Massively Parallel Screens to Identify Splice Disruptive Variants in Human Disease Genes

by

Cathy Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Jacob O. Kitzman, Chair
Professor Sally A. Camper
Associate Professor Hui Jiang
Professor Jeffrey M. Kidd
Professor Maureen A. Sartor
Professor Nils G. Walter

Catherine S. Smith

Cathy Smith

smithcat@umich.edu

ORCID iD: 0000-0003-1518-9320

Dedication

This dissertation is dedicated to those who have gone before me and helped pave the way for me to be where I am and who I am today especially Fluffy Smithmas, Sebastian Smokovitz, Steve and Lillian (Mazur) Wodkowski, and Charles and Barbara (Fellows) Smith.

Acknowledgements

It definitely took a village to get me to where I am today and put me in a place to complete this dissertation. I will forever be grateful to Rowell Huesmann and Eric Dubow who both took a big risk on hiring me after my undergraduate floundering. Their guidance, mentorship, and support got me through my master's level work and led me on this career in research. Beverly Strassmann's passion for science, love of debunking logical and biological fallacies within research, determination, and stubbornness in the face of adversity was also a major source of inspiration for me in the beginning of my career in research. She took the time to get to know me personally and offered blunt and invaluable advice which led me to pursue my Ph.D. The computational statistics courses taught by Hyun Min Kang during my master's in biostatistics inspired me to pursue a bioinformatics Ph.D., and under his suggestion, I met my Ph.D. mentor Jacob Kitzman. Jacob has patiently stood by as I've wrestled with proper genomics terminology, memory management on the server, and simple math mistakes. I think we've made a great team and his passion for strong supporting examples, careful data analysis, and vivid data visualizations has led a lot of my growth during my Ph.D. work. He's also been there for life advice and has been a strong advocate for me throughout our time together. I'd also like to thank Tony Antonellis and Miriam Meisler along with both of their labs for their helpful feedback on my projects during joint lab meetings. Finally, I want to thank my dissertation committee for their contributions to my research

and collaboration – and a special thanks to Sally Camper whose guidance, encouraging words, and fruitful collaboration have been a highlight of my Ph.D. research.

I have been encouraged and supported by all of the past and present members of the Kitzman lab: Bala Burugula, Isaac Jia, Sajini Jacoby, Miriam Maksutova, Victor Chen, Anthony Scott, Shelby Hemker, Seba Vishnopolska, Adelaide Tovar, Lena Glick, Grace Clark, Veronica Glaser, and Jennifer Lai Yee. I knew the lab would be a great fit when Bala invited me to lunch on the first day of my rotation, and I will miss the laughs, stories, and gossip at lunchtime as I transition to my new job. Thank you to Margit Burmeister and Maureen Sartor for the energy and care they put into advising bioinformatics students.

Our research would not be possible without the hard work of the staff and administration teams. Starting with Mary Cullen Guttman, Diana Armistead, Debbie Bourque, and Laura Reynolds at the ISR who always invited me to lunch and to the fabled admin parties during St. Patrick's Day and before the winter holidays. Your wit, conversation, life guidance, and insight into the inner-workings of research within UofM has been invaluable for my growth both personally and professionally. Thank you to all of the staff and administrators who helped with all of the critical details of my graduate career notably Julia Eussen, Kati Ellis, Aaron Bookvich, Mary Freer, Theresa Nester, Jeff Holden, Molly Martin, and Dhammika Dewasurendra.

I was extremely nervous about feeling isolated during my tenure in the biostatistics department for my master's especially after my experiences in undergrad. But those fears were completely unfounded as I found a great group of friends who helped me with homework, went out about the town, and refreshed my outlook on

friendships in adulthood. My chili cook-off champions: Holly Hartman, Michelle McNulty, and Jenny Nguyen. My friends for sporting events, dinners at Revolver, late night homework help, and nights out on the town: Christina Zhou, Steve Salerno, Dan Molling, Emily Roberts, Abhay Hukku, Jon Boss, Holly Hartman, Michelle McNulty, Jenny Nguyen, Brett Morris, Chris Fitzpatrick, and Sarah Hanks. And my fellow bunny lover Amy Lasher.

The social connections continued into my Ph.D. and I'm especially proud of the community that I built in BGSA with Ford Hannum. We put a lot of energy into creating a space to socialize and unite the student body, and I definitely could not have done it without Ford. The guidance I got from students in older cohorts including Zena Lapp, Rucheng Diao, Brooke Wolford, Peter Orchard, Marlena Duda, Kevin Hu, and Chris Castro were invaluable in navigating the department and program. I also appreciated the support and deep conversations with Kelly Sovacool and Ali Farhat and working on Matchathon with Zena Lapp, Scott Barolo, Haley Amemia, and Maggie Durdan. Finally, I have to thank the squad who was always down for last minute pop up food, helping me host social events, sporting events, and whatever else we could think of doing: Brodie Mumphrey, Ford Hannum, Kevin Yang, Stuart Castaneda, Brad Crone, and honorary bioinformatics member Ben Hillebrand.

I could not have completed my Ph.D. without my friends who were just a text away to send memes, laugh about internet hijinks, complain about life, and cheer each other on. The pandemic pod of Holly Hartman, Nicky Wakim, Erin Loomas, and Meg Banker who were one of the few we socialized with in person during early COVID has been one of the best group chats, friends for spontaneous hangs and lunches, and a

source of endless excursions as a relief to the stress of life and research. My older and most enduring friendships with Sam Primeau, Lauren Boumaroun, and Ryan Smokovitz have filled my undergraduate and graduate school years and life with sarcasm, quick wit, emotional support, shade, and side splitting comic relief. The three of them have really been there through it all and I treasure my connections with them – they are some of the most genuine and hilarious people. And my newer local friends outside of academia have been a welcome break from the atmosphere of academia: Vanoo Gant, Kate McNamara, Erin Jones, Sekai Ward, Heather Robert, Sarah Watkins, Martha Siegmund, and Maggie Reuter.

My parents have been a huge support and, at times incessantly, encouraging of my education. My mom, Carol (Wodkowski) Smith, was one of the first female engineers at Ford Motor Company and grew up in a time when she was the only woman in her chemistry classes. There was always an expectation, especially at the large Polish gatherings, that I would excel in math and science and become an engineer. Despite my protests, I'm glad that she didn't let me quit on my math degree in undergrad no matter how arduous it was at times, as that math degree has opened doors for me that a degree in general psychology may not have. My dad, Dave Smith, has always been there to support me, help me move around Ann Arbor seemingly a million times, clean my apartment, and go out to spontaneous lunches. They have both been a huge financial support – in gas money, rent, dinners out, cooking classes, vacations, and kielbasa - throughout my graduate career and I could not have done it without them – although at times perhaps there could have been less nagging.

My wife, Erin Loomas, has been my biggest cheerleader during this process. We made a decision for me to pursue this career move for our advancement as a family, and she alone knows the full breadth of the peaks, valleys, doubts, triumphs, and difficult decisions along the way. She has supported me emotionally, helped me hydrate, debated pros and cons, and all along the way has reminded me that this is the path I am supposed to purse and that I can succeed. This degree is truly shared between us, and I look forward to experiencing the fruits of our labor within my career path. Her level of steady emotional support is rivaled only by that of Fluffy and Sebastian. Although they didn't make it to see the end of my journey to earn this degree, they and Wawie provided a wealth of cuddles and were excellent administrative assistants as I worked from home.

Thank you to everyone – proud to be this much closer to Dr. Smith!

Table of Contents

List of Tables

List of Figures

Abstract

Splicing is a critical step in mRNA maturation with roles in gene regulation and proteome diversification. Splice disruptive variants (SDVs) are implicated in diverse human diseases, and 10-33% of exonic variants may disrupt splicing. However, identifying SDVs remains challenging due to the degeneracy and redundancy of the underlying sequence code. Experimental splicing measurements from patient cells or mini-gene assays can detect SDVs but have traditionally been low-throughput.

Massively parallel splicing assays (MPSAs) systematically measure splicing impacts at scale and could clarify variant pathogenicity and inform models of splicing regulation. In this assay, complex barcoded libraries of mutant exons are synthesized, cloned into minigene constructs, and transfected into human cells. Splicing outcomes of each mutation are quantified en masse using targeted RNA-seq of minigene-derived transcripts, and analyzed with a custom python package.

In Chapter 2, I apply this assay to the pituitary transcription factor gene *POU1F1* (in collaboration Dr. Sally Camper's lab). Mutations in *POU1F1* cause combined pituitary hormone deficiency (CPHD), a clinically and genetically heterogenous disorder with prevalence ~1:4000. We targeted exon 2, which has two alterative isoforms (alpha and beta) using competing splice acceptors that encode mutually antagonistic proteins. We measured the splicing effects of 1,070 SNVs across the exon and surrounding introns and identified 96 SDVs - 14 of which were synonymous substitutions. Our measurements were concordant with six nearby heterozygous missense and

synonymous variants seen in unrelated hypopituitarism patients. This map identifies a putative splice silencer motif that represses the use of the normally lowly expressed beta isoform.

In Chapter 3, I apply a MPSA to a critical developmental renal transcription factor gene, *WT1* (in collaboration with clinical nephrologist Dr. Jen Lai Yee). Mutations in *WT1* are implicated in nephrotic syndrome and sexual differentiation phenotypes. I focus on exon 9 which is alternatively spliced at competing donor sites resulting in two isoforms (KTS+ and KTS-). KTS+ and KTS- are normally expressed in ~2:1 ratio, but perturbation of the ratio can lead to Frasier's syndrome – a rare nephrotic syndrome. We tested 518 SNVs for splicing defects and identified 8 known Frasier's Syndrome variants as well as 16 additional variants that similarly lowered the KTS ratio. We also detected 19 variants increasing the KTS ratio, two of which have been observed in patients with sexual differentiation phenotypes.

Although MPSAs can measure splicing effects of hundreds of variants simultaneously, the current scale of variant discovery via exome and genome sequencing demands efficient and accurate computational approaches to identify splice disruptive variants genome-wide. To evaluate the state of the art within contemporary splice prediction algorithms, in Chapter 4 I employed the results of five high throughput splicing assays and one literature curated variant set. A unique advantage of MPSAs over typical training and validation datasets is that they avoid bias towards essential splice site variants. I found the latest deep learning tools, SpliceAI and Pangolin, were most concordant with the measured splicing effects. However, all tools showed less agreement with exonic splicing outcomes compared to intronic. Some tools' predictions,

like SpliceAI's, were sensitive to specified annotation files. Therefore, there is still room for improvement within the next generation of splice prediction algorithms which future MPSA studies may facilitate. Thus, MPSAs are critical to identify clinically relevant SDVs and improve computational splice prediction.

Chapter 1 Introduction

1.1 Molecular biology of splicing

Splicing is a key process during mRNA maturation in eukaryotic cells, in which introns are excised from pre-mRNAs which can then be exported to the cytoplasm for translation (in the case of coding genes). The process of splicing is catalyzed by five small nuclear ribonucleoproteins (snRNPs) each composed of one small nuclear RNA (snRNA) and additional proteins which form a complex called the spliceosome[1-3]. Two different evolutionarily conserved spliceosome complexes exist in eukaryotes, termed the major and minor spliceosomes, each composed of four distinct snRNPs and U5, the one shared snRNP[3]. The major spliceosome directs excision of the large majority of introns (>99%), so called U2-style introns (named for the major spliceosome component; **Figure 1-1**)[3]. The minor spliceosomes excises distinct, U12-style introns which in most cases reside within the same gene as other U2-style ones[3].

Both spliceosomes remove intronic portions of pre-mRNAs through an ordered, multi-step process involving the recognition of conserved motifs, at the 5' (donor) and 3' (acceptor) splice sites, facilitated by components of the spliceosome as well as conserved sequence features in the pre-mRNAs undergoing splicing. Major spliceosome assembly begins when a 5' splice site with the essential sequence motif 5'-GU-3' is recognized via base pairing between those bases in the mRNA and the reverse complementary binding sites in the U1 snRNP[2-4] (**Figure 1-1**). Next, upstream of the 3'

splice site, a conserved branchpoint motif (5'-CURAY-3'; R=A/G; Y=U/C), an intronic pyrimidine rich polypyrimidine tract downstream of the branchpoint, and the essential 3' splice site (5'-AG-3') sequence adjacent to downstream exon are bound by the splicing factors SF1, U2AF65, and U2AF35 respectively, via RNA binding sites which target each of the respective conserved motifs[2-4]. The U2 snRNP then displaces SF1 and binds to the branchpoint with reverse complementary base pairing to recruit a complex of three snRNPs – U4, U5, and U6[2-4]. The U6 snRNP replaces the U1 snRNP at the 5' splice site bringing the branchpoint within the vicinity of the 5' splice site[2-4]. Next with the aid of additional proteins and splicing factors, the branchpoint attacks the proximal 5' splice site forming an intron lariat and excising the upstream exon. Then in a second step, the 5' splice site attacks the 3' splice site which is bound through complementary base pairing to the spliceosome which excises the intronic material from the downstream exon and ligates the exons together[1-4].

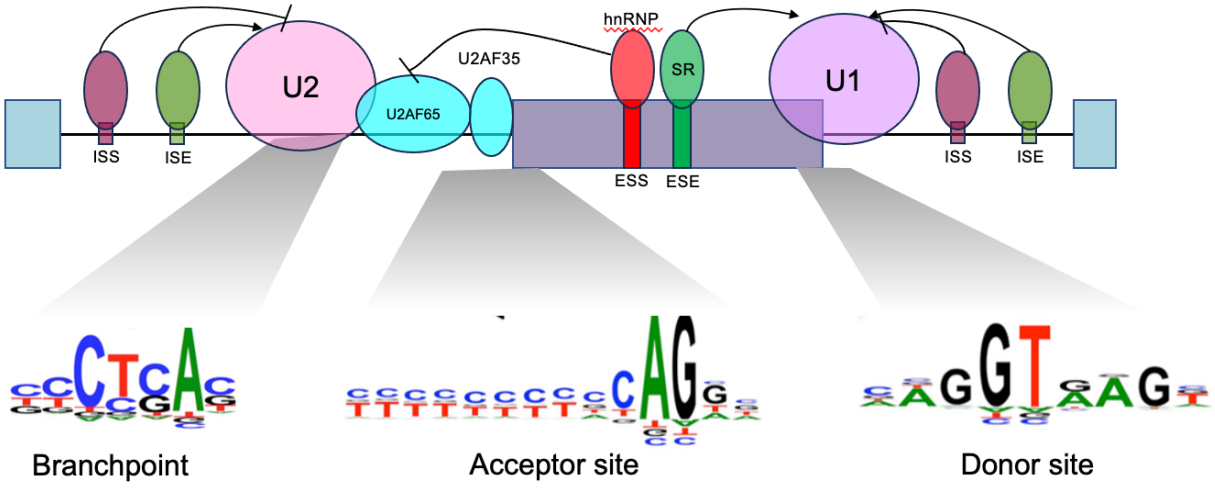


Figure 1-1: Splicing of U2 style introns.

Diagram of components of the major spliceosome along with models of the motifs needed to maintain splicing fidelity. Adapted from Scotti & Swanson, *Nat Rev Genet,* 2016[1].

The U12-dependent minor spliceosome removes a minority (<1%) of introns characterized by highly conserved - but distinct - motifs around the splice sites and at the branchpoint[3]. However, the minor spliceosome does not depend on recognition of the polypyrimidine tract and can remove introns not demarcated by the essential canonical dinucleotides[3]. Although this dissertation will focus on the action of the major spliceosome, disruption of the proteins comprising the minor spliceosome as well as variants altering the cis sequence elements recognized by the minor spliceosome and components of the minor spliceosome itself, are implicated in human developmental and neurological disorders[5,6].

Much of the information content needed to define exons and specify their splicing is provided by sequence motifs at the branchpoint site, polypyrimidine tract, and essential splice site motifs[1]. The first of these, the branchpoint, is a highly conserved, degenerate 5-mer motif (CURAY; R=A/G; Y=U/C) with an obligate adenosine from which a 2'-5' linkage to the first intronic base downstream of the 5' splice site results in a lariat structure[7]. The majority of branchpoints reside ~30 bp upstream of the 3' splice site, and almost all exons have multiple, redundant branchpoints which may be specific to certain tissues or developmental states[8,9]. Characterizing branchpoint sequences has been challenging both computationally and experimentally due to the redundancy and degeneracy of the branchpoint sequence and the constant turnover and low stability of lariat-containing transcripts, but recently high throughput experimentally validated

branchpoints[9,10] have led to more sophisticated computational branchpoint prediction tools, for instance BranchPointer and BPP[11-13].

Downstream of the branchpoint lies the polypyrimidine tract, which does not have a specific motif but is instead cytosine and uracil rich, and is involved in modulating splicing efficiency through branchpoint and 3' splice site selection[14].

Specific sequence motifs are required at the 5' and 3' splice sites, with the most extreme constraint at the first and last two intronic bases, termed the essential or canonical dinucleotides. For U2 type introns, the required sequences are GU and AG, respectively, and mutations in these dinucleotides almost without exception abolishes their usage, leading to outcomes including exon skipping, intron retention, or usage of another nearby splice site. Although not strictly essential, the conserved intronic and exonic sequences proximal to the canonical dinucleotides (<8 bp for introns and <3 bp for exons) also contribute to exon definition.

Since splice sites are in fixed locations bordering internal exons, consensus motifs for the 5' and 3' splice site regions can be defined by tallying nucleotide representation at each of the splice site adjacent positions across the genome and then identifying the most common nucleotide at each site, to produce a position-weight matrix (PWM). Regions more closely resembling the consensus sequence at either site give rise to exons which are more constitutively included in the mature mRNA[2], implying that while these proximal 5' and 3' splice site regions undergo less evolutionary constraint than the canonical dinucleotides, they help maintain splicing fidelity.

Despite the well-defined conserved motifs at the branchpoint and both splice sites, these elements only contain about half of the information necessary to define

authentic exons against the large background of cryptic splice sites present in non-exonic sequences which comprise >95% of the human genome[15]. Thus, it was inferred that additional sequence motifs beyond the splice sites, polypyrimidine tract, and branchpoint motifs must be required to maintain splicing fidelity[16]. One potential mechanism is by the action of RNA binding proteins (RBPs) which recognize and bind short (<10 bases) sequences located within introns or exons to antagonize or promote splicing[2]. Motifs bound by such RBPs are categorized by location and direction of impact on splicing: exonic splice enhancers (ESE), exonic splice silencers (ESS), intronic splice enhancers (ISE), and intronic splice silencers (ISS). Families of RBPs involved in splicing regulation include hnRNPs (heterogeneous ribonucleoprotein particles), SR (serine- and arginine-rich) proteins, or tissue specific splicing factors such as the neuronal NOVA-family proteins and the Rbfox family of factors expressed in brain and skeletal muscle[2]. Although there are notable exceptions, hnRNP family proteins generally have suppressive effects while SR family proteins stimulate splicing[2]. The regulatory motifs bound by the RBPs are thought to act as a redundant mechanism to maintain and hone splicing efficiency and are found at higher density in and around constitutively included exons[17-19].

As one example, an exonic splice silencer (ESS) helps to guide proper splicing of exon 11 of the oncogene *RON*. The putative ESS near the exon 11 splice acceptor was initially identified due to sequence homology to another ESS found in the oncogene *MADD*[20]. The implicated sequence (5'- AT<u>TGGG</u>C<u>TGGG</u>C-3') contains two proximal, repeated motifs matching the preferred binding site of hnRNPH (underlined) which, like most hnRNP family proteins, tends to repress splicing[20]. A mini-gene assay was

implemented to test the potential splice silencing effect of the sequence. The assay inserts the native sequence of the targeted exons along with their internal introns into a plasmid which was then transfected into human cells. Total RNA is collected and analyzed with RT-PCR to detect alterations in the splicing pattern. At baseline, exon 11 is included at ~50%, both in the endogenous mRNA, and when tested via transfected minigenes. Knockdown of hnRNPH by co-transfecting with small interfering RNA (siRNA) led to near-complete exon 11 inclusion, and conversely, over-expression of hnRNPH promoted skipping of exon 11[20]. Mutating two sites within the putative hnRNPH motifs (5'- ATTG**A**GCTG**A**GC-3'; variants bolded) either separately or simultaneously encouraged exon 11 inclusion, and the effects were most pronounced when both motifs were altered[20]. Similarly, when both motifs were abolished, both the knockdown and over-expression of hnRNPH had no impact on splicing[20]. Since exon 11 encodes the transmembrane domain of *RON,* the skipping of exon 11 leads to localization to the cytoplasm and homodimerization resulting in a constitutively active receptor implicated in promoting epithelial-mesenchymal transition and tumor invasiveness[21,22].

Dissecting the effects of an individual RNA binding site on an exon's proper splicing remains a challenge. RBPs cannot be mapped 1:1 onto splicing regulatory motifs as multiple expressed RBPs often share very similar sequence specificities as defined by biochemical methods such as SELEX or via occupancy assays such as CLIP-seq. In addition, the motifs bound by RBPs tend to be short and degenerate making it difficult to map precise binding sites. The grammar – the number and arrangement of such sites required to support splicing – is often unclear. Finally, exonic

sequences evolve under additional constraints required to encode functional protein, making it challenging to identify functional RBP sites by sequence conservation across evolution[23-25].

Since many regions harbor multiple potential regulatory motifs, distinguishing essential motifs from redundant binding regions has been difficult[19,24]. Furthermore, a motif's expected impact on splicing may be location dependent – the same regulatory element could enhance splicing if positioned within the exon, but silence splicing from an intronic region[17,24,26]. Attempts have been made to comprehensively identify and list potential splicing regulatory motifs both experimentally and computationally[17,24,25,27-29], but enumerating essential regulatory motifs outside of splice sites sufficient to maintain splicing has been elusive and many of the resulting sets of putative regulatory elements show little overlap[30]. So, although splicing regulatory sequences outside of splice sites have been thoroughly investigated, predicting the impact of sequence variation in these regions and identifying which trans acting RBP factors may be implicated remains challenging.

1.2 Alternative splicing and dysregulation of splicing in disease

Splicing goes beyond simply processing pre-mRNAs, and also serves to diversify the proteome, providing tissue specific gene expression, and a layer of gene regulation in development. Almost every human gene undergoes some type of alternative splicing event: whole exon skipping, alternative splice site use to truncate or extend existing exons, inclusion of pseudo-exons, or whole intron retention (**Figure 1-2**). The combinatorial effect of these events is that a single gene locus may encode multiple isoforms which translate into distinct proteins[18,31] and have untranslated regions with

different regulatory sequence context. The proportion of alternatively spliced genes scales with organism complexity[32] and neuronal genes in particular express different isoforms throughout development[33] implicating alternative splicing as a potential important functional substrate for evolution. Some alternative splicing events shift the frame of the transcript, potentially leading to premature truncation, nonsense mediated decay, and consequently, changes in gene dosage. In other cases, splicing changes encode stable proteins that have dominant negative activity or toxic gain of function[1].

