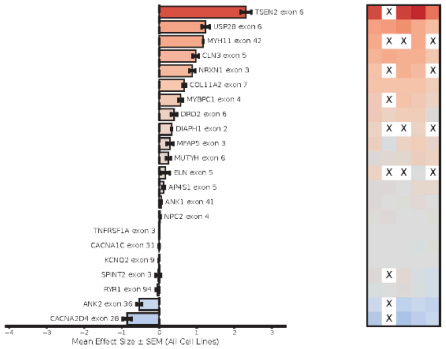
Appendix 3


Motif Effects Across Intron1 Contexts (Mean ± SEM, All Cell Lines)
Motif Logo
RBP Motifs in Cluster
Effect Size (Mean ± SEM per Exon)
Per-Cell Effects
Number of Motifs in cluster 026: 8
KHDRBS2, PCBP1/2, ZCRB1
Number of Motifs in cluster 063: 6
HNRNPK, RBM6, SRSF2/3/5
Number of Motifs in cluster 115: 4
HNRNPK
Number of Motifs in cluster 143: 5
PTBP1, RBM5
Number of Motifs in cluster 148: 3
FMR1, RBM45
Number of Motifs in cluster 087: 4
DAZ3, TARDBP
Number of Motifs in cluster 089: 4
MEX3C, RBM24, TARDBP
Number of Motifs in cluster 053: 4
SFPQ
Mean Effect Size ± SEM (All Cell Lines)
Per-Cell Effect Size
Δlogit Effect Size
2
1
0
−1
−2
−3

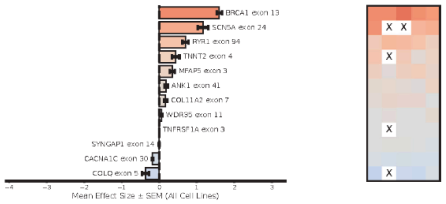
Number of Motifs in cluster 009: 35
BOLL, CPEB1/2/4, ELAVL1, FMR1
HNRNPC/CL1, RALY, RALYL, RBM15B
RBM24, TIA1/L1, TRNAU1AP, U2AF2



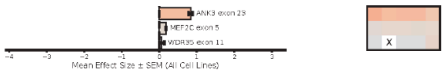
Number of Motifs in cluster 094: 4
PTBP1/3



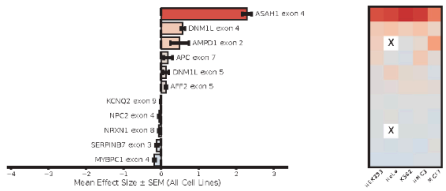
Number of Motifs in cluster 183: 2
PTBP1, TIA1

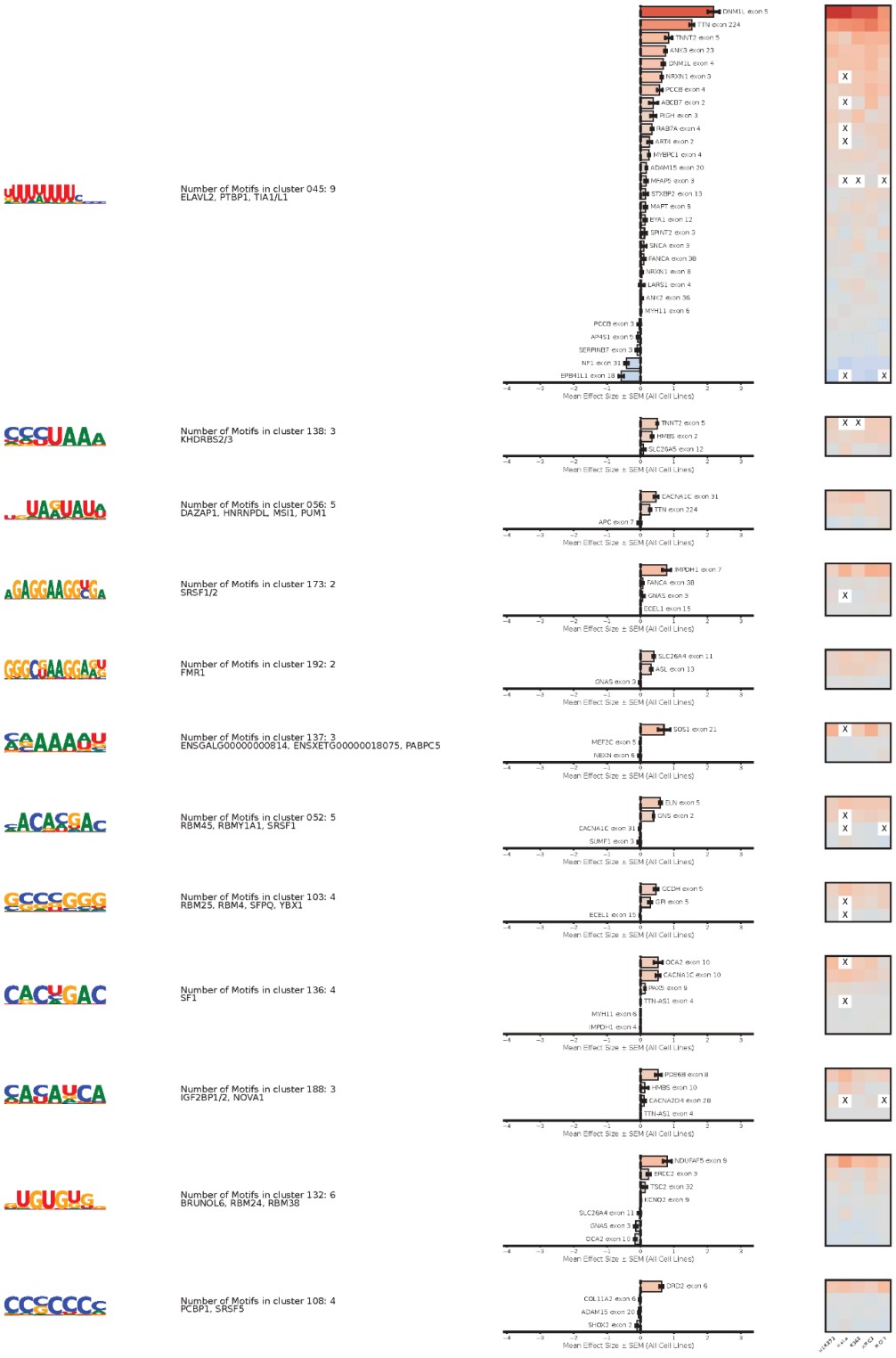


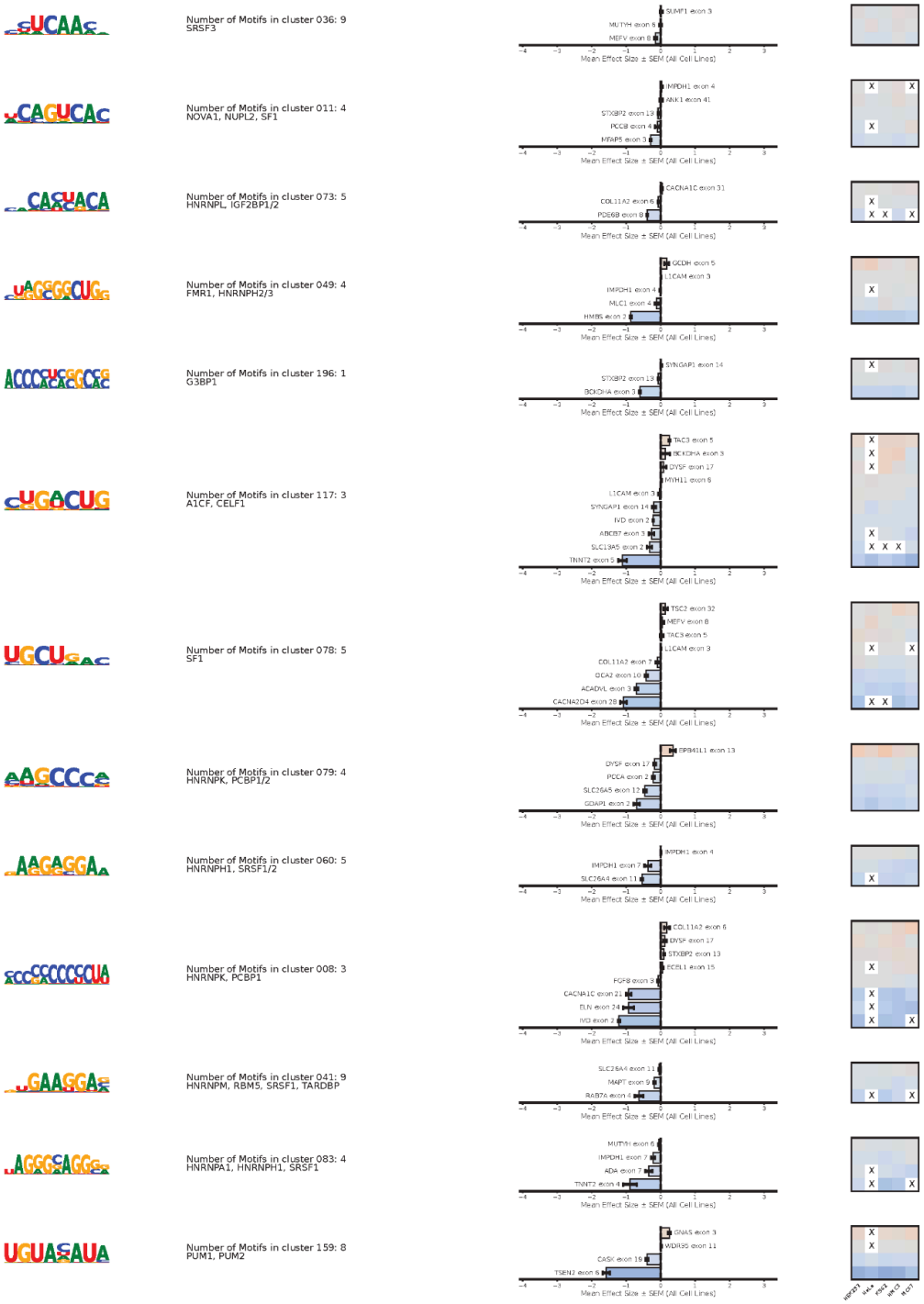
Number of Motifs in cluster 091: 4
FUBP1, KHSRP