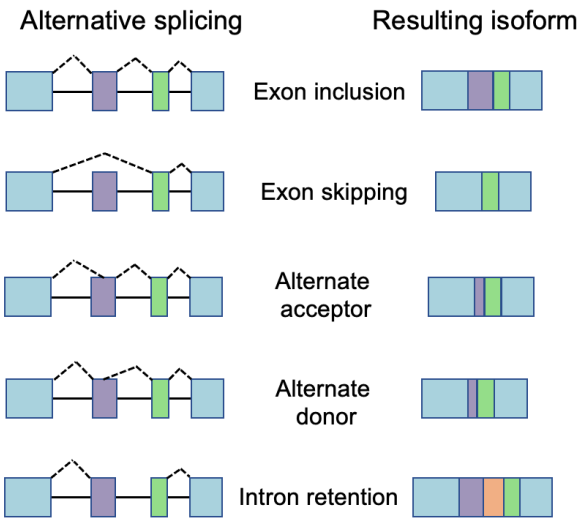


Figure 1-2: Possible outcomes of alternative splicing.

Examples of alternative splicing of the purple exon and the resulting isoforms. Adapted from Keren, Lev-Maor. & Ast, *Nat Rev Genet* 2010[18].

Proper mRNA splicing is disrupted through multiple mechanisms in disease. Mutations to the genes that encode components of the splicing machinery may cause coordinated changes in exon definition and splice site usage. For instance, in myelodysplastic syndromes, somatic mutations to the core spliceosomal factor *SF3B1* are a common driving event[34]. In another example, retinitis pigmentosa is a disorder characterized by gradual vision loss and blindness[35] with a variety of underlying loci and

modes of inheritance. One of its autosomal dominant forms is caused by missense variants in *SNRP200,* which encodes an RNA helicase responsible for splice site recognition as part of the spliceosome complex[35]. The resulting mutant spliceosome preferentially selects cryptic splice sites instead of the native site implying that the variants alter the proofreading ability of the spliceosome[35].

What is more common, and the focus of this dissertation, are individual variants which disrupt the proper splicing of the single gene in which they reside. Such variants may disrupt one of the authentic splice sites, or newly create a splice site in a disruptive context. They may occur in the conserved acceptor or donor sequences, or in RBP-bound splicing regulatory elements, or can act by yet other mechanisms (e.g., by altering RNA secondary structure to perturb interactions with the spliceosome)[1].

Although the true proportion of variants which cause Mendelian disorders by altering splicing is unknown, recent estimates implicate 10-30% of SNVs causing human disease as disrupting splicing[36,37]. Variants that exert a pathogenic effect primarily via splicing disruption are expected to conform to the inverse relationship between risk allele frequency and effect size – so, splice disruptive variants with a large impact on Mendelian disease phenotypes would be rare in the population due to purifying selection[38].

Variants altering splicing in disease can abolish splicing regulatory motifs or create decoy sites resulting in mis-spliced mRNA. An intronic variant creating a cryptic 3’ splice site in *HBB* was one of the earliest SNVs linked to disease through altered splicing[39]. In the presence of this mutation, the spliceosome selects the cryptic acceptor site, extending the exon by 19 bp, and resulting in a frameshifted protein leading to

degradation of the mRNA by nonsense mediated decay[39,40]. The reduction in beta-globin expression causes haploinsufficient ß+ thalassemia – a disease linked to anemia and in many cases requiring lifelong blood transfusions[40]. Intronic variants can also cause disease by altering the essential dinucleotides which is seen in Duchenne's muscular dystrophy (DMD) – an X-linked disorder typified by rapid, progressing muscle loss in childhood[41]. Several point variants at essential splice sites have been identified across *DMD* with more severe clinical presentations occurring when the exon skipped is not a multiple of three bases in length. Such disruption events cause frameshift and premature truncation rather than in-frame deletions[41].

Variants outside of the canonical sites can also exert pathogenic effects by disrupting splicing regulatory motifs. A missense variant in exon 13 of *CFTR* impacts a purine rich sequence (3'-GAAAGAAGAAA-5') in the middle of the exon. Substitution of adenines to thymines (3'-GA**TT**G**TT**G**TT**A-5'; variants bolded) within a mini-gene construct containing exons 12-14 of *CFTR* with ~100 bp of flanking introns lead to preferential selection of cryptic splice site 248 bp downstream of the native 3' splice site, and this alternate splice site usage was recapitulated with the missense variant cloned on a WT background (3'-GAA**T**GAAGAAA-5'; variant bolded)[42]. The use of the alternate 3' splice site shifts the transcript out of frame creating a premature termination codon subject to nonsense mediated decay[42]. When these variants are either in a homozygous state or compound heterozygous with another deleterious *CFTR* variant, patients display a cystic fibrosis phenotype with elevated sweat electrolytes and severe fluid buildup in the lungs[42].

1.3 High-throughput splicing assays

Although there are many splice disruptive variants already associated with Mendelian disorders[41,43-46], the advent of relatively inexpensive and rapid whole genome sequencing has created a bottleneck at the step of interpreting and properly classifying the resulting deluge of variation. Across the genome, over 9 billion single-nucleotide variants (SNVs) are possible – leaving aside the large space of other variant types. Most of the 4.6 million missense variants currently identified in gnomAD are rare (minor allele frequency <0.5%) making any statistical association with a disease phenotype underpowered[47]. Only 2% of the variants identified in gnomAD have a clinical interpretation in ClinVar[47], and of those represented in ClinVar, almost half are listed as variants of uncertain significance (VUS) or have conflicting and as yet unresolved interpretations (**Figure 1-3B**). Without functional evidence, most newly discovered variants - even those in disease-associated genes - do not reach the burden of proof to be reliably classified as deleterious or benign. Currently, over 40% of *BRCA1* point variants in ClinVar are classified as VUS and another 8% have conflicting interpretations despite the long-recognized role of *BRCA1* loss of function variants in early onset breast and ovarian cancer (**Figure 1-3B**). This proportion of *BRCA1* VUS has remained relatively stable since 2017[48], so the resolution of these variants is at a standstill. The lack of clear designation for these variants is frustrating for patients and clinicians especially in the case of medically actionable genes like *BRCA1* in which knowledge of a pathogenic variant would invoke a risk management treatment protocol.

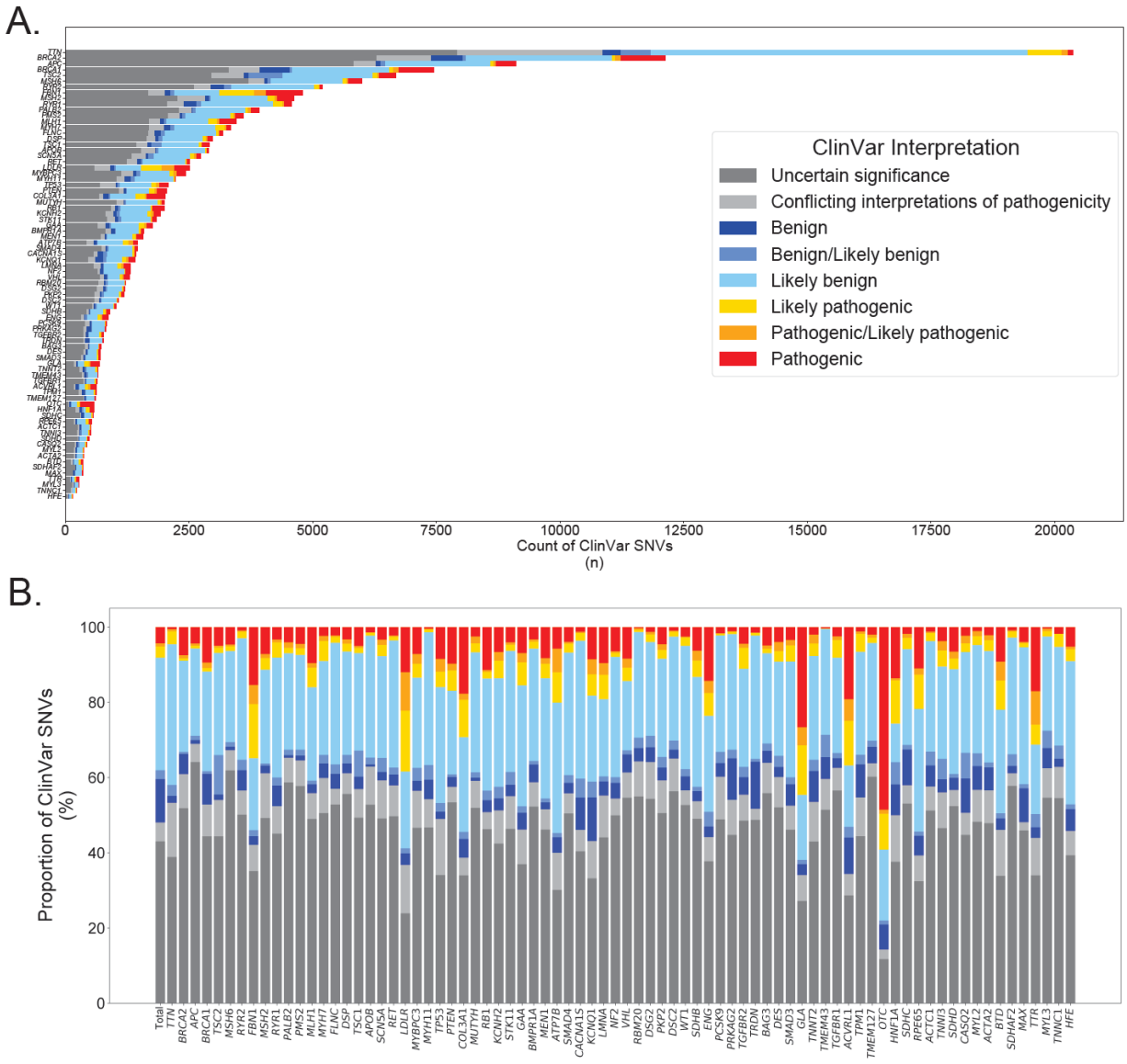


Figure 1-3: SNVs within ClinVar by interpretation.

A. Counts of ClinVar SNVs for the American College of Medical Genetics (ACMG) actionable genes (v3.1)[49] colored by variant interpretation. B. Proportions of variants, by ClinVar variant interpretation, within each of the ACMG actionable genes (v3.1)[49] and colored as in A.

Typically, variants discovered in patients would be classified using statistical association tests or individual functional assays. However, for variants that are rare in the population or private to a family, there is insufficient power for statistical associations, and the scale of unclassified variants makes individual functional assays

for each variant infeasible. Recently, functional assays have been designed which allow testing of multiple variants in parallel within a single experiment[50,51]. These assays can assess protein function of DNA damage response genes[48,52], DNA methyltransferase function[53], enhancer function[54], and splicing effects[17,19,36,48,55-63] to name a few. These approaches use high throughput DNA or RNA sequencing to measure outcomes of specific variants, each often tagged with a uniquely identifying barcode. The results of a high-throughput assay assessing loss of function effects of missense variants within *MSH2* have been applied to reclassify ~400 ClinVar VUS as either pathogenic or benign[64]. Since heterozygous loss of function variants in *MSH2* cause Lynch Syndrome, a predisposition towards early onset colon and endometrial cancers that can be mitigated with frequent, early cancer screenings[65,66], the reclassification of these variants via a high-throughput functional assay can have an immediate impact for patients undergoing genetic testing.

Since a substantial minority of all pathogenic variants act by disrupting mRNA splicing, high throughput investigation of splicing effects within disease associated genes is warranted. Synonymous variants (which do not alter the protein sequence), and intronic variants outside of the essential splice sites dinucleotides are of particular interest with respect to candidate pathogenic variants as they are sometimes given low priority due to an assumed lack of protein-coding impact.

One conventional means of assaying variants' splicing impact is via minigene assays. In this approach, cells are transfected with plasmids containing a cloned intron-exon-intron sequence, optionally between two constitutively included exons, surrounded by constant first and last exons. The cloned region contains the variant of interest. After

its transient transfection, the mini-gene plasmid is transcribed, RNA is recovered, and the resulting splicing outcomes can be measured in low-throughput using RT-PCR and gel electrophoresis of the RT-PCR product. Recently, high-throughput versions of this approach have been developed, in which multiple variants are cloned in a library with each variant being tagged with an identification marker within the mini-gene plasmid sequence. The variants are then linked to their identification tag through sequencing and transfected into cells within a single experiment. RNA is collected and the splicing effects are measured in a high throughput manner using RNA-seq or employing a reporter construct such as GFP and using FACS.

The high throughput splicing assays can measure the effects of individual variants across multiple exons simultaneously[36,61,62] or the effects of multiple variants within a single exon[17,19,48,55-60,63], which provide complementary views on variant effects in splicing. This dissertation will focus on saturation splicing screens in which the splicing effect all possible point variants within a single exon of interest are measured in parallel. Since high throughput splicing saturation screens have recently identified a handful of putative exonic splicing regulatory regions[17,48,55,59], measuring the full landscape of splicing effects across an exon provides a means to elucidate the impacts of rare variants implicated in disease, and also to better understand the dynamics of key splicing motifs outside of the splice sites. The implementation of these screens varies slightly across groups, but the assay itself is generally agnostic to the specific exon or gene model being interrogated unlike high throughput assays to evaluate protein function which are very gene specific (see Jia et al., 2021[52] vs. Lue et al., 2022[53] for example). To measure splicing outcomes of all possible point variants en masse, a

barcode is added to the 3' UTR of the mini-gene construct and these barcodes are linked to a library of individual variants[55] (**Figure 1-4**). Then after transfection, the barcodes are read out using RNA-seq and the splicing outcomes are tallied within each barcode[55](**Figure 1-5**). Finally, splicing results for each barcode are aggregated into variant specific splicing effects allowing the nomination of candidate splice disruptive variants[55](**Figure 1-5**).

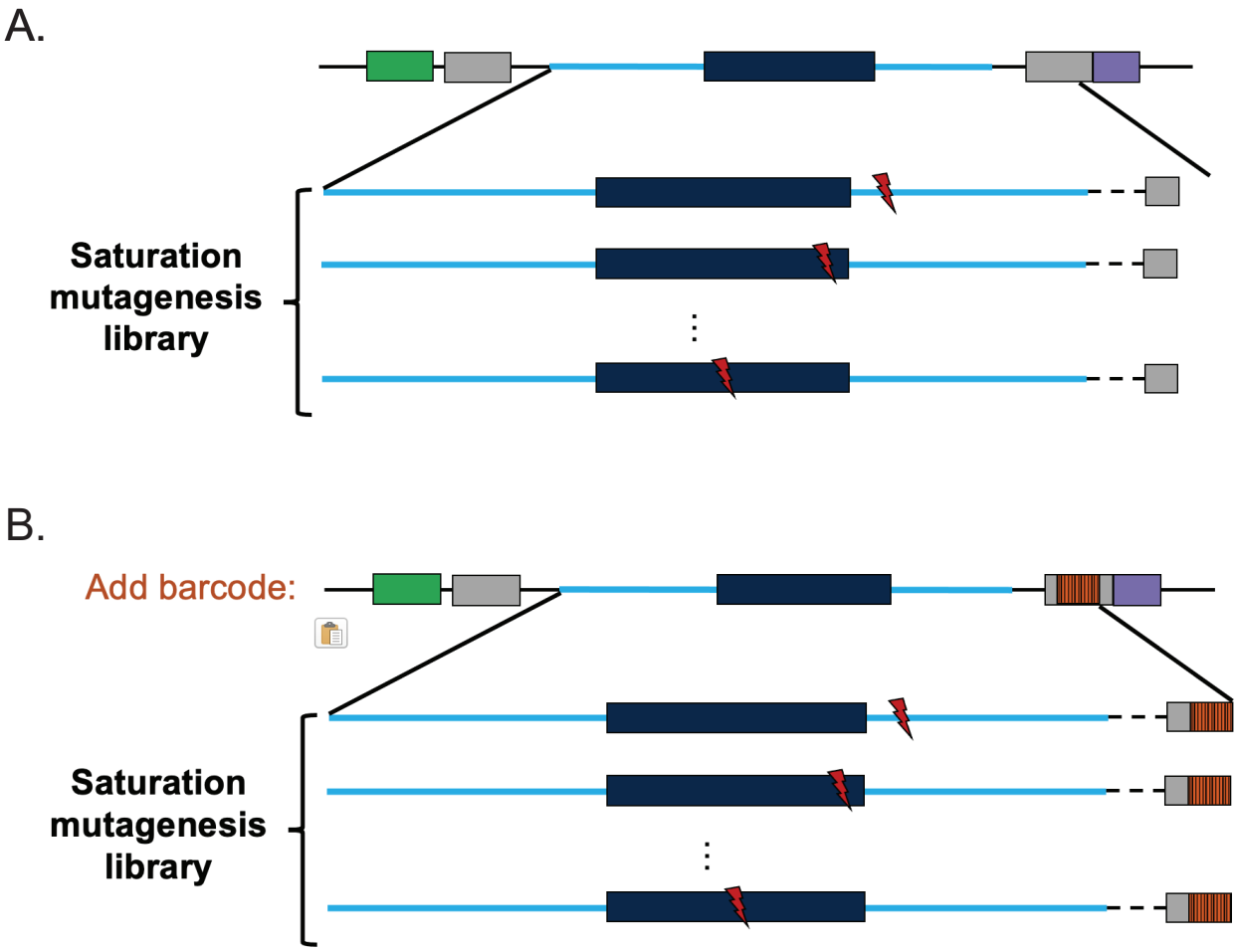


Figure 1-4: Library creation for massively parallel splicing assays.

A. Mini-gene construct with promoter (green), constant upstream and downstream exons (gray), target exon (dark blue), and 3'UTR (purple). Variant library is constructed targeting every single nucleotide variant within the cloned region (light blue) of the exon and surrounding introns. Hypothetical variants shown as red lightning bolts. **B.** To track the variants, a barcode (orange) is added to the 3' UTR.

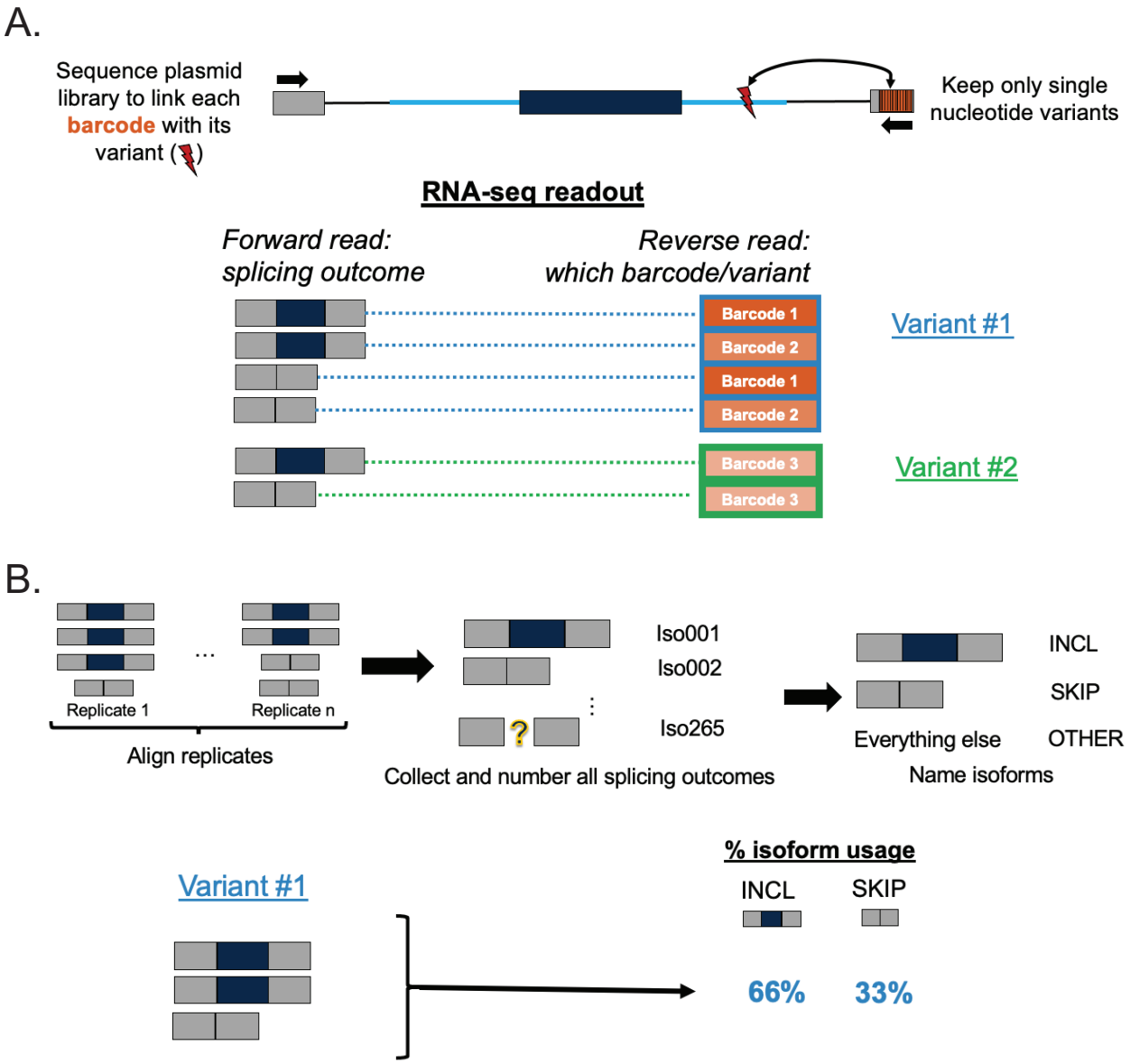


Figure 1-5: Processing RNA-Seq data for MPSAs.

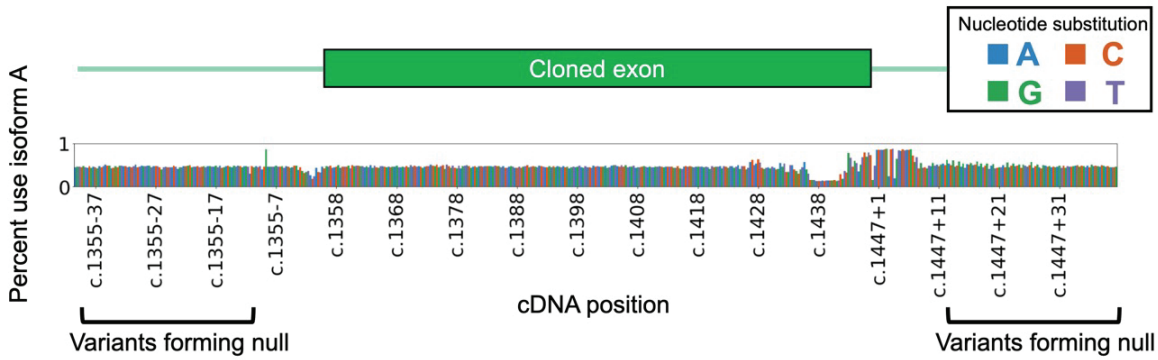
A. Deep sequencing links barcodes (orange) to individual variants (red lightning bolts). Plasmids are transfected into cells and the resulting RNA-seq dataset has a forward read representing the splicing outcome and a reverse read which contains the identifying barcodes and, sometimes, the splicing outcome. Variants are linked to multiple barcodes and each individual barcode can contain a diverse set of isoforms. **B.** Each biological replicate is first aligned. The, all possible splicing outcomes are collected, numbered, and assigned to named categories. Finally, isoform representation is tallied within each barcode and aggregated within variants to summarize the percent isoform usage.

I implemented a custom python module to process the RNA-seq reads, compute percent isoform usage for each variant, nominate variants as splice disruptive, and visualize the resulting data. Relying heavily on pysam and pandas, the module takes the aligned RNA-seq reads as input and starts by collecting, numbering, and finally categorizing the isoforms represented within each biological replicate (**Figure 1-5**). Typically, a small number of isoforms are named (exon inclusion and exon skipping for instance), and many of the isoforms are ushered into a catch-all other category of lowly represented, rare isoforms ('OTHER'). Within each barcode, the counts of each isoform category are tallied. Then, using a look-up table that links identifying barcodes to individual variants, barcode level isoform counts are aggregated into variant specific percent isoform usage. To account for differences in read coverage, each barcode's isoform counts are weighted by the number of associated reads before computing variant level summaries. In this way, barcodes represented by more reads have more influence over the final isoform usage values. We finally take the median of the percent isoform usage across replicates as our final outcome measure.

Next, the percent isoform usage tallies are used to nominate candidate splice disruptive variants. To discern which variants have isoform usage outside of the baseline usage within a given exon, we must first construct a null distribution. To do this, we sample barcodes from the distal intronic variants, which usually have minimal impact on splicing, with replacement (bootstrapping; **Figure 1-6**). Since each variant is represented by a different number of barcodes, we bootstrap 1,000 samples with a matched number of null distribution barcodes for each variant. In this way, our null distribution can naturally reflect our uncertainty for variants with sparser barcode

representation (**Figure 1-6B**). Then, within each isoform and variant, we can compare the standardized, observed isoform usage with our standardized bootstrapped null distribution to arrive at a *p*-value. We perform this procedure separately for each biological replicate to account for any possible batch effects, and we compute isoform specific *p*-values to allow one variant to disrupt multiple isoforms. Finally, our resulting *p*-values are aggregated across replicates using Stouffer's test.

A.



B.

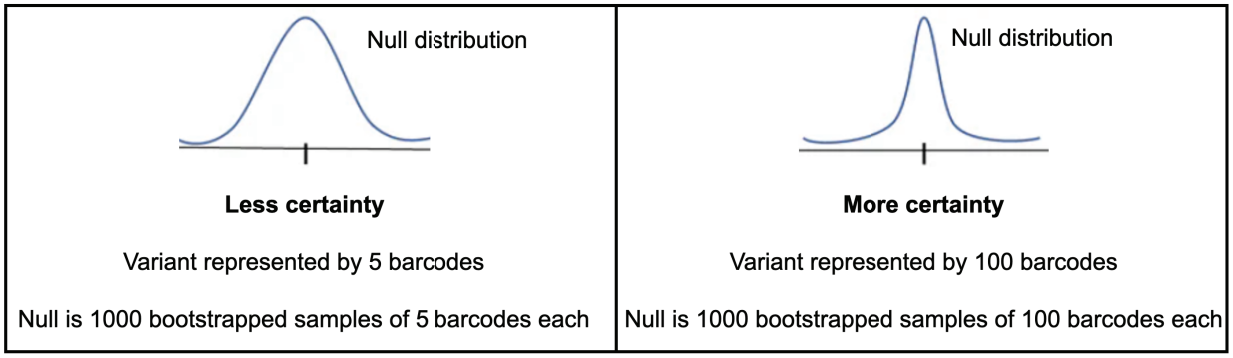


Figure 1-6: Constructing a null distribution.

A. Splicing effect map for a hypothetical exon. Percent isoform use (y-axis) by transcript position (x-axis) for individual variants. Each variant is represented by a bar and shaded by nucleotide substitution. Distal intronic variants used to form the null distribution are boxed. **B.** Boot-strapped null distributions reflect uncertainty about experimental measurements. Hypothetical null distributions for variant with sparse (left) and dense (right) barcode coverage. Bootstrapped null distributions have a variant matched number of barcodes within each sample.

Next, we create an exon-specific effect size threshold as another indicator of splice disruption. We inform our effect size threshold with expert knowledge about the exon, known variants associated with pathological phenotypes, dosage tolerance of the exon, and population level frequencies. By carefully selecting an effect size threshold, we aim to separate pathogenic variants from those which may alter splicing without creating a phenotype. Variants with a significant *p*-value after FDR correction and meeting our effect size threshold criteria are then nominated as splice disruptive. We can then plot and visually inspect the resulting splicing effect maps and SDV classifications.

In Chapters 2 and 3, we deploy two massively parallel splicing assays to measure the splicing effects of every point variant in and around exon 2 of *POU1F1*[55] and exon 9 of *WT1*[56] respectively. *POU1F1* is a key pituitary specific transcription factor involved in the regulation of growth hormone, prolactin, and other pituitary hormones[67]. Loss of function variants in *POU1F1* are associated with hormonal deficiencies, short stature, and alteration of pubertal timing[68]. The use of two competing 3' splice sites of exon 2 of *POU1F1* results in two mutually antagonistic isoforms – alpha and beta[69]. The alpha isoform is the predominant, activating isoform while the beta isoform is normally lowly expressed and serves a repressive role[69-71]. Four adjacent missense variants within the beta specific region of exon 2 were identified in patients with growth hormone deficiencies, and we hypothesized that those variants could be altering splicing by increasing the expression of the minor beta isoform[55]. Chapter 2 outlines the results of our massively parallel splicing assay measuring the effects of the four missense

variants identified in hypopituitary patients along with the 1,070 other possible point variants in and around exon 2 of *POU1F1*.

In Chapter 3, we examine the splicing effects of variants within exon 9 of *WT1* which, unlike *POU1F1* exon 2, has two alternatively spliced isoforms using competing 5' splice site donors. *WT1* is an essential transcription factor that regulates kidney function and sexual development[72]. Loss of function variants within *WT1* are associated with an array of overlapping nephrotic syndrome phenotypes including Denys-Drash and Frasier's syndrome[73]. Frasier's syndrome is caused by splicing dysregulation which alters the ratio of two isoforms: KTS+ and KTS-[73,74]. The function of each isoform has not yet been fully elucidated but the KTS- isoform is generally associated with DNA-binding function as a transcription factor while the KTS+ may bind at a distinct set of targets, and separately, has been implicated as an RNA-binding factor, with functional roles that remain unclear[72,75]. The two isoforms differ by three amino acids – KTS – and are normally expressed in mature kidneys at a ~2:1 KTS+/KTS- ratio, but in Frasier's syndrome a higher expression of the KTS- isoform is linked to pseudo-hermaphroditism, nephrotic syndrome, and tumors[73,74,76]. Seven adjacent intronic variants downstream of the KTS+ donor site have been previously shown to lower the KTS+/KTS- ratio in Frasier's syndrome patients[76], and in Chapter 3, we measure the splicing effects of those seven variants as well as the other 512 possible point variants in and around *WT1* exon 9 to identify other potential splice altering variants that could cause a nephrotic or sexual development phenotype.

1.4 Computational prediction of splicing effects

High throughput functional assays can classify hundreds of variants simultaneously and RNA analysis of patient blood or tissue can provide strong evidence of splicing defects. However, RNA analysis may not be feasible for tissue specific splicing effects within difficult to access tissues, and the extent of unclassified variants makes the task of measuring splicing effects genome-wide daunting even with a high throughput approach. So, accurate and reliable computational predictions of splicing effects are needed to identify pathogenic variants at scale. The challenge of bioinformatic prediction of splice sites has been long standing with initial attempts using newly annotated genetic sequences to first define the bases comprising consensus 3' and 5' splice sites and then to design a position weight matrix (PWM) to identify putative splice sites with similar motifs in random sequences[77]. Similar to the identification of consensus splice site sequences, PWMs are built by counting the frequency of each nucleotide at each position surrounding the fixed splice sites across the genome and then converting those frequencies to probabilities. These PWMs can then be applied to cryptic or existing splice sites to measure their splice site strength – that is, sequences resembling the consensus motif would be more likely to be selected by the spliceosome as an authentic splice site. Thus, exons with high inclusion rates would be expected to be flanked by splice site motifs closely resembling the consensus sequence and with a corresponding high probability score from the PWM. However, classic PWMs treat every position as independent, so the next wave of splice prediction tools employed Markov models, early neural nets, decision tree methods, and maximum entropy models to incorporate information about dependencies between adjacent and non-

adjacent positions at the splice sites[78-81]. MaxEntScan is one of the enduring tools from this era and uses a maximum entropy distribution trained on the short motifs surrounding splice sites to predict relative splice site strength[78]. The maximum entropy model underlying their method uses a greedy search algorithm to select the best set of constraints modeling the interdependent nucleotide probabilities. Fitting those constraints results in a likelihood ratio interpreted as the probability each motif represents a true splice site or the splice site strength of each motif compared to a set of cryptic motifs.

Since there is not enough information content within the conventionally recognized splicing motifs (splice sites, branchpoint, polypyrimidine tract) to distinguish authentic splice sites from decoys and thereby define exons[16], the next task was to computationally and experimentally define the short regulatory motifs outside of the splice site region. Various groups looked for computational signals of potential splicing regulatory motifs by identifying motifs enriched in exons with weak vs. strong splice sites[27,29,82], motifs enriched in constitutive vs. alternatively spliced exons[82,83], motifs enriched in exons vs. pseudo exons or UTRs[84], motifs depleted in authentic exons[84], or through sequence conservation[24,27,29,85]. On the experimental side, the enumeration of key splicing regulatory motifs has proceeded through mutagenesis of random short motifs within mini-gene constructs. Putative splicing regulatory motifs have been identified with the iterative SELEX method paired with RT-PCR[86-88], FACS selection in which exon skipping would reconstitute a fluorescent reporter[89], and more recently, saturation mutagenesis of all possible hexamers within two different cloned exons read out using RNA-seq[17] and saturation mutagenesis of short motifs around the 3' and 5'

splice sites read out with RNA-seq[90]. Although the resulting lists of potential regulatory motifs was almost exhaustive of all possible motifs and there was little overlap across the various sets of putative short regulatory motifs[30], the catalogs of short motifs have been used as features in a number of bioinformatic splice prediction tools[90-93], and computational algorithms have been built to detect potential regulatory elements within random sequences[92,94]. One of these, HAL, uses an additive linear model to predict the change in percent spliced in (PSI) of the impacted exon based on the presence of synthetic short motifs in and around the 5' and 3' splice sites derived from saturation mutagenesis measurements at the splice sites of entirely synthetic sequences subjected to MPSAs[90].

Contemporary bioinformatic splice prediction algorithms rely on deep learning or more conventional machine learning algorithms to predict splice site usage, and are trained on annotated sequences, curated benign and pathogenic variant sets, and evolutionary constraint layered, and features derived from existing prediction algorithms. Large-scale genetic variation databases such as gnomAD, reflective of a general adult population[95], and ClinVar, which is a clearinghouse for clinical variants and their expert interpretations (benign, pathogenic, or uncertain)[96] serve as one source of training data for these tools. However, there are several potential problems with this type of training data. For instance, defining benign splice variants based on a high population frequency could include most modestly common variants (e.g., minor allele frequency 0.1-1%) that alter splicing without causing a Mendelian phenotype. Moreover, not all ClinVar entries' interpretations are based on functional evidence – for instance a synonymous variant might be presumed benign without further investigation but could in fact alter splicing.

Finally, the available training data have a severe class imbalance, with the sets of known pathogenic variants tending to over-represent variants at the essential dinucleotides of canonical splice sites, which are straightforward to predict and could lead to inflated performance during the testing phase of tool creation. SQUIRLS, MMSplice, and S-Cap were all trained and tested using sets of benign and pathogenic variants that were either measured in previous massively parallel splicing assays or curated from databases like ClinVar and gnomAD[91,97,98]. All three algorithms use sequence features derived from small-scale, defined exonic and intronic regions within machine learning frameworks to create their predictions. MMSplice trained on the same short motif dataset as HAL and additionally incorporated information from annotations of the impacted exons, disease states of ClinVar variants, and splicing effects from broad splicing saturation screens as part of training. S-Cap bolstered the predictions of SPANR – another bioinformatics splice prediction algorithm – with additional sequence features based on short defined sequence regions including evolutionary conservation as measured by phyloP, CADD, and LINSIGHT, and disease states from curated clinical pathogenic sets v. presumably benign variants found in gnomAD. SQUIRLS trained on a custom pathogenic variant set vs benign ClinVar variants and modeled the information content within small regions of the impacted exon as well as incorporating evolutionary conservation and measured scores of exonic splicing regulatory regions.

Similar to earlier tools like MaxEntScan[78], another set of algorithms is trained using information from annotated sequence which avoids some of the inherent flaws related to curated sets of deleterious and neutral variants. Bayesian methods drive SPANR's splice effect predictions which were trained from sequence features derived

from the annotations of alternatively spliced exons and estimates of exon specific percent spliced in (PSI) values[99]. SPANR extracts sequence features within the impacted exon and proximal introns including splice site strength, nucleosome positioning, and existence of splicing regulatory motifs. Two recent tools – SpliceAI and Pangolin – use deep learning on a long flanking sequence context (10,000 bp) to predict the probability that each position in the genome is an acceptor or donor[100,101]. Neither tool used pathogenic/benign labels during training, instead deriving truth labels for each chromosomal position in the training set based upon whether that position is an acceptor or donor within a set of gene model annotations. These tools' predictions of splice disruption are based on the change in probability between WT and variant sequences that each position is a splice site. Pangolin advances the previous efforts of SpliceAI by training on both human and other mammalian (primate, rat, and mouse) sequence annotations and by providing tissue specific predictions. SpliceAI in particular has outperformed other contemporary splice prediction algorithms in recent benchmarking studies[102-107], so SpliceAI predictions have been used as a feature in CADD-Splice, a metaclassifier which predicts overall pathogenicity of each variant and is not restricted to splice altering variants[108], and in ConSpliceML, which added a metric of constraint against splice-disrupting population variation using information from gnomAD and SpliceAI predictions. Combining these constraint metrics with SQUIRLS and SpliceAI predictions into an ensemble classifier may help to differentiate pathogenic and splice disruptive variants since not all splice altering variants create a deleterious phenotype[105].

In Chapter 4, we will systematically benchmark several current bioinformatics tools using saturation splicing screens, which have not yet been extensively used to that end. Specifically we will test HAL[90], SQUIRLS[91], MMSplice[93], S-Cap[98], SPANR[99], SpliceAI[100], Pangolin[101], and ConSpliceML[105] using datasets derived from massively parallel splicing assays in *POU1F1*[55] (Chapter 2), *WT1*[56] (Chapter 3), *BRCA1*[48], *FAS*[57], and *RON*[59] as well as one curated set of neutral and splice disruptive variants from the literature with functional evidence in *MLH1*. By testing each tool against saturation splicing screens, we avoid the bias associated with the over-representation of canonical sites within manually curated sets of variants. We then report the best performing tool overall and within different variant classes and provide specific recommendations for areas of improvement within computational splice prediction.

Chapter 2 High-Throughput Splicing Assays Identify Missense and Silent Splice-Disruptive *POU1F1* Variants Underlying Pituitary Hormone Deficiency

2.1 Abstract

Pituitary hormone deficiency occurs in ~1:4,000 live births. Approximately 3% of the cases are due to mutations in the alpha isoform of POU1F1, a pituitary-specific transcriptional activator. We found four separate heterozygous missense variants in unrelated individuals with hypopituitarism that were predicted to affect a minor isoform, POU1F1 beta, which can act as a transcriptional repressor. These variants retain repressor activity, but they shift splicing to favor the expression of the beta isoform, resulting in dominant negative loss of function. Using a high throughput splicing reporter assay, we tested 1,070 single nucleotide variants in *POU1F1*. We identified 96 splice disruptive variants, including 14 synonymous variants. In separate cohorts, we found two additional synonymous variants nominated by this screen that co-segregate with hypopituitarism. This study underlines the importance of evaluating the impact of variants on splicing and provides a catalog for interpretation of variants of unknown significance in the *POU1F1* gene.

2.2 Introduction

POU1F1 (formerly PIT-1, OMIM: 173110) is a signature pituitary transcription factor that directly regulates the transcription of growth hormone (*GH,* OMIM: 139250*)*,

prolactin (*PRL,* OMIM: 176760*)*, and both the alpha (*CGA,* OMIM: 118850) and beta (*TSHB,* OMIM: 188540) subunits of thyroid stimulating hormone[67,109]. In mice, *Pou1f1* is expressed after the peak expression of *Prop1* (OMIM: 601538) at E14.5 and remains expressed into adulthood[110,111]. A well-characterized mutant of *Pou1f1* (*Pou1f1*$^{dw/dw}$) carries a spontaneous missense mutation (p.Trp251Cys) in the homeodomain that disrupts DNA binding[111,112]. The homozygous mutant mice have no somatotrophs, lactotrophs or thyrotrophs except for the *Pou1f1*-independent rostral tip thyrotrophs[111,113,114]. In humans, loss of *POU1F1* function typically results in GH, TSH and PRL deficiency[68].

POU1F1 undergoes an evolutionarily conserved program of alternative splicing[115,116], resulting in a predominant isoform, alpha, that acts as a transcriptional activator and a minor isoform, beta, that acts as a transcriptional repressor[69,70,109]. In the human pituitary gland, the beta isoform comprises approximately 1-3% of *POU1F1* transcripts [115,117]. The POU1F1 beta isoform transcript is created by utilization of an alternative splice acceptor sequence for exon 2, located 78 bp upstream of the alpha acceptor, resulting in a 26 amino acid insertion that encodes an interaction domain for the transcription factor ETS1 (OMIM: 164720). This insertion, which is absent in the alpha isoform, disrupts the POU1F1 transactivation domain at amino acid 48. The POU1F1 alpha and beta isoforms have different activities depending on the context of the target gene[69]. For example, the POU1F1 alpha isoform activates its own expression, but the beta isoform does not, and the beta isoform interferes with alpha isoform mediated auto-activation[70]. Although alternative splicing of POU1F1 is

evolutionarily conserved among vertebrates, the functional significance of the minor, beta isoform remains unclear[115].

The first case of a recessive *POU1F1* loss of function was described in a child with combined pituitary hormone deficiency (CPHD, OMIM: 613038, 262600, 221750, 262700, 601538, 173110, 615849, 600577, 182230, 612079, 602146) born to consanguineous parents[118]; since then, many unique variants in *POU1F1* have been reported in people with CPHD or isolated growth hormone deficiency (IGHD, OMIM: 307200, 262400, 173100, 612781, 139250, 618157,139191, 262500, 615925, 618160)[119-125](reviewed in [126]). A few dominant negative mutations have been reported that likely act by interfering with the function of POU1F1 dimers. The variant p.Pro76Leu alters the transactivation domain and causes completely penetrant IGHD[127], p.Lys216Glu interferes with the ability of POU1F1 to interact with retinoic acid receptors and CREBBP (p300, OMIM: 600140)[128], and p.Arg271Trp interferes with the ability of POU1F1 to be tethered to the nuclear matrix through MATR3 (OMIM: 164015), SATB1 (OMIM: 602075) and CTNNB1 (OMIM: 116806)[129]. All of the reported mutations are located in domains shared by the alpha and beta isoforms of POU1F1 and were functionally tested using the alpha isoform only.

We found four missense variants, in four independent families, that shift splicing to favor the POU1F1 beta isoform almost exclusively, while retaining its transcriptional repressor activity on the *POU1F1* enhancer. We used a high throughput assay to identify in total 132 variants in and around exon 2 that cause exon skipping, isoform switching, or cryptic isoform use. With this splicing effect catalog, we evaluated additional families with hypopituitarism and identified two unrelated individuals carrying

synonymous *POU1F1* variants that affect its splicing without changing the amino acid sequence. This study underscores the importance of evaluating splicing defects as a disease mechanism.

2.3 Methods

2.3.1 Informed consent

The studies were approved by ethical committees: the local Comite de Eticae Pesquisa da Faculdade de Medicina da Universidade de São Paulo (CEP-FMUSP) and the national Comite nacional de etica em pesquisa (CONEP) CAAE, 06425812.4.0000.0068; the Ethics Committee of the Faculty of Medicine, University of Leipzig (UL), Karl-Sudhoff-Institute for Medical History and Natural Sciences, Käthe-Kollwitz-Straße 82, 04109 Leipzig, Germany; and the Comité de Ética en Investigación (Research Ethics Committee) of the Hospital de Niños Ricardo Gutierrez (HNG), Gallo 1330, Ciudad autónoma de Buenos Aires, Argentina (CEI Nº 16.06). The GENHYPOPIT network collected anonymized information in a database declared to health authorities in accordance with local regulations at Aix-Marseille Université (AMU) - Conception Hospital (Assistance Publique - Hôpitaux de Marseille, AP-HM), and a declaration was made to the National Commission for Data Protection and Liberties (CNIL-France): 1991429 v 0. Adult individuals or the parents of children signed a written informed consent to participate. Families 1, 3, and 6 are historical cases that were referred to the GENHYPOPIT network for genetic testing. Limited information is available for Families 1 and 3, and they were lost for follow up. The University of Michigan Institutional Review Board (UM) found the study exempt because DNA samples were anonymized before exome sequencing at UM.

2.3.2 Genomic DNA sequencing

Individuals from Families 1, 2, 4, and 5 underwent whole exome sequencing (WES). Representative *POU1F1* variants in Family 3 and 6 were discovered in a traditional CPHD candidate gene screening using Sanger sequencing (*PROP1*, *POU1F1*, *LHX3* and *LHX4*) (*LHX3* OMIM 600577, *LHX4* OMIM 602146). WES of Families 1 and 5 was carried out at University of Michigan as previously described[119]. WES of Family 2 was performed at the Broad Institute as previously described[130]. WES of Family 4 was performed at the Institute of Human Genetics at University of Leipzig.

2.3.3 Expression vectors and cell culture

The open reading frame of either *POU1F1* isoform alpha (RefSeq: NM_000306.3) or beta (RefSeq: NM_001122757.2) was cloned into pcDNA3.1+/C-(K)-DYK. Site directed mutagenesis was used to obtain each of the variant POU1F1 beta isoforms: p.Ser50Ala, p.Ile51Ser, p.Leu52Trp, and p.Ser53Ala in the beta isoform (Genscript). A firefly luciferase reporter gene was constructed in pNBm81-luc with 14 kb of the mouse *Pou1f1* 5' flanking sequences that includes early and late enhancers and the promoter, and 13 bp of the 5'UTR. Cloning was performed with Infusion HD (Clontech) or NEBuilder HiFi DNA Assembly (New England Biolabs). Plasmid sequences were confirmed by Sanger sequencing. The pRL-TK renilla (Promega) was used as a normalization control and pcDNA3.1(-) (Thermo-Fisher) to keep the total DNA constant. COS-7 and GH3 cells were purchased from the American Type Culture Collection. Cells were maintained in Dulbecco's modified eagle medium (DMEM, Gibco, Grand Island, NY, USA) containing 10% fetal bovine serum and pen-strep (Gibco). Plasmids were transiently transfected using ViaFect Transfection Reagent (Promega,

Madison, WI, USA). Luciferase activities were measured as suggested by the manufacturer (Dual-luciferase assay system; Promega).