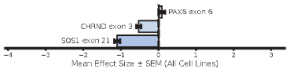
Number of Motifs in cluster 144: 3
TIA1/L1



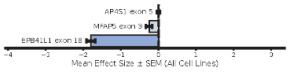

Number of Motifs in cluster 045: 9
ELAVL2, PTBP1, TIA1/L1
DNM1L exon 5
TTN exon 224
TNNT2 exon 5
ANK3 exon 23
DNM1L exon 4
NRXN1 exon 3
PCCB exon 4
ABCB7 exon 2
PIGH exon 3
RAB7A exon 4
ART4 exon 2
MYBPC1 exon 4
ADAM15 exon 20
MPAP5 exon 3
STXBP2 exon 13
MaPT exon 9
EYA1 exon 12
SPINT2 exon 3
SNCA exon 3
FANCA exon 38
NRXN1 exon 8
LARS1 exon 4
ANK2 exon 36
MYH11 exon 6
PCCB exon 3
AP4S1 exon 5
SERPINB7 exon 3
NF1 exon 31
EPB41L1 exon 18
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 138: 3
KHDRBS2/3
TNNT2 exon 5
HMBS exon 2
SLC26A5 exon 12
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 056: 5
DAZAP1, HNRNPDL, MSI1, PUM1
CACNA1C exon 31
TTN exon 224
APC exon 7
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 173: 2
SRSF1/2
MPDH1 exon 7
FANCA exon 38
GNAS exon 3
ECEL1 exon 15
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 192: 2
FMR1
SLC26A4 exon 11
ASL exon 13
GNAS exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 137: 3
ENSGALG00000000814, ENSXETG00000018075, PABPC5
SOS1 exon 21
MEF2C exon 5
NBEN exon 6
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 052: 5
RBM45, RBMY1A1, SRSF1
ELN exon 5
GNS exon 2
CACNA1C exon 31
SUMF1 exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 103: 4
RBM25, RBM4, SFPQ, YBX1
GCDH exon 5
GPI exon 5
ECEL1 exon 15
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 136: 4
SF1
OCA2 exon 10
CACNA1C exon 10
PAX5 exon 9
TTN-AS1 exon 4
MYH11 exon 6
IMPDH1 exon 4
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 188: 3
IGF2BP1/2, NOVA1
PDE8B exon 8
HMBS exon 10
CACNA2D4 exon 28
TTN-AS1 exon 4
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 132: 6
BRUNOL6, RBM24, RBM38
NDUFAF5 exon 9
BRCC2 exon 3
TSC2 exon 32
KCNQ2 exon 9
SLC26A4 exon 11
GNAS exon 3
OCA2 exon 10
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 108: 4
PCBP1, SRSF5
DRD2 exon 6
COL11A2 exon 6
ADAM15 exon 20
SHOX2 exon 2
Mean Effect Size ± SEM (All Cell Lines)