2.3.4 Exon trapping assay

Human *POU1F1* exon 2, flanked by partial intron 1 (85 bp upstream) and intron 2 (178 bp downstream), was cloned into the BamHI cloning site of the pSPL3 plasmid (Invitrogen) to create an exon trapping plasmid with a total insert size of 413 bp. Similarly, a minigene exon trapping plasmid was constructed that included the last 85 bp of intron 1 and the first 85 bp of intron 5, for a total insert size of 3,442 bp including exons 2, 3, 4, and 5. Site directed mutagenesis was used to create the desired variants. Plasmids were transiently transfected into COS-7 cells. Total RNA was purified with RNeasy mini (Qiagen). After reverse transcription, we analyzed exon trapping using RT-PCR with following primers; Primers SD6 Forward (5'-TCT GAG TCA CCT GGA CAA CC- 3') and SA2 reverse (5'- ATC TCA GTG GTA TTT GTG AGC - 3')[131].

2.3.5 POU1F1 Saturation Mutagenesis

The cloned *POU1F1* fragment in pSPL3 was divided into four overlapping tiles of 150 bp each, spanning exon 2 plus flanking introns (79 bp upstream to 131 bp downstream). Mutant tile libraries containing every possible single nucleotide variant were synthesized as a single 150mer oligonucleotide pool by Twist Bio. HiFi Assembly was used to replace each wild type tile with the respective mutant tile library amplified from the oligo pool. The resulting mutant minigene library pools were transformed in 10b *E. coli* (New England Biolabs), with a minimum coverage of 90 clones per mutation.

2.3.6 Mutant library barcoding and sequencing

To tag each mutant minigene clone with a unique barcode, a random barcode sequence (N_{20}) was inserted by HiFi Assembly into the MscI site within the common 3' UTR. Subassembly sequencing[132] was used to pair each 3' UTR barcode with its linked variant(s) in cis. Briefly, a fragment starting with the POU1F1 insert and ending at the N_{20} barcode (2.2 kb downstream) were amplified from the plasmid library DNA by PCR using 5'-phosphorylated primers. The resulting linear fragment was re-circularized by intramolecular ligation using T4 DNA ligase (NEB), to bring each barcode in close proximity to the mutagenized region. From this re-circularized product, paired-end amplicon sequencing libraries were generated, such that each reverse read contained a plasmid barcode and the paired forward read contained a sequence from the associated POU1F1 insert. Barcode reads were clustered with starcode[133] (arguments "-d 1 -r 3") to generate a catalog of known barcodes. Variants were called within each barcode group using freebayes[134] and filtered to require majority support, and read depth ≥4 along the entire region targeted for mutagenesis. Barcode-variant pairing was confirmed by Sanger sequencing of 15 clones selected at random from the POU1F1 library; of those, 13/15 were found in the final catalog of reconstructed sequences and associated barcodes, and all 13 perfectly matched the Sanger-sequenced clones.

2.3.7 Pooled exon-trap transfection and RNA-seq

COS-7 cells were plated at $5x10^6$ cells/60 mm plate. Each was transfected with 4 ug of the barcoded mutant exon-trap library using ViaFect reagent (3:1 ratio to DNA). After 24 hours, RNA was purified as above, and 5 ug of total RNA was used to prepare first-strand cDNA using the SuperScript III First-Strand Synthesis kit (Invitrogen) with

oligo dT primers. Spliced transcript was amplified using nested PCR, initially for 6 cycles using the SD6F/SA2R primers, followed by 20 cycles using primers SD2F/jklab0046 (TGTAGTCAGTGCCATCTTGGATCT). Paired-end Illumina sequencing libraries were generated by tailing PCR (6 cycles) with a forward primer within the constant upstream exon (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT AGGGCATAGTGCCATCTTGGATCT) and a reverse primer immediately downstream of the N_{20} barcode (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGAACTGCACTGTG ACA AGCTGC). Unique dual i5/i7 indices were added by a second round of tailing PCR (6 cycles), and the resulting products were purified by SPRI bead cleanup and submitted for Illumina sequencing on a Hiseq 4000 and/or Novaseq instrument.

2.3.8 RNA-seq processing pipeline

Reverse reads containing the plasmid barcode were searched for exact match to a known barcode from the plasmid library. Forward reads containing the spliced sequence were mapped to a variant-specific reference consisting of the POU1F1 exon trap reference sequence with the respective mutation introduced in silico, using GMAP[135] (arguments " -t 8 -f samse --microexon-spliceprob=1.0 --allow-close-indels=2"). From the spliced reads, an isoform catalog was tallied requiring each isoform to be represented by at least three distinct barcodes and nine reads. Spliced reads associated with each barcode were tallied to produce per-barcode isoform usage counts, and percent spliced in (PSI) fractions. Barcodes corresponding to the same POU1F1 variant were then aggregated (weighted by the number of reads obtained) to

generate for each variant a mean PSI score for all known isoforms. Isoforms not matching a known isoform (beta, skip, or alpha) were placed in a catch-all category called “OTHER”. Barcodes represented by fewer than three reads were discarded from further analyses.

2.3.9 Fold-change and significance testing

PSI distributions under the null hypothesis (no splicing difference) were approximated by bootstrap sampling. For each tested variant, the equivalent number of barcodes was drawn (with replacement) from intronic background region variants (defined as intronic variants >20 bp from exon boundaries), repeated 1,000 times, and used to derive a null distribution against which each per-variant observed PSI values was converted to a z-score. For each variant, the z-scores were combined across replicates using Stouffer’s test. This process was repeated separately for each isoform. For each of the three tested alternative isoforms (beta, skip, other), a fold-change over background was calculated for each variant. This was taken as the PSI value for that variant and isoform, divided by the sampling mean PSI for that isoform derived from the intron background region barcodes; the median of these values was then computed across replicates. Variants with a *z*-score > 4.16 (Bonferroni-corrected threshold for p = 0.05) and at least a 3-fold change from the average null distribution PSI for the beta, skip, or other isoform in at least half the replicates were nominated as splice disruptive variants (SDV). Variants which met the *z*-score threshold but had a fold-change between 2 and 3, or which met the SDV criteria overall but failed to meet it individually in ≥7 replicates, were labeled as intermediate.

2.3.10 Comparison of bioinformatic predictors

HAL delta_psi scores[90], SPANR zdelta_psi scores[99], SpliceAI ds_max scores[100], and MMSplice delta_logit_psi scores[93] were obtained from their original publications without modification. To compute per-variant ESRseq scores[17], we took the difference between the mean ESRseq z-scores of hexamers overlapping a variant position from that of hexamers overlapping the corresponding wildtype position. Precision-recall curves were obtained to summarize each algorithm's ability to predict the experimental determination of splice disruptiveness. For algorithms which output signed scores, area under the curve (prAUC) was separately computed using signed and absolute scores as input and the higher prAUC was taken.

2.3.11 Selection of candidate RNA binding proteins (RBP)

RNACompete z-scores[23] were obtained from the cisBP-RNA database. At each position, wild-type and variant-containing z scores were taken as the maximum among the overlapping k-mers, and the difference taken between the wild-type and variant scores. Motifs with high scoring matches (wildtype $z \geq 3$) to the wild-type sequence in the beta variant cluster (c.143 to c.167) were then pursued further.

2.3.12 Data availability

Custom python scripts and notebooks used to process the data are available at https://github.com/kitzmanlab/pou1f1_splicing. A look up table of variant effects is available both as Table S1 within the original publication[55] and on Zenodo[136].

2.4 Results

2.4.1 Mutations in the POU1F1 beta coding region cause hypopituitarism

We initially focused on four cases of hypopituitarism from different cohorts in Europe and South America (**Figure 2-1A**). Affected individuals' presentation was variable, ranging from multiple hormone deficiency with pituitary stalk interruption (Family 1) to isolated GH deficiency (Family 2) (**Table 2-1**; **Figure 2-2)**. The affected individuals had severe short stature and responded well to GH therapy (**Figure 2-1B**). To identify causal variants, we performed whole exome sequencing (WES) for individuals in three families. Combined with conventional Sanger sequencing in another family, this revealed four missense variants in exon 2 of the *POU1F1* beta isoform, each in an unrelated family (**Figure 2-1A, 2-3A**). The four individual *POU1F1* missense variants are absent from Genome Aggregation Database (gnomAD) and in-house population-matched exome databases[137,138], and they are predicted to be damaging by several bioinformatic algorithms (**Table 2-1**). Remarkably, these variants clustered in four consecutive codons within the beta isoform: c.148T>G (p.Ser50Ala), c.152T>G (p.Ile51Ser), c.155T>G (p.Leu52Trp), and c.157T>G (p.Ser53Ala) (**Table 2-1**). Only one of these (c.155T>G, Family 3) appears to be *de novo*; the others were dominantly inherited and co-segregate with hypopituitarism phenotypes, except for c.148T>G which was inherited from the apparently unaffected parent in Family 1, indicating that if causal, this variant is incompletely penetrant. The other parent in Family 1, the two affected children, and one unaffected relative also carried a variant of uncertain significance, *SIX3* p.Pro74Arg (OMIM 603714). No other variants in known hypopituitarism genes were detected.

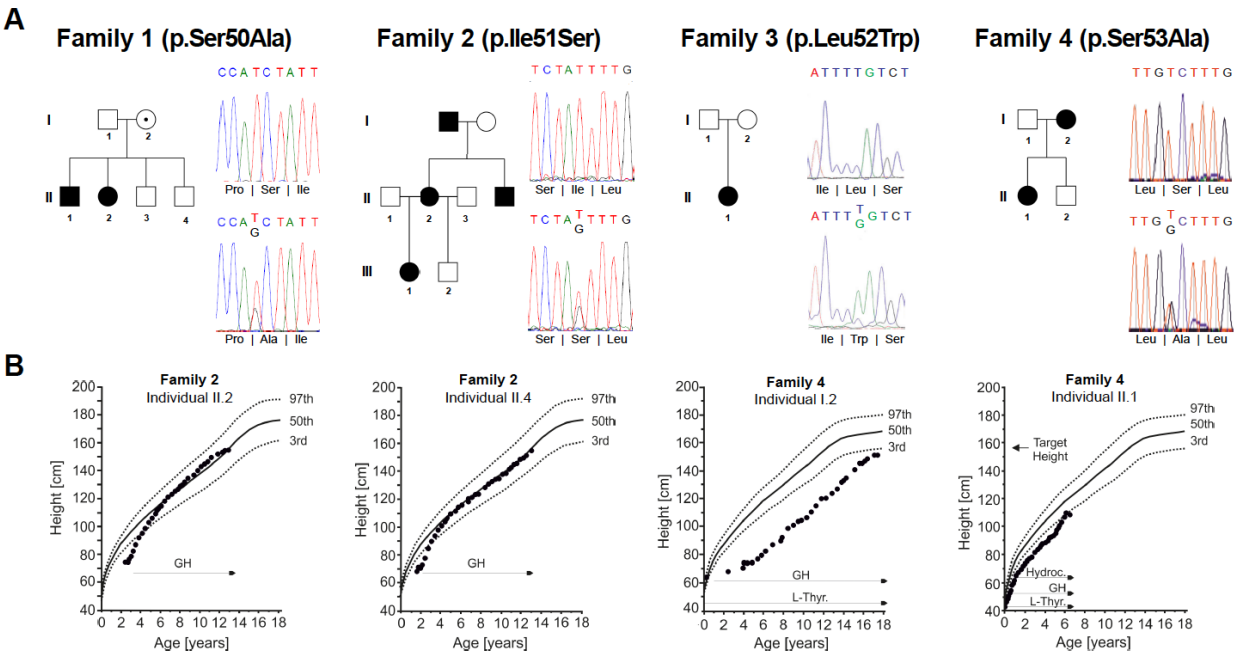


Figure 2-1: Clinical characteristics of the variants of POU1F1 beta coding region.

Pedigrees and the sequence of *POU1F1* variants. Family 1-4 have variants in the POU1F1 beta coding region: c.148T>G (p.Ser50Ala), c.152T>G (p.Ile51Ser), c.155T>G (p.Leu52Trp), and c.157T>G (p.Ser53Ala). B. Growth curve of the affected individuals from Family 2 and 4. GH replacement therapy was effective in reaching ideal height.

Table 2-1: Clinical and molecular features of affected individuals

Feature	Family 1		Family 2				Family 3	Family 4		Family 5		Family 6	
Cases	**II.1**	**II.2**	**I.1**	**II.2 (index)**	**II.4**	**III.1**	**II.1**	**1.2**	**II.1**	**I.1**	**II.1 (index)**	**II.1 (index)**	**II.2**
Sex	Male	Female	Male	Female	Male	Female	Female	Female	Female	Male	Male	Male	Male
Age at diagnosis 1st hormone deficiency / Hormone	<5 yr / GH, TSH	<5 yr / GH, TSH	40s / GH	<5 yr / GH	< 5 yr / GH	<5 yr / GH	<5 yr / GH, TSH, PRL	<5 yr / TSH	<5 yr / TSH, GH	preteen / GH	<5 yr / GH	preteen / GH	preteen / GH
Height at diagnosis of GHD (SDS)	na	na	-3.7	-5	-5.3	na	-4	-5.42	-3.45	-4.2	-4.15	na	na
rhGH treatment (Yes/No)	na	na	No	Yes	Yes	Yes	na	Yes	Yes	Yes	Yes	Yes	na
Final height (cm / SDS)	na	na	150 / -3.7	156.5/-0.9	165/-1.5	na	na	150.9 /-2.67	still growing	147.9 / -3.66	still growing	na	na
Pituitary hormone deficiencies	**GH, TSH**	**GH, TSH**	**GH**	**GH**	**GH**	**GH**	**GH, TSH, PRL**	**GH, TSH, PRL, (ACTH†)**	**GH, TSH, PRL, (ACTH†)**	**GH**	**GH**	**GH**	**GH**
Biochemical assessment													
GHSTs	na	na	clonidine, ITT	clonidine, ITT	clonidine	na	na	glucagon, insulin, clonidine	Basal neonatal	Insulin-Ldopa	Arginine-Clonidine	ITT	ITT
Maximum GH peak (ng/ml)	na	na	7.6	0.9	0.5	na	na	ND	ND	2.9°	2.7**	3.7 mUI/l	3.6 mU/l
TSH (U/L)	na	na	0.6	0.7	1.7	na	na	na	0.28	na	3.4	normal	normal
Total T4 (ug/dL)	na	na	6.4	6	5.6-7.3	na	na	na	na	na	na	na	na
Free T4 (ng/dL)	na	na	0.7	0.6-1.1	0.6-0.9	na	na	na	0.4	0.9	1.2	na	na
Prolactin (ng/mL)	na	na	na	3.8 (pTRH 12)	3.2-6.6	na	na	2.0 mU/l	9 mU/l	8.3	4.4	na	na
Cortisol (ug/dL)	na	na	na	peak ITT 42	normal	na	na	treated in 20's	36.2 nmol/l	13	13.2	na	na
LH/FSH	na	na	na	early puberty	normal puberty	na	na	Delayed puberty, spontaneous pregnancy	Normal	Spontaneous puberty	0.1/0.75	spontaneous puberty	na
Pituitary MRI	Disrupted stalk	Disrupted stalk	Normal	Normal	Normal	na	Normal	Normal	Normal	na	APH	Normal	na
Extrapituitary brain MRI	na	na	na	na	na	na	na	1 cm left frontal and parietal lobe abnormality	Normal	Normal	Normal	na	na
Dysmorphic features	none noted	none noted	none noted	large forehead	none noted	na	na	Intellectual disability, delayed puberty, strabismus, astigmatism, nystagmus, dysplastic thyroid gland	Macroglossia, bilateral hearing impairment, developmental delay, dysplastic thyroid gland	none noted	Short stature, frontal bossing, high pitched voice	na	na
Molecular findings (all in heterozygous state)	**c.148T T>G**	**c.148T T>G**	**c.152T>G**	**c.152T>G**	**c.152T>G**	**c.152T>G**	**c.155T>G**	**c.157T>G**	**c.157T>G**	**c.150T>G**	**c.150T>G**	**c.153T>A**	**c.153T>A**
	p.Ser50Ala	**p.Ser50Ala**	**p.Ile51Ser**	**p.Ile51Ser**	**p.Ile51Ser**	**p.Ile51Ser**	**p.Leu52Trp**	**p.Ser53Ala**	**p.Ser53Ala**	**p.Ser50=**	**p.Ser50=**	**p.Ile51=**	**p.Ile51=**
***In silico* predictions**													
CADD	22.00		23.4				25.8	18.65					
SIFT	damaging		damaging				damaging	damaging					
PP2	benign		benign				probably damaging	benign					
Mutation taster	disease causing		disease causing				disease causing	disease causing					

na: not available; GHSTs: growth hormone stimulation tests, ND: non-detectable, MRI: magnetic resonance imaging, rhGH: recombinant human growth hormone, APH: anterior pituitary hypoplasia; ITT: insulin-tolerance test; *Cut-off 4.8 ng/ml; °*cut-off 10 ng/ml; †diagnosed in late 20s for I.2 and newborn for II.1

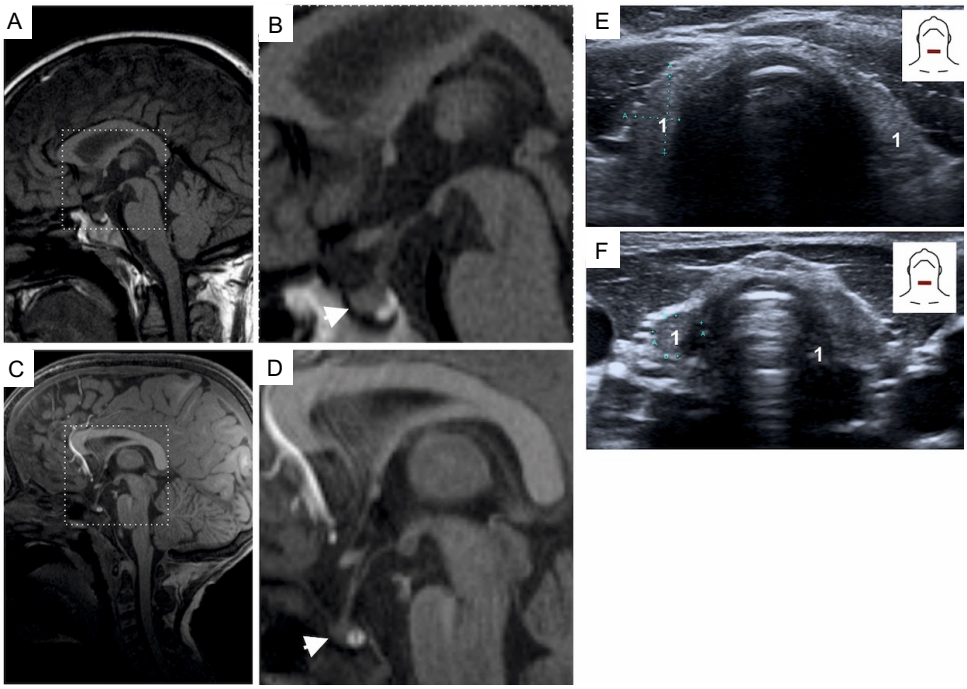


Figure 2-2: Clinical information for Family 4.

Brain MRI of individual I.1 (A, B; as a teenager) and II.1 (C-D; as a pre-teen). Thyroid ultrasound of individual I.2 (E) and II.1 (F). 1Arteria carotis communis.

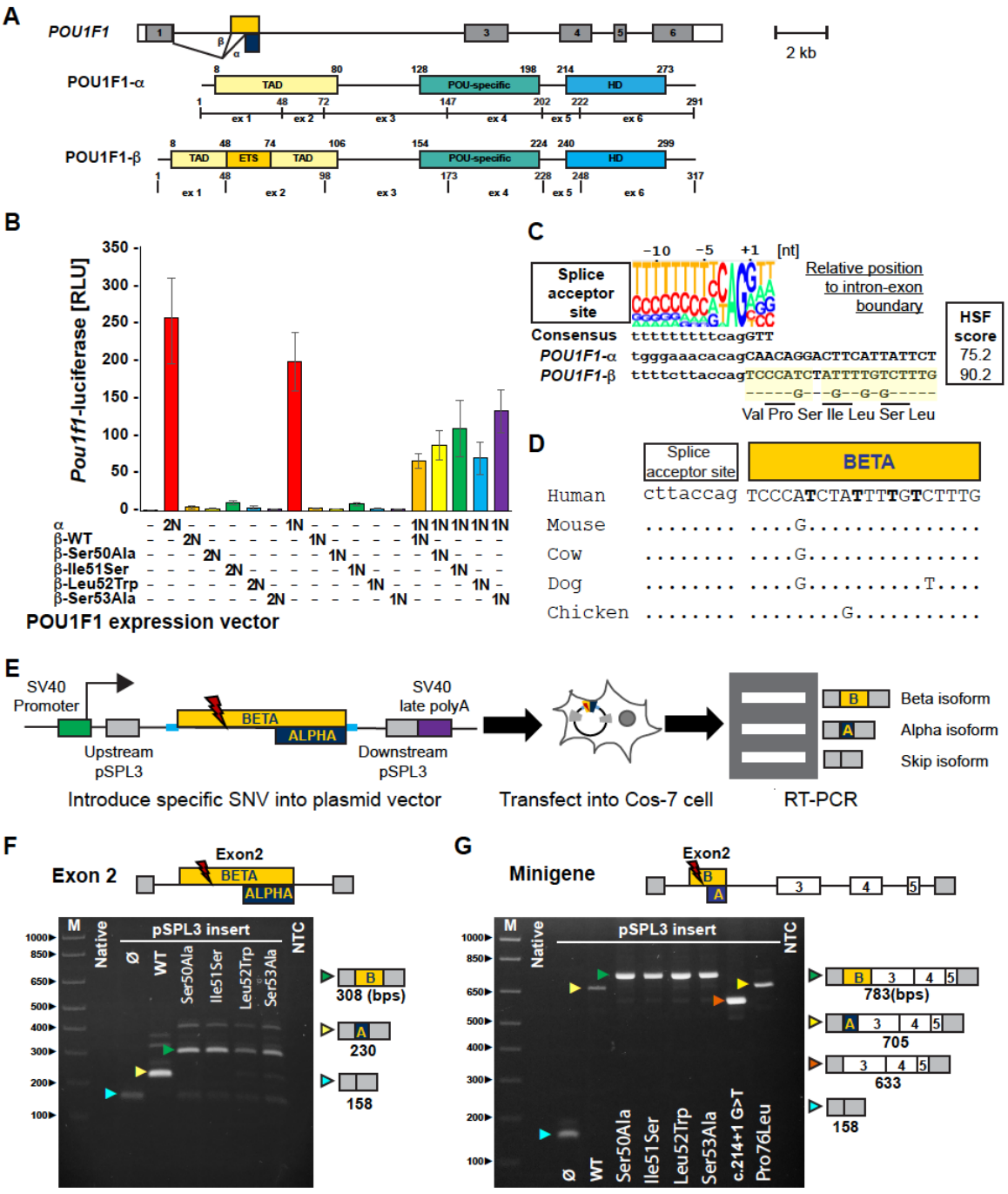


Figure 2-3: Variants in the POU1F1 beta coding region suppress the function of alpha isoform and lead to splicing abnormality.

A. Schematic of the human POU1F1 gene and protein isoforms produced by use of alternate splice acceptors at exon 2. The Pou1f1 beta isoform has an insertion of 26 amino acids located at amino acid 48 in the transactivation domain. B. COS-7 cells were transfected with a Pou1f1-luciferase reporter gene and expression vectors for POU1F1 alpha or beta isoforms either singly or together in the ratios indicated (2N and

1N). WT POU1F1 alpha has strong activation at 2N and 1N dosages. WT and variants of POU1F1 beta isoform have no significant activation over background. A 50:50 mix of alpha and WT beta isoforms exhibited reduced activation. The variant beta isoforms suppress alpha isoform mediated activation to a degree similar to WT. C. Diagram of the splice acceptor site consensus and the genomic DNA sequence at the boundary between intron 1 and splice sites utilized in exon 2 of the POU1F1 gene[139] D. Evolutionary conservation of the genomic sequence encoding POU1F1 beta isoform in mammals and chicken. E. Exon trapping assay with pSPL3 exon trap vector containing exon 2 of POU1F1 and portions of the flanking introns. F. Ethidium bromide-stained gel of exon trap products from cells transfected with the indicated plasmid. Arrowheads indicate the expected products for exon skipping (Blue), alpha isoform (Yellow), and beta isoform (green). G. POU1F1 minigenes spanning from intron 1 to 5, with all of the intervening exons, were engineered with the indicated variants and assayed for splicing. WT and p.Pro76Leu POU1F1 splice to produce the alpha isoform, the G>T change in the splice acceptor causes exon skipping (red arrow) and the other variants all splice to produce POU1F1 beta isoform.

2.4.2 Sequence variants retain POU1F1 beta isoform repressor function

We used a transient transfection assay to determine whether these variants disrupt the ability of POU1F1 to transactivate its own, highly conserved distal enhancer element[140-142] (**Figure 2-3B**). As expected, a *Pou1f1* promoter reporter was strongly activated when co-transfected with cDNA of POU1F1 alpha isoform, which does not include the variant sites. Neither WT POU1F1 beta isoform, nor any of the four missense variants found in affected individuals, showed significant activation of the *Pou1f1*-luc reporter. Consistent with a repressive role for POU1F1 beta, co-transfection with alpha at a 1:1 ratio significantly suppressed activation compared to the equivalent amount of alpha isoform alone. The four POU1F1 beta variants and WT beta repressed POU1F1 alpha activity to a similar degree.

2.4.3 Missense variants disrupt normal POU1F1 splicing to favor the beta isoform

Alpha is normally the predominant *POU1F1* isoform, comprising 97-99% of the *POU1F1* transcripts in human pituitary gland[115], but its splice acceptor is predicted to be

much weaker than the beta isoform acceptor 78 bp upstream (MaxEntScan[94]; scores, alpha: -3.63, beta: 6.96) (**Figure 2-3C**). The beta isoform splice acceptor sequence and coding region are evolutionarily conserved in mammals and birds (**Figure 2-3D**). We reasoned that splice repressor and/or enhancer sequences in *POU1F1* may dictate the normal balance of alpha over beta isoforms, and these may be disrupted by the four T>G transversions detected in individuals with hypopituitarism. To test the effect of these variants directly, we cloned *POU1F1* exon 2 beta and portions of the flanking introns into the exon trap splice reporter pSPL3 and introduced each variant by site directed mutagenesis (**Figure 2-3E**). These small minigenes were transfected into COS-7 cells, and RNA was analyzed by RT-PCR. As expected, the wild type minigene produced almost exclusively alpha isoform, while variants carried by affected individuals predominantly produced the beta isoform (**Figure 2-3F**). We tested these small minigenes in GH3 cells, a rat pituitary tumor cell line that secretes growth hormone. The results were the same, suggesting that the splicing of rat exon 2 is the same in pituitary and non-pituitary cell lines (**Figure 2-4**). This is consistent with previous studies of *Pou1f1* splicing [143]. Finally, we tested splicing with larger minigenes, which contain portions of intron 1 and intron 5 with intact exons 2, 3, 4 and 5 as well as introns 2, 3 and 4, and obtained similar results, indicating the additional sequence context does not strongly influence the observed splicing pattern (**Figure 2-3G**). We also tested two previously reported *POU1F1* variants in the longer minigene context. The c.214+1G>T caused skipping of exon 2, as expected, resulting in an in-frame POU1F1 protein that lacks 80% of the transactivation domain[144]. This variant is associated with mild hypopituitarism. The p.Pro76Leu variant is located in the transactivation domain,

enhances POU1F1 interaction with other proteins, and is associated with severe, dominant IGHD[127]. The effect of this variant on splicing had not been assessed previously, and we found that it produced predominantly alpha isoform expression, indistinguishable from wild type.

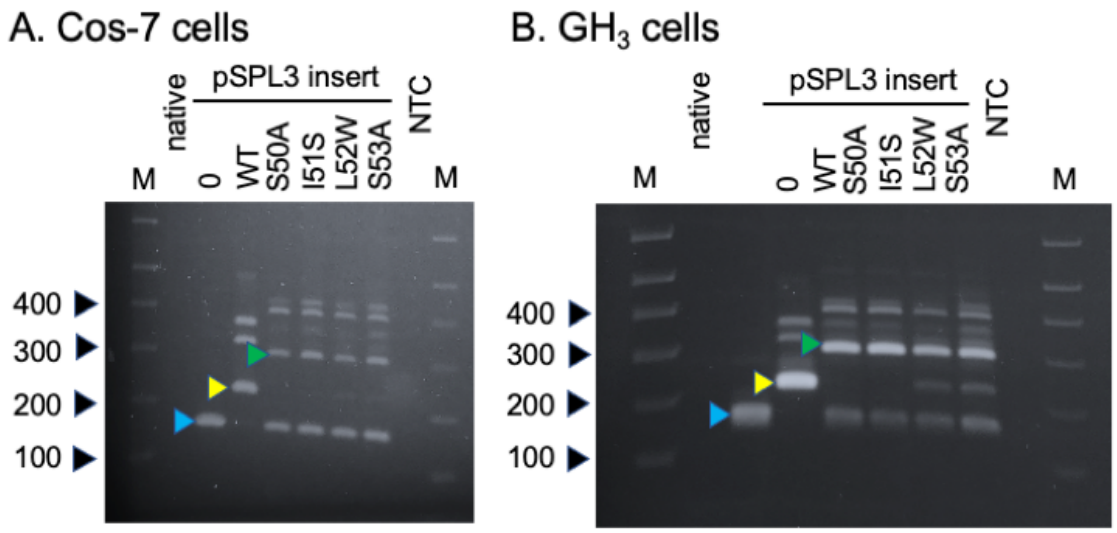


Figure 2-4: Comparison of splicing in pituitary and non-pituitary cell lines.

Cos-7 cells were transfected with the small minigene containing either wild type (WT) or the indicated variants in *POU1F1* exon 2. RT-PCR products were separated by gel electrophoresis. Arrowheads indicate the expected products for exon skipping (Blue), alpha isoform (Yellow), and beta isoform (green). B. The same experiment was conducted in GH3 cells. M = 100 bp marker ladder, native = untransfected cells, NTC = no template control for RT-PCR.

2.4.4 Saturation mutagenesis screen for splice disruptive effects

We set out to systematically identify splice disruptive variants in *POU1F1* exon 2 using a massively parallel splice reporter assay. We designed oligonucleotide pools containing every possible single nucleotide variant across exon 2 beta (150 bp) and 210 bp of the flanking introns (*N*=1080 variants), and generated libraries of this allelic series placed into the pSPL3 reporter. To track the splicing outcomes associated with each

mutation, we placed a degenerate 20mer barcode in the downstream 3' UTR. The mutant plasmid library was subjected to subassembly sequencing[132] to establish the pairing between each unique barcode and its associated *POU1F1* mutation. In total, the mutant library contained 255,023 distinct barcoded clones, among which 188,772 (74.0%) had exactly one programmed mutation. Nearly every targeted mutation appeared in this library (1070/1080, 99.1%), with a high degree of redundancy (median 75.0 distinct barcodes/mutation, **Figure 2-5**).

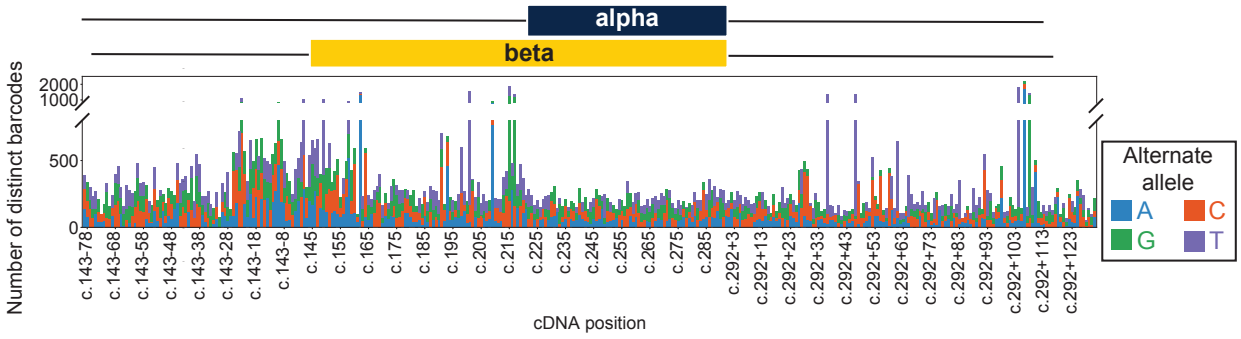


Figure 2-5: Completeness and uniformity of saturation mutagenesis.

Stacked barplot showing, for each POU1F1 variant by position (x-axis) and allele (color), the number of distinct barcodes detected in RNA-seq data (median across replicates).

The splice reporter library was transfected as a pool into COS-7 cells and processed similarly to the single mutation constructs. Spliced reporter transcripts were read out *en masse* using paired-end RNA-seq (**Figure 2-6A**), with each forward read measuring an individual splicing outcome and the paired reverse read containing the 3' UTR barcode which identifies the mutation(s) present in the primary transcript. We performed 14 biological replicates, across which 94.2% (81.8-93.4%, mean 87.4%) of barcodes associated with single nucleotide variants in the clone library were detected. As expected, alpha was the predominant *POU1F1* isoform (69.2% of reads overall),

followed by exon 2 skipping (25.6%), and beta (1.6%). We created a catch-all category ('Other') for the remaining reads (3.6%) derived from the 262 other isoforms detected. Most of those noncanonical isoforms were only scarcely used; among them, the top 20 accounted for >80% of the reads from that category. For each *POU1F1* variant, a percent spliced in (PSI) value was computed for each isoform (alpha, skip, beta, other), averaged over the associated barcodes. PSI values were highly reproducible across replicates (median pairwise Pearson's *r:* 0.92; **Figure 2-7**), and the effects measured in the pooled screen were corroborated by individual assays of 17 variants selected for validation (**Figure 2-8**).

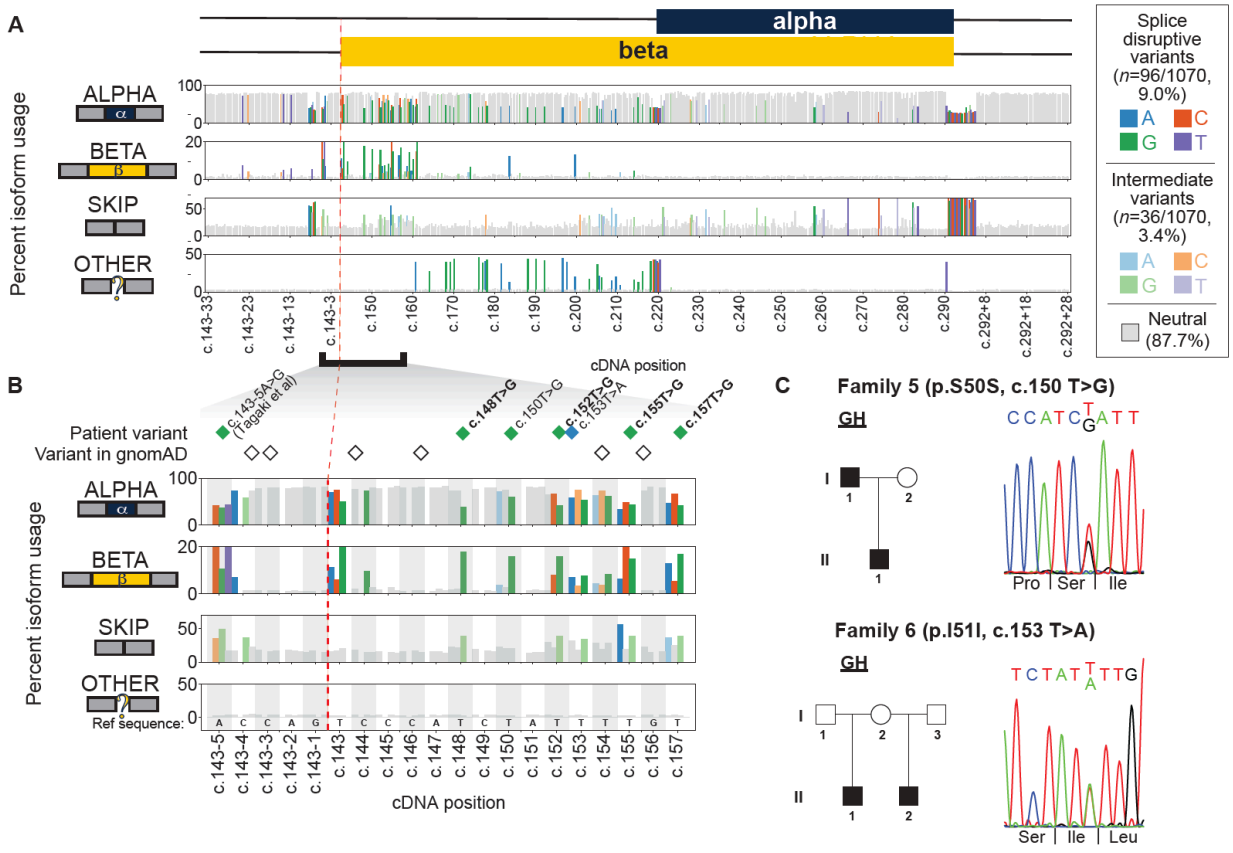


Figure 2-6: Splicing effect map in *POU1F1* exon 2 and flanking introns, and identification of IGHD families with synonymous changes.

A. Percent usage of *POU1F1* exon 2 alpha (top panel), beta (second panel), skip (third panel), and other isoforms (bottom panel) by variant position, as measured by massively parallel minigene assay. Gray bars denote splicing-neutral variants, while shaded bars indicate the base pair change of each splice disruptive variant (dark colors) and intermediate variant (light colors). Cropped intronic regions are shown in **Figure 2-9**. B. A cluster of SDVs near the beta isoform splice acceptor leads to increased usage of the beta isoform, and in some cases, intermediately increased exon skipping. Diamonds colored by the alternate allele indicate variants in individuals with hypopituitarism, and empty diamonds indicate variants reported in gnomAD. Missense variants' labels are in bold text. C. Families 5 and 6 each had two individuals affected with IGHD and synonymous variants that were splice disruptive. Pedigrees and Sanger sequence confirmation of variants are shown.

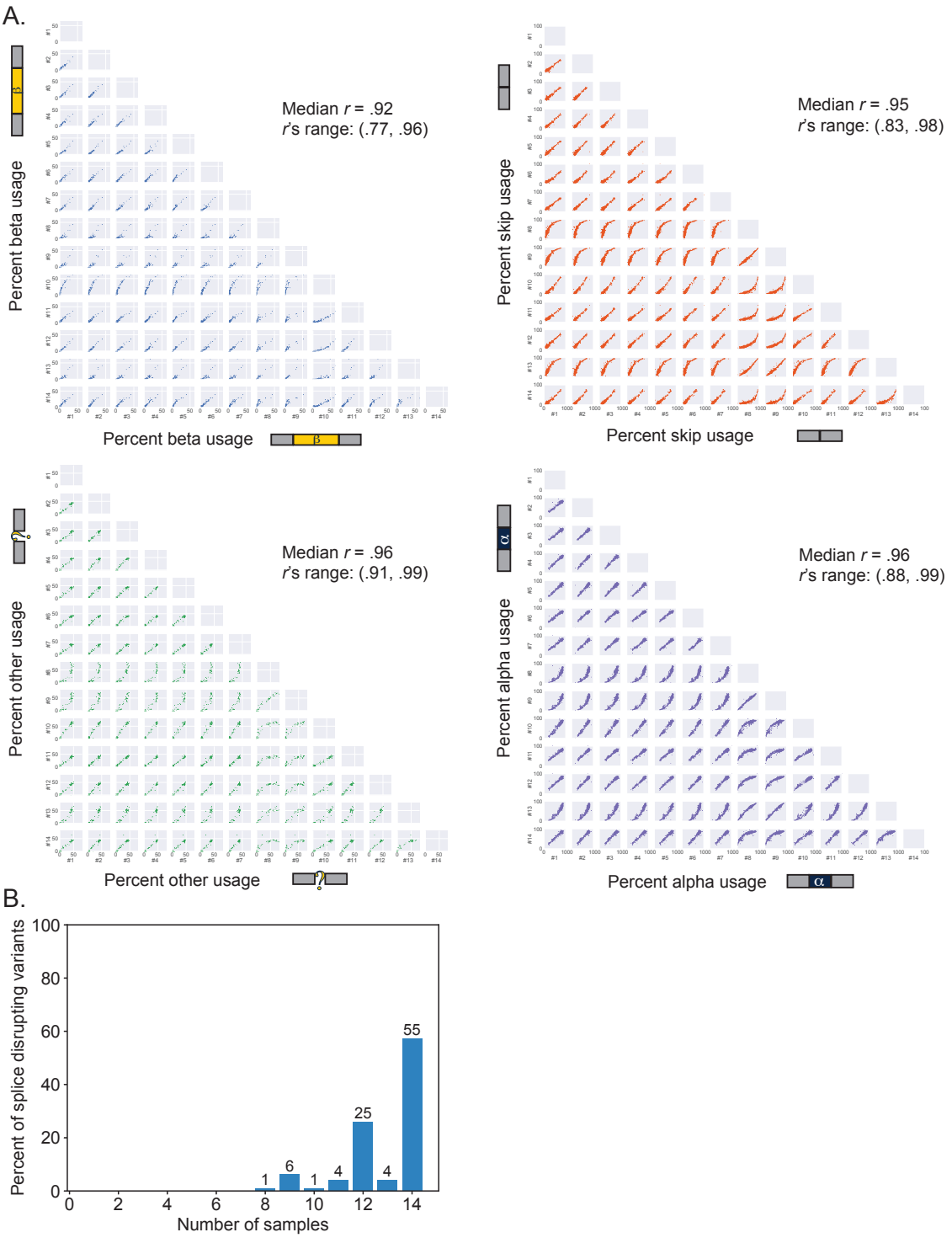


Figure 2-7: Inter-replicate correlation.

A. Pairwise scatterplots of percent isoform use for beta, skip, other, and alpha isoforms among the fourteen biological replicates. The median and range of Pearson's correlation values across samples are shown for each isoform. B. Histogram plotting the

number of replicate samples in which variants met the splice disruptive variant (SDV) criteria; all SDVs met threshold in ≥8 replicates, with 55/96 found in all 14 replicates.

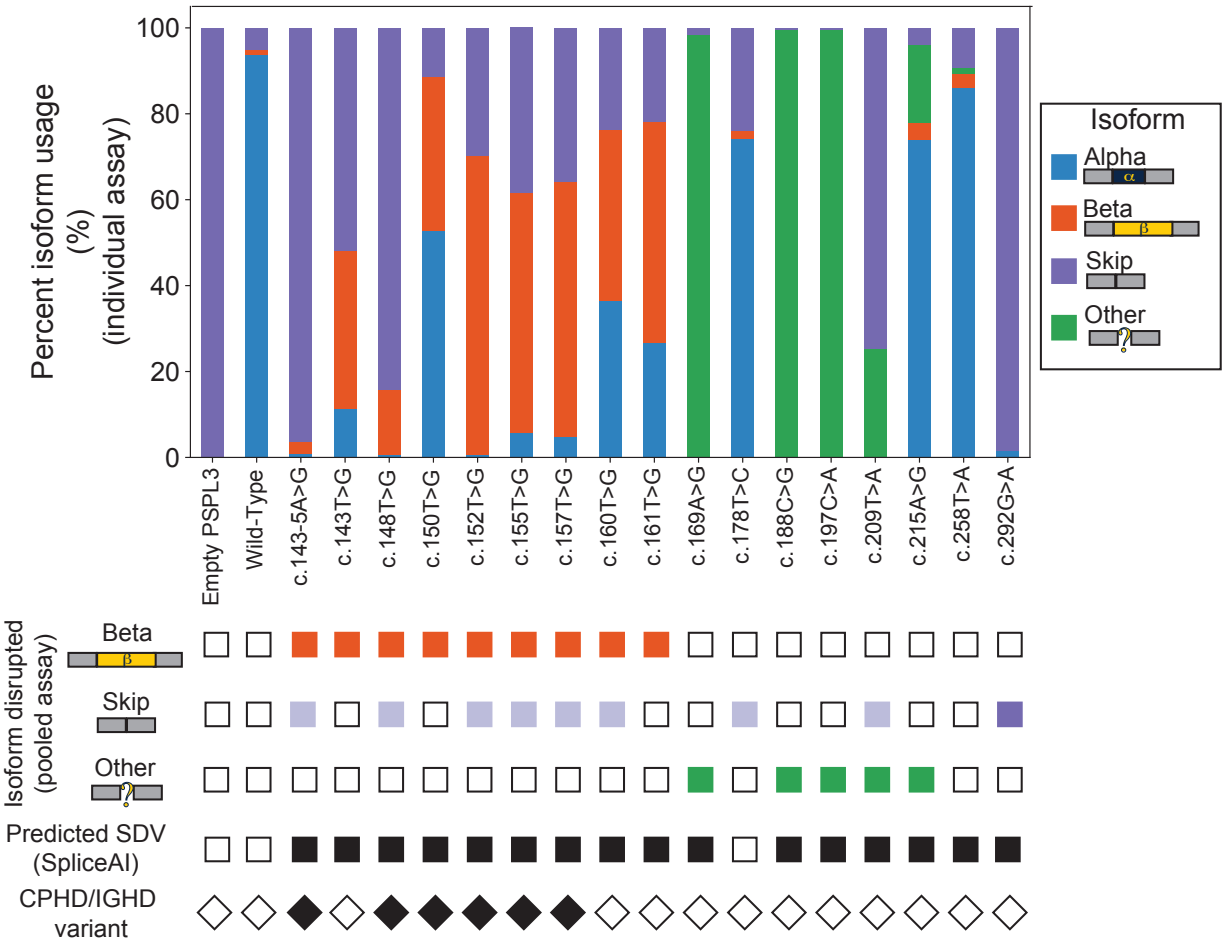


Figure 2-8: Validation by individual minigene assays.