Number of Motifs in cluster 036: 9
SRSF3
SUMF1 exon 3
MUTYH exon 8
MEFV exon 8
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 011: 4
NOVA1, NUPL2, SF1
IMPDH1 exon 4
ANK1 exon 41
STXBP2 exon 13
PCCB exon 4
MFAP5 exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 073: 5
HNRNPL, IGF2BP1/2
CACNA1C exon 31
COL11A2 exon 6
PDE6B exon 8
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 049: 4
FMR1, HNRNPH2/3
GCDH exon 5
L1CAM exon 3
IMPDH1 exon 4
MLC1 exon 4
HMBS exon 2
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 196: 1
G3BP1
SYNGAP1 exon 14
STXBP2 exon 13
BCKDHA exon 3
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 117: 3
A1CF, CELF1
TAC3 exon 5
BCKDHA exon 3
DYSF exon 17
MYH11 exon 6
L1CAM exon 3
SYNGAP1 exon 14
IVD exon 2
ABCB7 exon 3
SLC19A5 exon 2
TNNT2 exon 5
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 078: 5
SF1
TSC2 exon 32
MEFV exon 8
TAC3 exon 5
L1CAM exon 3
COL11A2 exon 7
OCA2 exon 10
ACADVL exon 3
CACNA2D4 exon 28
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 079: 4
HNRNPK, PCBP1/2
EPB41L1 exon 13
DYSF exon 17
PCCA exon 2
SLC26A5 exon 12
GDAP1 exon 2
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 060: 5
HNRNPH1, SRSF1/2
IMPDH1 exon 4
IMPDH1 exon 7
SLC26A4 exon 11
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 008: 3
HNRNPK, PCBP1
COL11A2 exon 6
DYSF exon 17
STXBP2 exon 13
PCEL1 exon 15
FGF8 exon 3
CACNA1C exon 21
ELN exon 24
IVD exon 2
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 041: 9
HNRNPM, RBM5, SRSF1, TARDBP
SLC26A4 exon 11
MAPT exon 9
RAB7A exon 4
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 083: 4
HNRNPA1, HNRNPH1, SRSF1
MUTYH exon 8
IMPDH1 exon 7
ADA exon 7
TNNT2 exon 4
Mean Effect Size ± SEM (All Cell Lines)
Number of Motifs in cluster 159: 8
PUM1, PUM2
GNAS exon 3
WDR35 exon 11
CASK exon 19
TSEN2 exon 6
Mean Effect Size ± SEM (All Cell Lines)
HEPG2
K562
SK-N-SH
A549

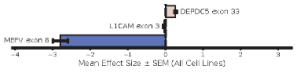
Number of Motifs in cluster 075: 9
AGO2, IGHMBP2, PABPC1/C4/N1/N1L, SART3



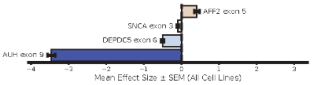
Number of Motifs in cluster 029: 7
RBM28, SNRPA, SRSF2



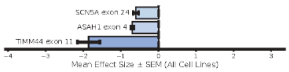
Number of Motifs in cluster 088: 4
EIF4G2, MBNL1, SRSF2



Number of Motifs in cluster 070: 6
A1CF, DAZAP1, ELAVL4, HNRNPD
TRNAU1AP



Number of Motifs in cluster 118: 2
MATR3, RC3H1



Appendix 3.1 | Concordant RBP motifs in intronic regions.
Motif clusters found in the 5’ intron with consistent effects across exon families are shown. Intronic splicing enhancers (ISEs) promote exon inclusion (red), whereas Intronic splicing silencers (ISSs) promote exon skipping (blue). The heatmap displays effect sizes for each exon family across individual cell lines, and the accompanying bar plots show mean effect sizes across all cell lines with error bars representing the SEM.