Barplots show the proportion of isoform expression for alpha (blue), beta (orange), skip (purple) and other (green) isoforms measured by shotgun sequencing of RT-PCR products of individual mutant mini-gene transfections. Colored boxes indicate isoforms with increased (darkly colored) and intermediate (lightly colored) use, called from the pooled screen. Variants predicted as disruptive by SpliceAI[100] (squares) or seen in individuals with hypopituitarism (diamonds) are shaded in black.

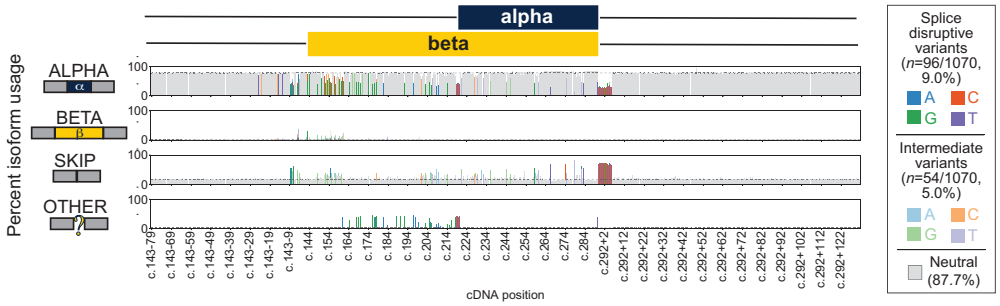


Figure 2-9: Uncropped *POU1F1* splicing effect map.

Uncropped version of **Figure 2-6A**, including cropped intronic regions lacking any splice disruptive or intermediate variants. Isoform usage and variants are plotted as in **Figure 2-6A**. Dashed black line indicates average isoform usage across null barcodes.

2.4.5 Splice disruptive variants (SDVs) across POU1F1 exon 2

We measured the impacts upon splicing of 1,070 single nucleotide variants (**Figure 2-6A** and **Figure 2-9**). Of these, 96 (9.0%) were splice disruptive variants (SDVs), which we defined as those which increased usage of beta, skip, or other isoforms by at least three-fold (Bonferroni-corrected $p<0.05$; mean observed fold-change 8.10). SDVs using other isoforms or increasing beta usage were the most frequent (n = 35/96 variants associated with each outcome) followed by those increasing exon skipping (n = 30/96), with some variants (n = 4/96) impacting usage of multiple isoforms (**Figure 2-10**). Variants leading to each outcome tended to cluster in distinct regions; notably, the beta-increasing SDVs were located near the 5' end of the beta isoform. Intronic SDVs tended to lead to skipping. A few variants that increased skipping were scattered across exon 2, and there was some enrichment in the 5' end of the beta isoform coding region, but most were enriched near splice donor and acceptor sites: 25 of 26 intronic SDVs were within +/- 20 bp of exon 2. We identified an additional 36 intermediate variants which had weaker but still significant effects (2 to 3-fold increase in beta, skip, or "other" isoforms usage; Bonferroni $p<0.05$). The majority

of these intermediate variants increased exon skipping (*n* = 22/36; 61.1%) and they clustered similarly to the SDVs associated with each isoform.

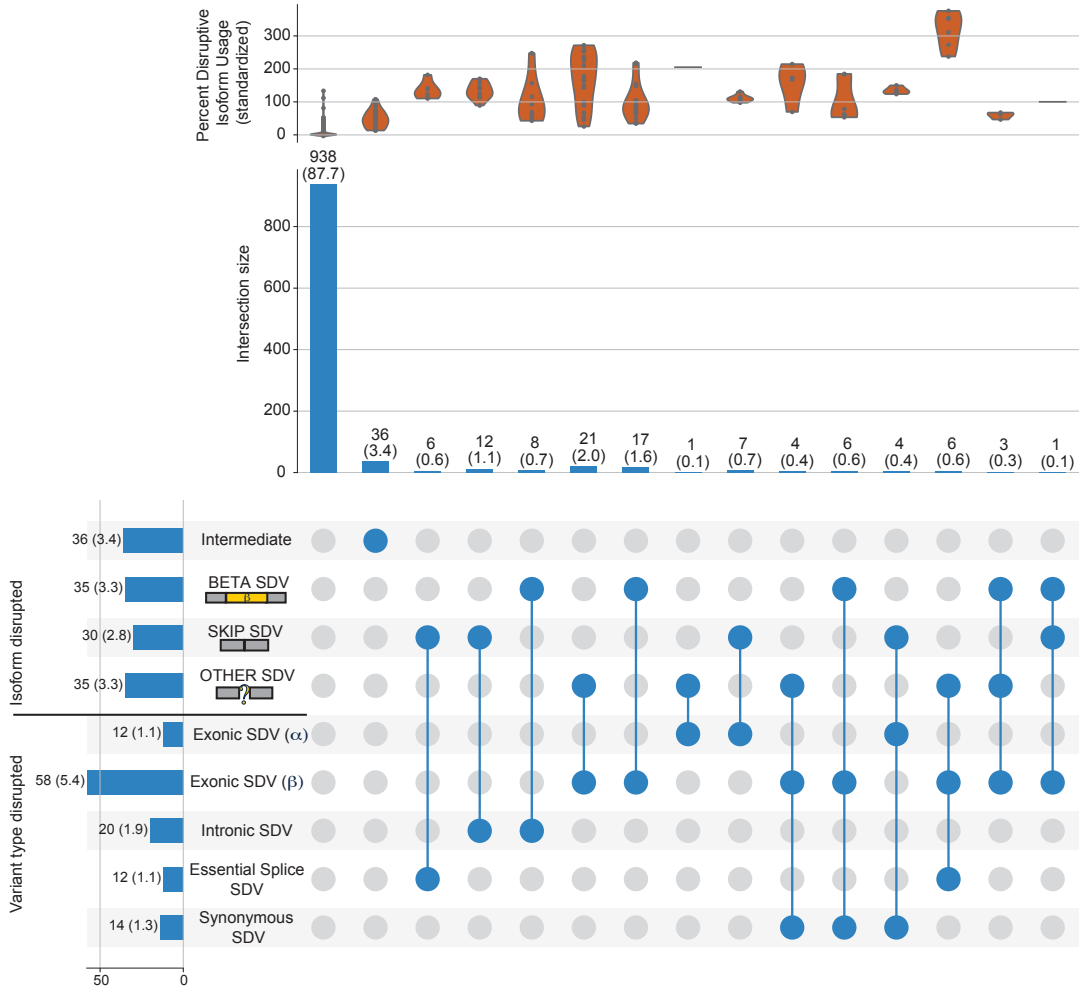


Figure 2-10: Splice disruptive variants by isoform and variant type.

Distributions of isoform usage *z*-scores for each subset of subsets are shown as violin plots. Count within each intersection (and % of total) are shown above vertical bars. Count within each subset prior to intersection (and % of total) are shown along horizontal bars. UpSet plot showing intermediate and splice disruptive variants (SDVs). SDVs are categorized by isoform (beta, skip, and other) and variant type (exonic, intronic, essential splice site, and synonymous). Filled circles denote membership in multiple categories (e.g., third column from the left indicates there are 6 essential splice site SDVs causing increased exon skipping).

We next examined the splicing isoforms in the “other” category. The associated 35 SDVs were nearly all located within the coding region unique to the beta isoform and

at the alpha isoform acceptor site (*n* = 34/35; 97.1%). Of these, most (*n*=28/34) create a cryptic acceptor AG dinucleotide that outcompetes the more distal, native alpha acceptor (**Figure 2-11**). Most of these (*n*=20/28) result in a frame-shifted transcript with a premature truncation codon predicted to result in non-sense mediated decay. In contrast, every one of the six possible variants in the native alpha acceptor "AG" dinucleotide activate a cryptic acceptor six bases downstream, leading to in-frame deletion of two codons (**Figure 2-12**). By contrast to the cryptic acceptors, of 99 SNVs creating a GT dinucleotide, only one was used as a novel splice donor, c.290:C>T located 4 bp upstream of the native exon 2 donor.

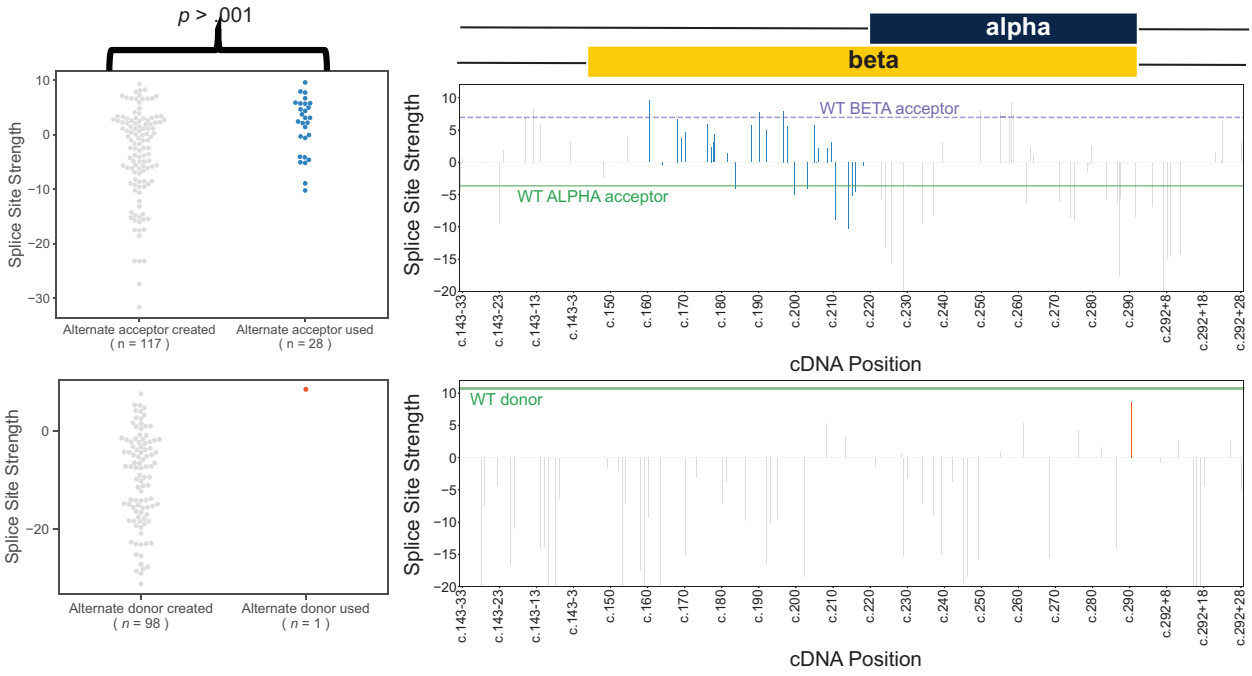


Figure 2-11: Splice site strength for novel alternate donors and acceptors.

Splice site strength as predicted by MaxEntScan[78] for novel alternate splice acceptor and donor sites. Upper: *P*-value corresponds to a *t*-test comparing the splice site strength at acceptor sites created mutation, comparing accepts not used (n = 117) vs. those used (n = 28). Dashed line (purple) represents the splice site strength of the native beta acceptor site. Solid lines (green) indicate the splice site strength of the native alpha acceptor site and native donor site respectively. Splice site strength is truncated at -20 in the positional plots, but minimum is as low as -31.6 for novel acceptors and donors within this exon.

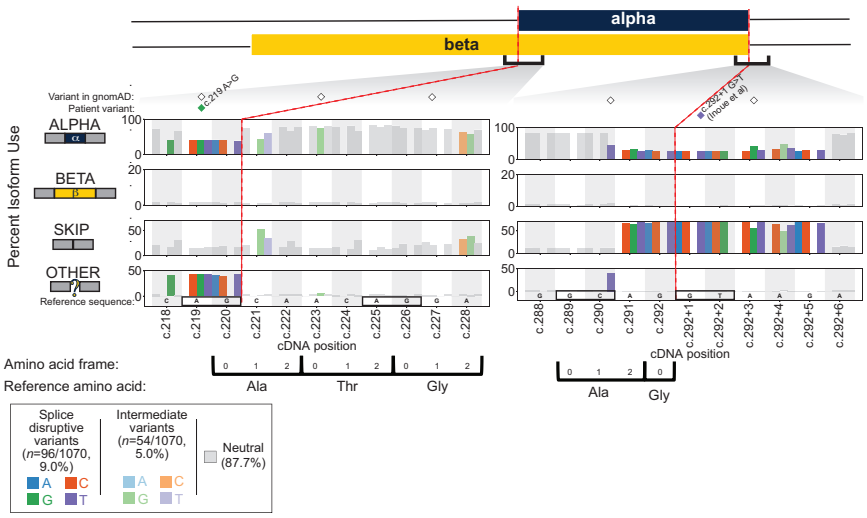


Figure 2-12: Alternate splice sites and frameshift mutations.

Detailed view of splicing effect measurements, plotted as in **Figure 2-6B**, focusing on native alpha acceptor site (left) and native donor site (right). Colored and unfilled diamonds indicate variants seen in individuals with CPHD/IGHD (colored by alternate allele) and gnomAD variants, respectively. Canonical and cryptic splice sites are boxed, red dashed lines demarcate canonical exon boundaries, and coding frame and corresponding amino acids are indicated below.

We next checked how the splicing disruption map scored the four *POU1F1* missense variants found in Families 1-4. All four showed strongly increased beta isoform usage (beta PSI increased 9.63 to 11.01-fold over background), as seen in individual minigene assays (**Figure 2-6B**). Our results also recapitulate previously described effects of two variants found in CPHD individuals: first, an upstream intronic variant c.143-5A>G[145] which led to increased beta usage and intermediately elevated skipping (**Figure 2-6C**), and an essential splice donor variant c.292+1G>T which led to near-complete skipping (**Figure 2-12**)[144].

We also examined the incidence of splice disruptive *POU1F1* variants in the general population. The gnomAD database contains 93 of the variants measured here; among those, six (6.5%) are splice disruptive and four (4.3%) are intermediate, with all being individually rare (minor allele frequency ≤ 1.6×10^{-5}; **Figure 2-13**). Overall,

variants found in gnomAD were not significantly depleted for splice disruptive/intermediate effects relative to randomly selected subsets of the tested single nucleotide variants (p=0.74 Fisher's Exact Test). Thus, *POU1F1* SDVs are tolerated to a similar extent as other predicted loss of function variants (stop gain, frameshift, splice site), which are observed throughout *POU1F1* at low frequencies in gnomAD.

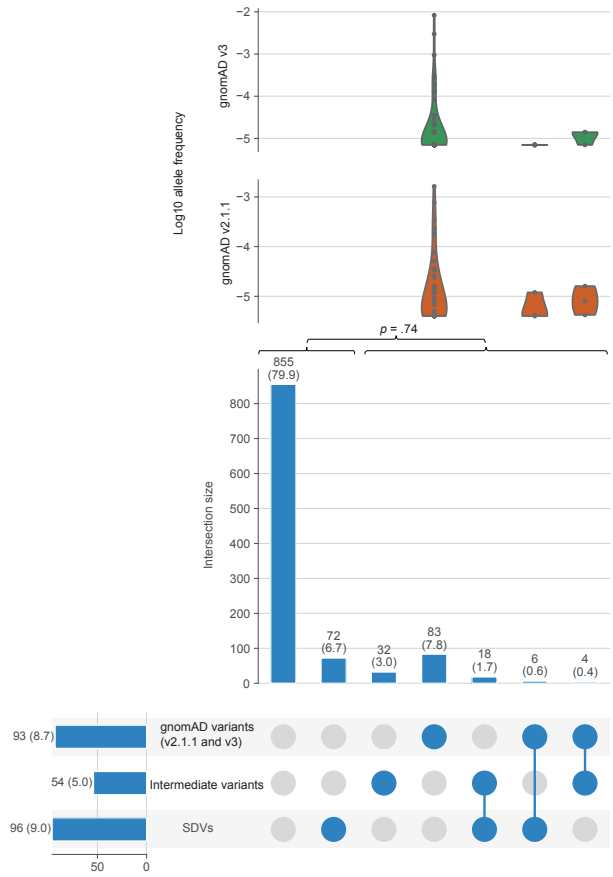


Figure 2-13: Splice disruptive variants (SDVs) in gnomAD.

Violin plots of the log10 allele frequency for each variant found in gnomAD v2.1.1 (orange) and v3 (green) within each subset are shown. Count within each intersection (and % of total) are shown above vertical bars. Count within each subset prior to intersection (and % of total) are shown along horizontal bars. *P*-value corresponds to a Fisher's exact test comparing the proportion of splice disruptive or intermediate variants between gnomAD variants and variants absent from gnomAD. UpSet plots showing intersection of neutral, intermediate, and splice disruptive variants with variants in gnomAD.

2.4.6 Additional SDVs, including silent variants, in individuals with hypopituitarism

We next examined the splicing impacts of synonymous variants, which would typically be given low priority during genetic screening due to their expected lack of coding impact. Of the 108 synonymous variants tested, 14 were splice disruptive and an additional 12 were intermediate (13.0% SDV; 11.1% intermediate; **Figure 2-13**). We identified unrelated individuals with IGHD carrying two of these synonymous SDVs in the beta isoform coding region near the 5' end of exon 2 (**Figure 2-6C**), both of which were absent in gnomAD and population-matched control databases. The first, c.150T>G (p.Ser50=), was found among an Argentinian cohort (*n*=171) in a family with two individuals with severe short stature and IGHD (**Table 2-1**), for whom WES did not reveal any likely pathogenic variants in known CPHD or IGHD genes. The index case had pituitary hypoplasia, and the individual responded well to recombinant GH treatment. The second, c.153T>A (p.Ile51=), was found in a French family in relatives with severe IGHD. The parent's DNA was not available for testing, and the parent could be an unaffected carrier or an example of gonadal mosaicism. Each of these two silent variants increased beta isoform usage to a degree similar to that of the four missense variants (beta fold change=10.7 and 4.15 for c.150T>G and c.153T>A, respectively).

2.4.7 Comparison to bioinformatic splicing effect predictions

We examined how scores from splicing effect prediction algorithms compared with these experimental measurements. We scored each single nucleotide variant in the targeted region of *POU1F1* using SpliceAI[93], MMSplice[93], SPANR[99], HAL[90] and ESRseq scores[17]. Among these, only SpliceAI predicted a high density of SDVs

specific to the exon 2 beta region surrounding the disease-causing variants (**Figure 2-14**). To benchmark each bioinformatic predictions, we took our SDV calls as a truth set and computed for each algorithm the area under the precision recall curve (**Figure 2-15**). SpliceAI was the most highly concordant with our results for both exonic variants (prAUC=0.843 vs other tools' range: 0.251-0.351) and intronic variants (prAUC=0.663 versus other tools' range: 0.549-0.585). Nevertheless, SpliceAI disagreed with our measurements for numerous variants: at the minimum threshold needed to capture all six variants seen in individuals with hypopituitarism as disruptive (SpliceAI score≥0.18), it achieved 80.2% sensitivity (*n*=19 SDVs according to the assay but not predicted by SpliceAI) and 97.3% specificity (*n*=26 variants predicted by SpliceAI but not identified by our assay) for predicting the SDVs we identified. The degree of concordance with SpliceAI was largely insensitive to the fold change threshold used to call variants as splice disruptive (**Figure 2-14**). Additional studies will be required to resolve the discordant predictions for variants observed during clinical screening.

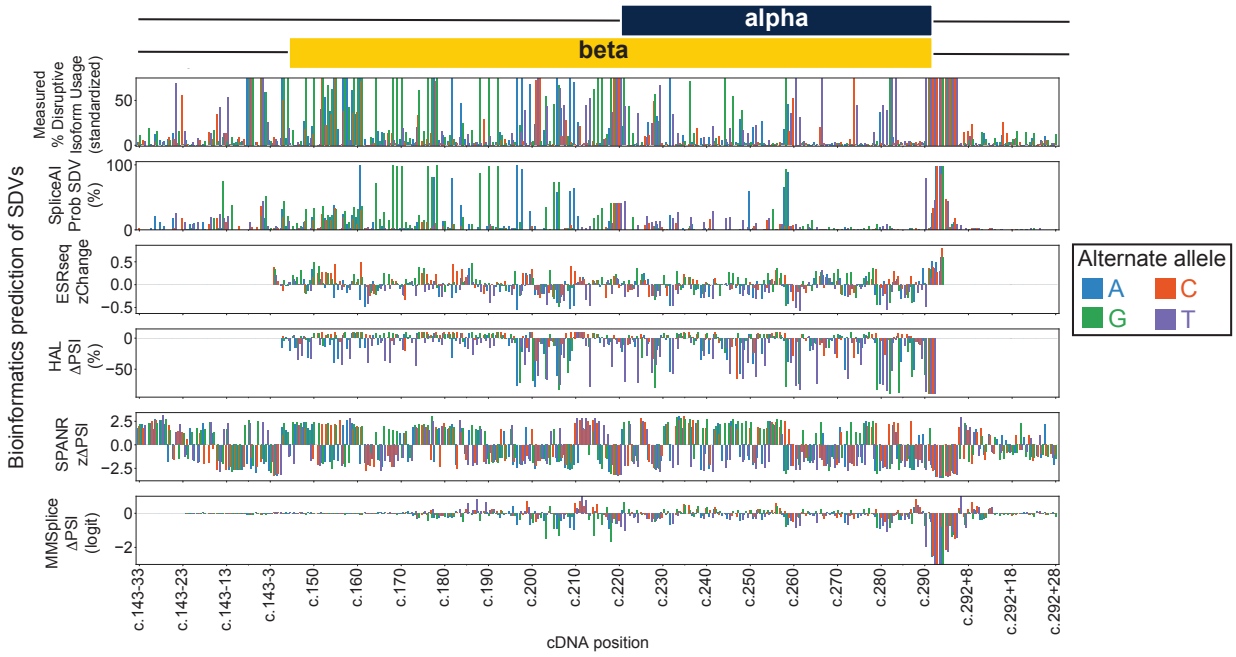


Figure 2-14: *In silico* predictions of splice disrupting variants (SDV).

Barplots showing for each variant (color) at every position (x-axis) the splicing effect measurements (top y-axis) and splice disruption as predicted by SpliceAI, ESRSeq, HAL, SPANR, and MMSplice (from second from the top to bottom y-axes).

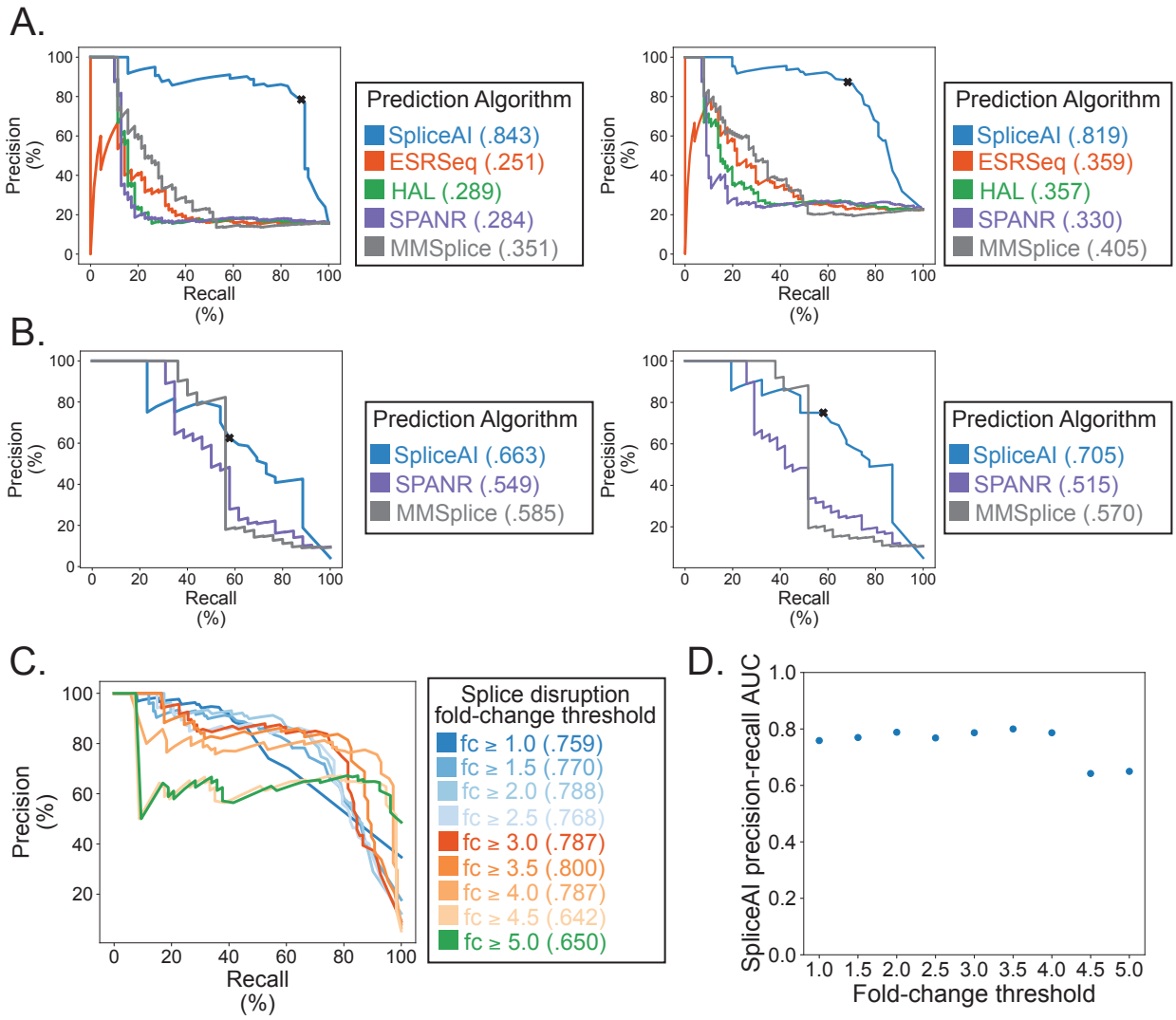


Figure 2-15: Evaluation of *in silico* splicing effect predictions.

A. Precision-recall curve showing the precision (y-axis) and recall (x-axis) of SpliceAI (blue)[100], ESRseq (orange)[17], HAL (green)[90], SPANR (purple)[99], and MMSplice (gray)[97] to predict SDVs (left) and splice disruptive or intermediate variants (right) in exonic regions. The 'x' is the at the minimum threshold where SpliceAI predicts all of the variants seen in individuals with CPHD/IGHD as disruptive (SpliceAI score ≥ 0.18). Area under the curve (prAUC) is shown within the legend B. Same as in A but for intronic variants. Since HAL and ESRseq values do not apply in noncoding regions so they are omitted from this plot. C. Precision-recall curve of the precision (y-axis) and recall (x-axis) for SpliceAI prediction of measured splice disruption across varying fold-change (fc) thresholds (range: 1 - 5) with the Bonferoni corrected *p*-value threshold held constant ($p < .05$) to call variants as disruptive. prAUC for each threshold is shown within the legend. D. Scatterplot of SpliceAI prAUC (y-axis) at varying splice disruption fold-change thresholds (x-axis).

2.5 Discussion

We found six unrelated cases with CPHD or IGHD that can be explained by variants that shift splicing to favor the repressive beta isoform POU1F1. The missense variants, p.Ser50Ala, p.Ile51Ser, p.Leu52Trp, and p.Ser53Ala, retain repressive function. They act in a dominant negative manner by suppressing the ability of the POU1F1 alpha isoform, expressed from the wild-type allele, to transactivate expression of *POU1F1* and other downstream target genes. Using saturation mutagenesis coupled to a high-throughput RNA-seq splicing readout, we systematically tested nearly every possible single nucleotide variant in or near *POU1F1* exon 2 for splice disruptive potential[136]. We identified 96 SDVs and an additional 36 intermediate SDVs which similarly activate usage of the beta isoform or cause other aberrant splicing outcomes such as exon skipping.

In addition to the four missense variants we identified initially, this screen also nominated 26 synonymous variants which were SDV or intermediately disruptive, together accounting for nearly a quarter of the possible synonymous variants in *POU1F1* exon 2. We identified two of these in unrelated families with IGHD, c.150T>G (p.Ser50=) and c.153T>A (p.Ile51=), each of which increased beta isoform usage similarly to the four missense variants that initially drew our attention. These findings underscore the need to closely examine variants for splice disruptive effects, particularly synonymous variants that could be overlooked by traditional exome sequencing filtering pipelines.

The clinical features varied amongst the six families, although they were consistent within a family. Families 1, 3, and 4 presented with CPHD, while Families 2,

5, and 6 had IGHD. Moreover, Family 4 developed cortisol deficiency. The reason for this variability in presentation is unknown. However, there are precedents for variable clinical features and incomplete penetrance with other cases of hypopituitarism[120]. Approximately 50% of IGHD progresses to CPHD, and this can even occur when the mutated gene is only expressed in GH-producing cells, i.e. *GH1*[146]. Even individuals with the same *POU1F1* mutation (i.e. p.Glu230Lys) can present with either IGHD or CPHD[147], indicating a contributing role for genetic background, epigenetic, and/or environmental factors. Both affected relatives in Family 1 had stalk disruption, a phenotype not currently associated with any other *POU1F1* variants. This feature may be due to the presence of an additional variant in *SIX3*, p.Pro74Arg, that was carried by two unaffected relatives. Heterozygous loss of function of *SIX3* is associated with incompletely penetrant and highly variable craniofacial abnormalities, including CPHD and holoprosencephaly, and there is precedent in mice for *Six3* loss of function to exacerbate the phenotype caused by mutations in other CPHD genes such as *Hesx1*[148-150].

Autosomal dominant inheritance is clear in Family 2, in which there were four affected individuals over three generations, as well as Families 4, 5, and 6. POU1F1 acts as a heterodimer[151]. Some other dominant mutations in *POU1F1* act as negative effectors due to the ability of the mutant protein to interfere with the action of the wild type protein produced from the other allele[128,152,153]. The negative effect of POU1F1 beta on the transactivation properties of POU1F1 alpha are context dependent, with differential effects on *Gh, Prl* and *Pou1f1* reporter genes[143]. The strongest effect was reported for autoregulation of POU1F1 expression via the distal, late enhancer;

dampening the auto-activation of *POU1F1* expression, and adversely affecting differentiation of the entire POU1F1 lineage and result in anterior lobe hypoplasia.

The lack of significant depletion for *POU1F1* SDVs among ostensibly healthy adult populations underscores the possibility of variable expressivity and/or penetrance for *POU1F1* splice-disruptive variants. This is consistent with the apparently unaffected parents in Families 1 and 6. A subset of these variants, like the variant c.219A>G which disrupts the alpha isoform acceptor and causes a frame-preserving two-codon deletion, may retain partial or complete function. Still others, may cause loss-of-function without dominant negative effects, and would not be expected to be strongly depleted.

In human genes, canonical splice site motifs contain less than half of the information content needed for proper splicing[16]. Additional specificity is provided by short (6-10 nt) motifs termed exonic or intronic silencers and enhancers, which are bound by RNA binding proteins that promote or antagonize splicing[154]. Although transcriptome-wide atlases have been developed to map these sites[17,61], and derive motif models[155], it often remains unclear how genetic variants impact their binding and in turn the eventual splicing output. Our splicing effect map identifies a cluster of SDVs at the 5' end of the *POU1F1* exon 2 beta, each of which increases the usage of that normally repressed isoform. These results suggest the presence of an exonic splice silencer (ESS) which may normally suppress utilization of the beta isoform acceptor. We do not expect any cell type specific factors to be involved because wild type minigene assays in pituitary cell lines and heterologous cell lines mimic the ratios of alpha:beta isoform transcripts found in normal pituitary gland [143]. We mined the cisBP-RNA database [23] and identified eight candidate motifs with strong matches to the U-rich

wild-type sequence in this region (c.143 to c.167) corresponding to known splicing factors including ELAVL1 (HuR), RALY, TIA1, and U2AF2 (**Figure 2-16**). All six hypopituitarism-associated variants replaced a U with another base (G in 5 of 6 bases), which may disrupt these motifs at high information content positions (**Figure 2-17**). Other variants predicted to disrupt these motifs tended to be beta-promoting more often than intermediate/neutral in our map ($p < 0.05$, Fisher's Exact test). These trends suggest that U-rich ESS serves to inhibit production of POU1F1 beta and this inhibition is disrupted by CPHD-associated variants, although conclusively identifying the specific cognate binding factor will require further study.

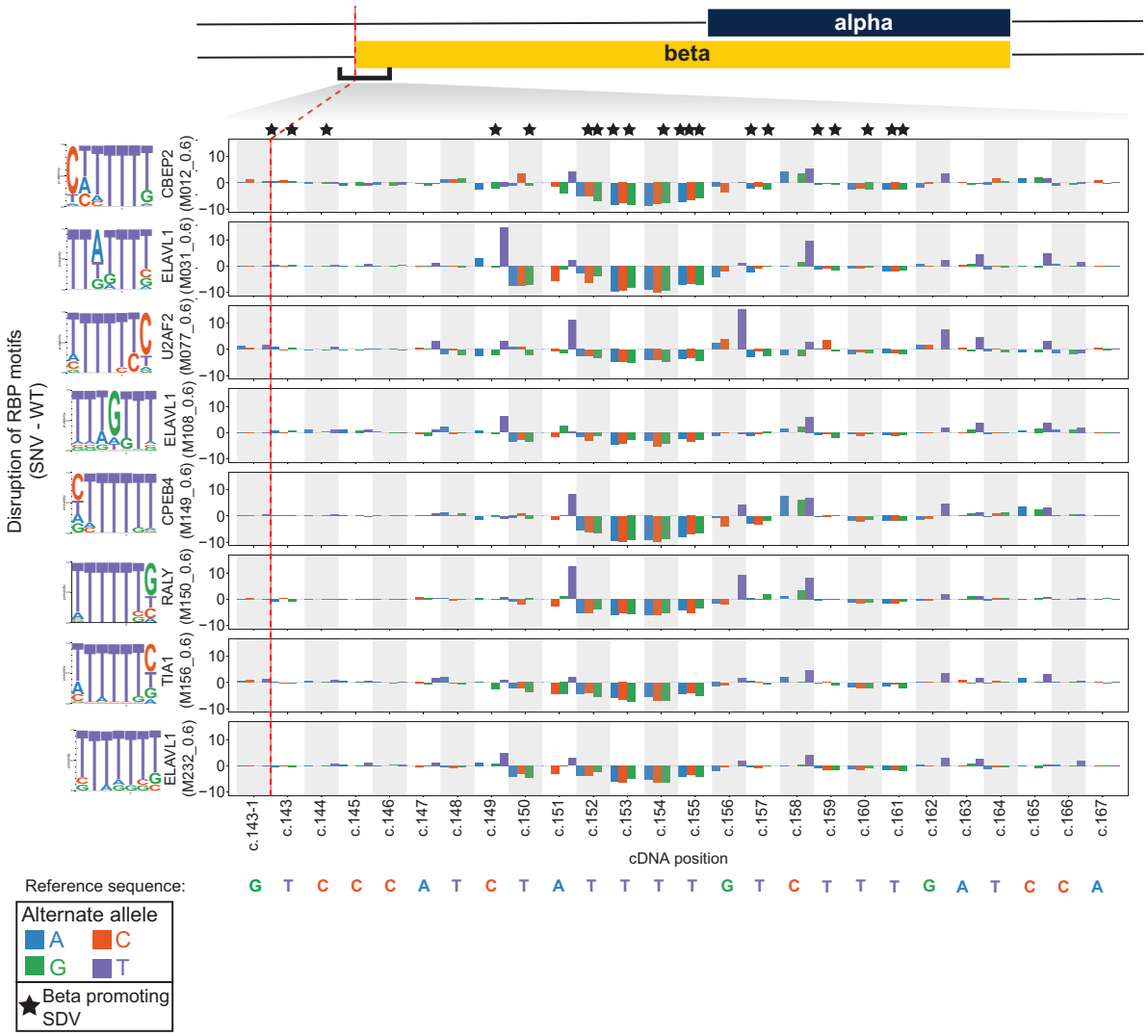


Figure 2-16: Changes in RNA binding protein motifs scores due to the SNVs in *POU1F1* beta.

Barplots show the change in maximal RNAcompete[38] kmer score (y-axis) by variant and position (x-axis), relative to the same motif scored against the wild-type *POU1F1* sequence. Black stars indicate SDVs that promote the use of the minor beta isoform.

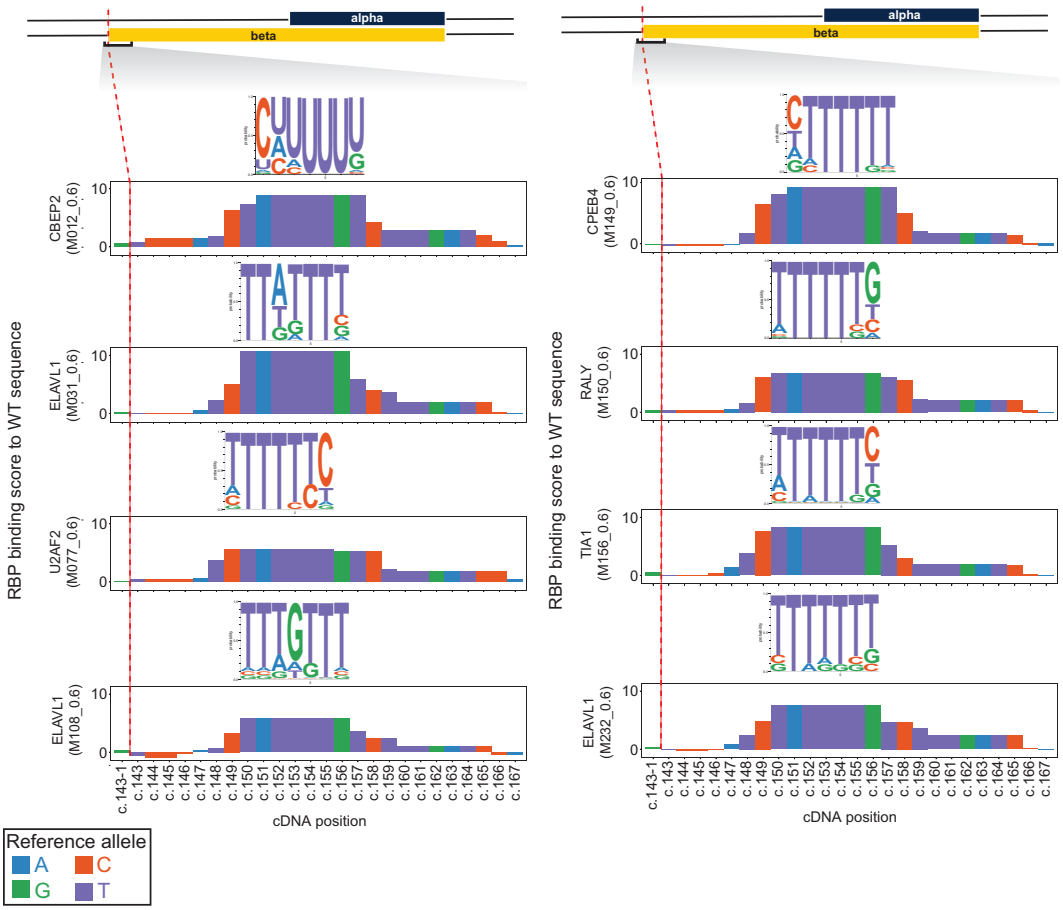


Figure 2-17: RNA binding protein (RBP) consensus binding motifs to wild-type (WT) sequence.

Barplots displaying match scores (y-axis) for selected motifs defined by RNACompete[156] scored against the wild-type *POU1F1* sequence beta region (positions c.143 to c.167).

These results extend the breadth of endocrine disorders caused by disrupted splicing. For example, in a large cohort with IGHD from Itabaianinha, Brazil, affected individuals are homozygous for a mutation in the splice donor dinucleotide (c.57+1G > A) in the growth hormone releasing hormone receptor gene (*GHRHR*)[157]. In addition, most mutations that cause dominant IGHD type II affect splicing of the growth hormone (*GH1*) gene[158]. Mutations in splice sites or splice enhancer sequences result in skipping exon 3 and production of a dominant-negative 17.5 kD isoform of growth

hormone that lacks amino acids 32-71[159]. The severity of the disease is variable and correlates inversely with the ratio of 17.5 to 20 kD GH. Finally, severe short stature associated with Laron syndrome, or GH resistance, can be caused by generation of a cryptic splice site in the GH receptor gene. Individuals from El Oro and Loja in southern Ecuador are homozygous for a p.Glu180= codon variant (GAA to GAG) that do not change the amino acid encoded but create a splice acceptor site 24 nt upstream of the normally utilized site[160]. It is notable that antisense oligonucleotide therapies hold promise for treating diseases caused by abnormal splicing, including IGHD[161,162].

Splicing disruption accounts for a significant minority of the genetic burden in endocrine disorders, as in human genetic disease more generally[1,163]. Some estimates from large-scale screens indicate that 10% of SNV within exons alter splicing, and a third of all disease associated SNVs impact splicing efficiency[36]. Variants at or near canonical spice sites are readily recognized as pathogenic[164], and these can be identified predicted with high accuracy by algorithms such as SpliceAI. However, for exonic variants, particularly those farther from exon junctions, splicing defects may be more challenging to identify bioinformatically[165-167]. Efforts to interpret these variants will need to account for the functional impacts of changing the encoded protein sequence as well as its splicing. Finally, as our results illustrate, different variants in a single gene may lead to distinct splicing outcomes with diverse consequences ranging from the straightforward loss-of-function to dominant negative effects.

2.6 Acknowledgments

The results presented in this chapter have been peer-reviewed and published[55]. I'd like to thank the other authors of this manuscript for their contributions especially my co-first authors Peter Gergics and Hironori Bando and Sally Camper who was the catalyst behind this project. I'd also like to thank those who completed the wet lab work: Peter Gergics, Hironori Bando, Mariam Maksutova, and Sajini Jayakody and the clinicians and scientists who discovered critical validation variants and analyzed the patients' whole exome/genome sequencing for this study: Ivo J.P. Arnhold, Thierry Brue, María Ines Pérez Millán, Roland Pfaeffle, Alexander A.L. Jorge, Frederic Castinetti; Frédérique Albarel, Alexandru Saveanu, Anne Barlier, Luciani Renata, Silveira Carvalho, Marilena Nakaguma, Berenice B Mendonça, Denise Rockstroh-Lippold, Julia Hoppmann, Rami Abou Jamra, Debora Braslavsky, Ana Keselman, Ignacio Bergadá, Qing Fang, A. Bilge Ozel, Qianyi Ma, Jun Z. Li, Michael H. Guo, Andrew Dauber, Sebastian Vishnopolska, Julian Martinez Mayer, and Marcelo Martí. Finally, I'd like to thank my mentor Jacob Kitzman for his guidance and assistance with analysis throughout the project, and for providing me with this wonderful collaboration to start my dissertation work. This work was supported by the National Institutes of Health (R01HD097096 to SAC, R01GM129123 to JOK), the Japan Society for Promotion of Science (HB), Grant 2013/03236-5 from the São Paulo Research Foundation (FAPESP) (IJPA), a grant from Pfizer (RP), and the Argentinean National Agency of Scientific and Technical Promotion, PICT 2016-2913 and PICT 2017-0002 (MIPM).

Chapter 3 High-Throughput Splicing Assays Identify Known and Novel WT1 Exon 9 Variants in Nephrotic Syndrome

3.1 Abstract

3.1.1 Background

Variants that disrupt mRNA splicing contribute to pathogenesis in nearly every human genetic disorder. This includes Mendelian forms of nephrotic syndrome (NS), such as Frasier Syndrome (FS), which is caused by splicing disruptive variants (SDVs) near *WT1* exon 9 splice donors resulting in decreased ratio of two natural splice isoforms, KTS+ and KTS-. However, beyond the few specific FS SDVs reported from case reports, accurately predicting other SDVs in *WT1* remains a challenge. *In vitro* splice minigene assay provides one means to test variants' splicing effect either one at a time or in highly multiplexed fashion. Therefore, we applied multiplex splice minigene assay across *WT1* exon 9 to prospectively identify *WT1* SDVs in a high-throughput manner.

3.1.2 Methods

WT1 exon 9 plus 200 bases of the flanking introns were cloned into an established minigene plasmid, in between constant synthetic exons. Large scale mutagenesis was performed to generate a variant library including every single nucleotide variant across the cloned region, each associated with a unique "barcode" sequence in the constant downstream exon. This variant library was then transfected into HEK293T and COS7 cells with multiple replicates. RNA was harvested after 24 hours of transfection and

spliced transcripts from the minigene library were analyzed by target RNA-seq. The splicing patterns associated with each variant were quantified from the aligned reads.

3.1.3 Results

Nearly every possible single nucleotide variant was represented (518/519; 99.8%) with a high degree of internal replication (mean=79.7 barcodes per variant). The splicing disruption was heavily concentrated near the canonical splice sites, the alternate KTS+ and KTS- donors. We successfully identified 8 known FS variants dramatically lowered log2 (KTS+/KTS-) to -2.1 or lower, and 16 additional SDVs which disrupted KTS+/KTS- comparably to the known FS variants. We also identified 19 variants that increased KTS+/KTS-, with two have been observed in patients with disorder of sex development (DSD).

3.1.4 Conclusions

The pooled minigene assay is highly sensitive and specific for identification of pathogenic *WT1* exon 9 splice disruption. Our multiplex screen identifies all known FS SDVs in *WT1* exon 9. We also nominate an additional 16 possible yet unseen FS variants with similarly decreased KTS+/KTS-. A set of variants significantly increase KTS+ expression, which might be related to DSD. In summary, multiplex functional analyses can prospectively score genetic variants in NS and guide the clinical decision.

3.2 Introduction

Variants that disrupt proper pre-mRNA splicing contribute a share of the pathogenic burden in Mendelian forms of nephrotic syndrome (NS). One NS gene sensitive to splicing disruption is WT1, which encodes a key genitourinary transcription

factor essential for podocyte development and integrity. Its disruption results in a phenotypic spectrum including isolated NS, syndromic NS with tumors and gonadal dysgenesis, differences of sexual development, Wilms Tumor and leukemia[72].

WT1 undergoes alternative splicing, including in exon 9 at a pair of donors which result in protein isoforms that differ by the three amino acids, KTS. These isoforms have overlapping and district functional roles: both act as transcription factors[75] but with partially different sequence motif specificities and gene targets[168-171]. Normally, they are expressed in the mature kidney[73,76] at a ~2:1 ratio (KTS+:KTS-). Variants which disrupt the KTS+ donor reduce this ratio and cause an extremely rare syndrome called Frasier Syndrome (FS), consisting of male gonadal dysgenesis, NS, Wilms tumor or gonadoblastoma (**Figure 3-1**)[73,74,76].

Figure 3-1: Frasier's syndrome and WT1 exon 9.
WT1 exon 9 alternative splice forms KTS+ (blue) and KTS- (yellow).

Currently, eight variants downstream of the KTS+ splice donor are known to cause FS or focal segmental glomerulosclerosis (FSGS)[74,76,172-175]. It remains unclear if other nearby variants are similarly splice disruptive. Accurate computational prediction of variants' splicing effects remains challenging, so we devised a massively parallel splice assay[55] to measure the splicing effects of every possible single nucleotide variant (SNV) in or near *WT1* exon 9 and the flanking introns (N=519 variants). This identified all eight known FS/FSGS variants and nominated an additional 49 *WT1* SNVs as splice disruptive, including two patient variants with uncertain interpretations (**Table 3-1**). This

splicing effect map can support clinical interpretation of novel *WT1* variants and improve the accuracy of genetic diagnosis.

Table 3-1: Splice assay scores for previously reported pathogenic variants for Frasier Syndrome, focal segmental glomerulosclerosis, or 46,XX OTDSD.

Variant	Genomic position (hg19)	Splice score log2(KTS+/KTS-)	Called splice disruptive?	ClinVar Interpretation	Literature report as pathogenic	Variant consequence
c.1437A>G	32413528	2.76	YES	Conflicting interpretations	YES	synonymous
c.1447+1G>A	32413517	-2.81	YES		YES	intronic
c.1447+1G>C	32413517	-2.56	YES		YES	intronic
c.1447+2T>C	32413516	-2.64	YES	Likely pathogenic	YES	intronic
c.1447+3G>A	32413515	1.72	YES	Uncertain significance	this study	intronic
c.1447+3G>T	32413515	-2.71	YES		YES	intronic
c.1447+4C>T	32413514	-2.21	YES	Pathogenic/Likely pathogenic	YES	intronic
c.1447+5G>A	32413513	-2.51	YES	Pathogenic	YES	intronic
c.1447+5G>T	32413513	-2.69	YES		YES	intronic
c.1447+6T>A	32413512	-2.58	YES	Pathogenic	YES	intronic

3.3 Materials and methods

3.3.1 Cell culture

HEK293T and COS-7 cells were obtained from American Type Culture Collection (ATCC) and cultured in Dulbecco's modified Eagle's medium with high glucose, L-glutamine and sodium pyruvate (DMEM; GIBCO, Grand Island, NY, USA) containing 10% fetal bovine serum and 1% penicillin-streptomycin (10,000 U/mL) (GIBCO). Media was checked monthly for mycoplasma contamination by PCR.

3.3.2 Saturation mutagenesis library construction

A *WT1* minigene construct was prepared by cloning a fragment with *WT1* exon 9 plus 414 bp of its flanking introns into the vector pSPL3 (Invitrogen, Carlsbad, CA) at the

BamHI site, an intronic context flanked by a synthetic first and last exon. This construct was subjected to saturation mutagenesis as previously described[55], targeting the full exon +/- 40 bp (173 bp). Briefly, a mutant oligonucleotide pool was designed in which each position across the targeted region was successively replaced by three other mutant bases. The pool was synthesized by Twist Biosciences (South San Francisco, CA), amplified by limited-cycle PCR, and cloned by HiFi Assembly (New England Biolabs, Ipswitch, MA) into the vector backbone linearized by inverse PCR.

3.3.3 Mutant plasmid barcoding

A library of random 20mer barcode sequences were synthesized by IDT (Coralville, IA) and cloned into the downstream 3' UTR at the MscI site by HiFi assembly. The resulting pools of mutant *WT1* exon 9 minigenes with barcodes were transformed by electroporation into NEB 10-b E. coli, reaching library complexity of hundreds of barcoded clones per designed mutation. To enumerate the 3'UTR barcodes and identify the specific variant paired with each barcode, subassembly sequencing libraries were generated as previously described[55,132].

3.3.4 Minigene library transfection

Mutant minigene libraries were transiently transfected into HEK293T (8 replicates) and COS-7 cells (2 replicates). At 24 hours post-transfection, cells were lysed by addition of Trizol and total RNA was purified with Direct-zol RNA Miniprep Kits (Zymo Research, Irvine, CA). A total of 3-5 ug total RNA was reverse transcribed using SuperScript III First-Strand Synthesis kit (Invitrogen) with oligo(dT)20 primer following the manufacturer's protocol. Afterwards, spliced transcript was amplified via semi-

nested PCR using outer primer pairs, first SD6 forward (5'-TCTGAGTCACCTGGACAACC-3') and SA2 reverse (5'-ATCTCAGTGGTATTTGTGAGC-3'), and inner primer pairs, JKlab232 (5'-AGTGAACTGCACTGTGACAAGCTGC) and SA2 reverse. Indexed Illumina sequencing adaptors were added by PCR and the resulting RNA-seq libraries were submitted for paired-end 150-bp sequencing on Illumina HiSeq or NovaSeq instruments.

3.3.5 RNA-seq processing and splice disruption calling

RNA-seq reads were processed as previously described[55]. Briefly, reads containing plasmid barcodes were selected with cutadapt, barcodes were clustered with starcode[176], and filtered to retain only those associated with a single-base variant. The paired, splice-informative read was aligned to the reference minigene sequence with the splice-aware read aligner STAR[177]. Custom python scripts (https://github.com/kitzmanlab/wt1_splice) were used to identify the isoform corresponding to each read: KTS+ (42.6% of all reads), KTS- (37.4%), exon 9 skipping ('SKIP', 19.1%), or all other isoforms ('OTHER'; collectively, <1% of all reads). The count of reads matching each isoform was tallied per barcode, then aggregated into a per-variant, per-isoform percent by taking a read-count weighted mean of the respective percentages across the associated barcodes.

To test the significance of splice disruption, we created for each variant a null distribution by bootstrap sampling a matching number of barcodes associated with intronic variants >10 bp outside the exon boundaries. Using this null distribution, we computed z scores for the observed per-isoform usage, then used Stouffer's method to aggregate z scores across replicates. Splice disruptive variants (SDVs) were defined as

those that were (a) significant at the p<0.05 level (after Bonferroni correction for multiple testing), and either (b) had either SKIP or OTHER usage at least 20% higher than the null or (c) an isoform log-ratio (calculated as log2(KTS+/KTS-)) of >=1.5 or <= -1. Variants were defined as intermediate if they (a) passed the same significance test and had either (b) SKIP or OTHER usage at least 10% higher than the null or (c) an isoform log-ratio of >=1 or <= -.5. Results were highly correlated across replicates; all SDVs were also called disruptive at least half of the replicates when processed individually.

3.3.6 Prediction of splice site strength

MaxEntScan scores[78] for variants at the common *WT1* exon 9 acceptor, KTS- donor, and KTS+ donor were computed using the maxentpy python module (https://github.com/kepbod/maxentpy). We first computed the splice site strength for the wild-type and mutant sequences for each and took the signed difference between the variant and wild-type scores.

3.3.7 Data availability

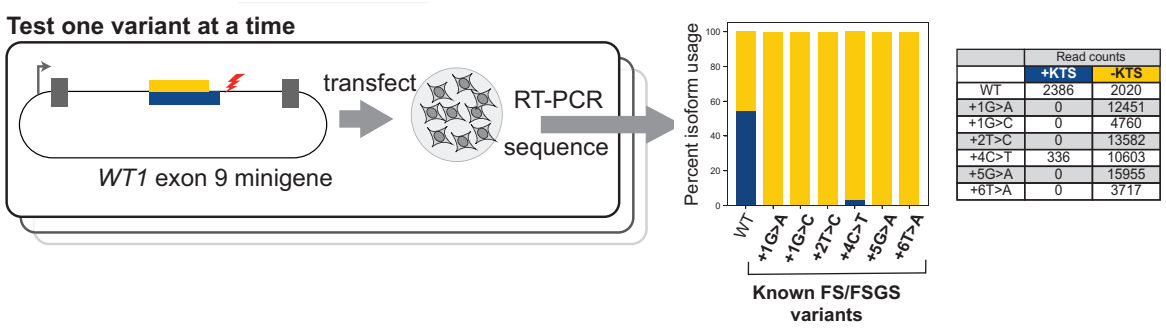
Custom python scripts are available at https://github.com/kitzmanlab/wt1_splice and a look up table of splicing effects will be available within the final publication and are currently online at Zenodo[136].

3.4 Results

3.4.1 Massively parallel splicing assay for WT1 exon 9

To systematically identify splice disruptive variants, we established a minigene assay with *WT1* exon 9 and flanking introns (~200 bp on either side). We first

individually tested the wild-type sequence and six pathogenic variants near the KTS+ donor known to cause FS or FSGS. The wild-type construct showed a roughly even balance between the two isoforms (1:1.2 KTS+:KTS- ratio), while each of the known pathogenic variants abolished KTS+ usage (**Figure 3-2**; **Figure 3-3**). Thus, consistent with previous reports[178], minigenes can faithfully model *WT*1 splicing defects associated with FS and FSGS.



	Read counts	
	+KTS	-KTS
WT	2386	2020
+1G>A	0	12451
+1G>C	0	4760
+2T>C	0	13582
+4C>T	336	10603
+5G>A	0	15955
+6T>A	0	3717

Figure 3-2: Variants in individuals with Frasier's syndrome alter the KTS ratio.

Six known Frasier or Focal segmental glomerulosclerosis syndrome (FS/FSGS) variants tested individually by minigene assays followed by sequencing, with the percent of spliced reads from each isoform shown.

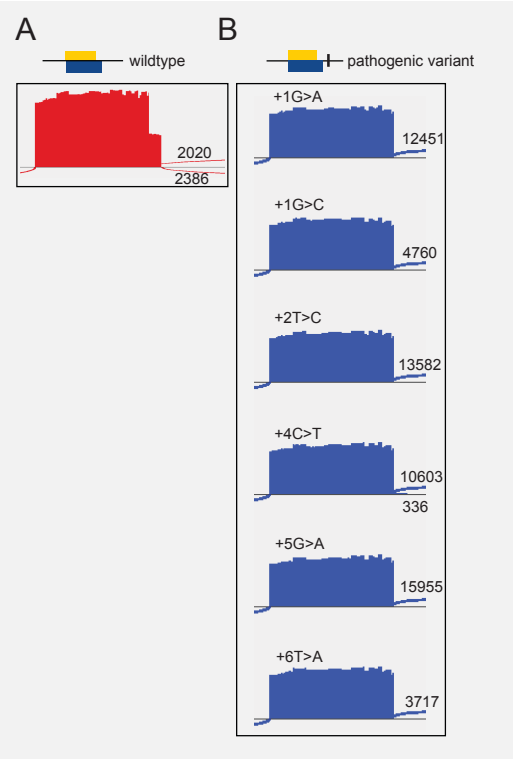


Figure 3-3: Altered splicing for Frasier's syndrome variants in minigene assay.

Sashimi plot from IGV showing read pileup and splice junction read counts from deep sequencing of RT-PCR products from individual minigene assays of **(A)** wildtype *WT1* exon 9 and flanking introns and **(B)** six mutant constructs each containing a different known FS/FSGS SDV. KTS- read counts are shown above each track, and KTS+ read counts (when present) are shown beneath each.

We next set out to test the splicing effects of every possible single nucleotide variant (SNV) in and around *WT1* exon 9 (**Figure 3-4A**). We applied saturation mutagenesis to create a library of all possible SNVs in the exon and for 40 bp into each flanking intron. The mutant library was tagged with random 20mers in the 3'UTR to serve as barcodes allowing for the splicing effect of each mutation to be tracked. Nearly every possible SNV was represented (518/519; 99.8%) with a high degree of internal replication (mean=79.7 barcodes per variant; **Figure 3-5**).

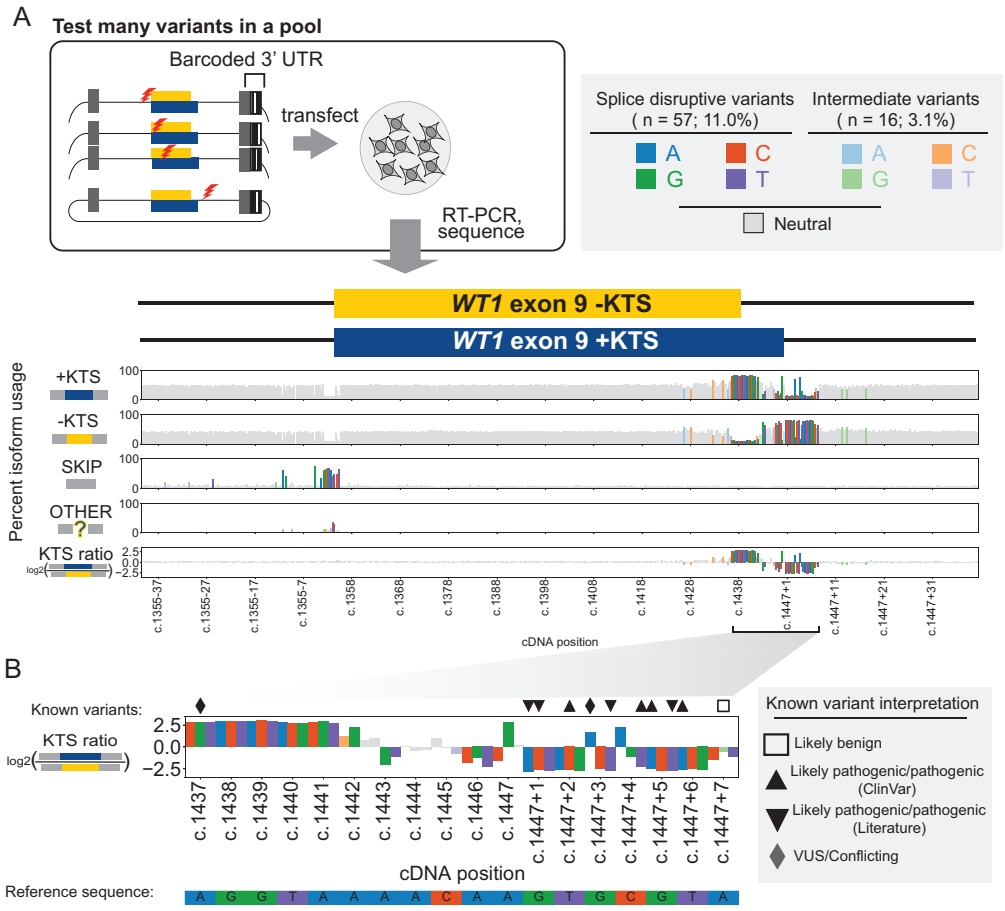


Figure 3-4: Screening for all possible splice disruptive variants in *WT1* exon 9.

A. Splicing effect map for all 518 single-nucleotide variants in/around *WT1* exon 9 from a massively parallel splice assay. Each bar represents a single variant plotted by its cDNA position (x-axis), with dark shading for splice disruptive variants, light shading for intermediate ones, and gray for variants with no effect upon splicing. The first four tracks show the percent usage of KTS+, KTS-, SKIP, and OTHER isoforms. Final y-axis track shows the log2(KTS+/KTS-) ratio. **B.** Zoom to the alternative donors showing KTS+/KTS- ratio and reference sequence. Known variants are shown above the plot with symbols denoting existing interpretation.

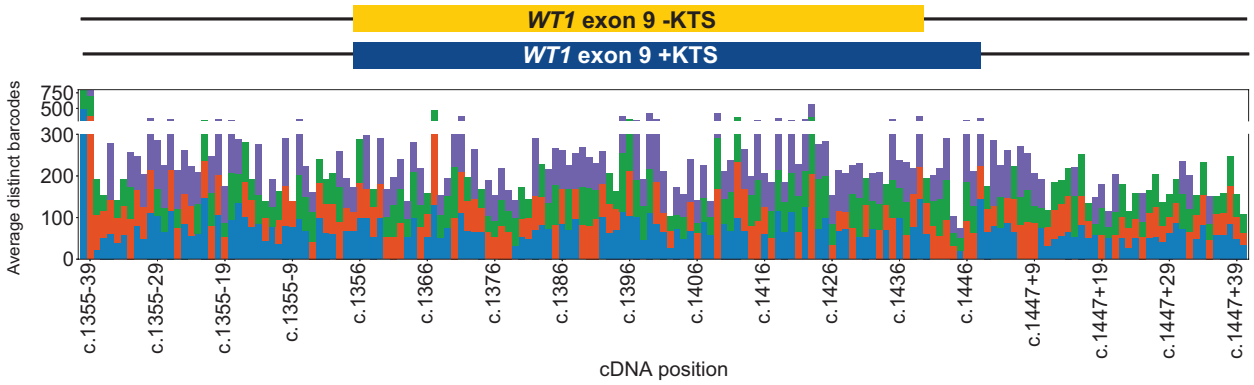


Figure 3-5: Completeness and uniformity of saturation mutagenesis.

Distinct barcode counts for each mutation (mean across replicates) are shown by position. Each shows the three different nucleotide substitutions per position.

We transfected HEK293T cells with the mutant minigene library pool and deeply sequenced the resulting spliced RNAs to quantify, for each mutation, the use of the KTS+ and KTS- isoforms, exon skipping ('SKIP') or all other splicing outcomes ('OTHER'). Mutations' effects upon isoform usage were reproducible within the HEK293T biological replicates (median pairwise Pearson's r=.94) and between HEK293T replicates and a second cell line, COS-7 (median between cell line pairwise Pearson's r=.93; **Figure 3-6**).

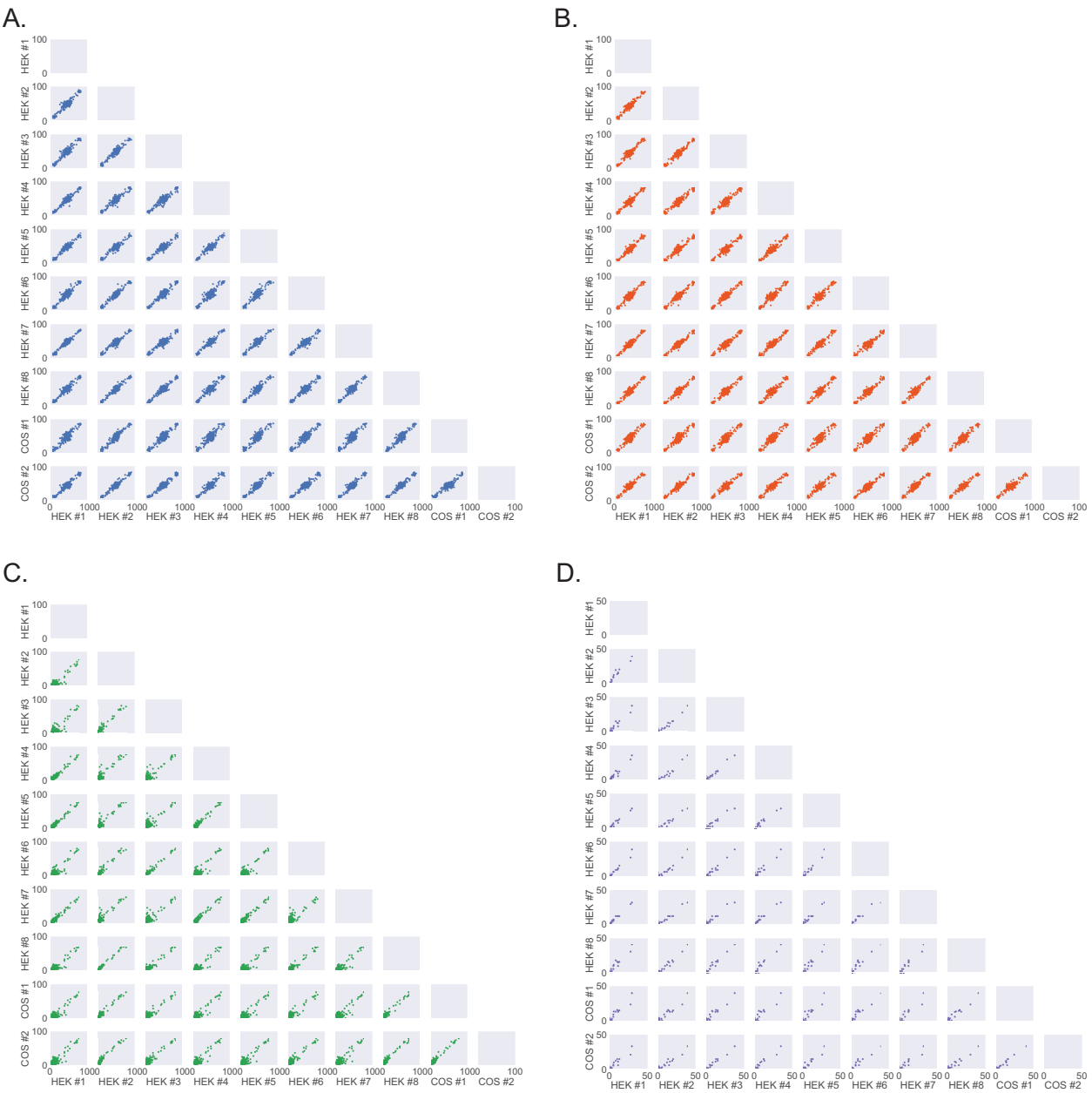


Figure 3-6: Correlations among replicates and across cell lines for each measured isoform.

A. KTS+ isoform usage for each replicate - different cell-types are indicated. Here, each point represents an individual variant. **B.** KTS- isoform usage for each replicate. **C.** SKIP isoform usage for each replicate. **D.** OTHER isoform usage for each replicate.

The resulting map shows that, as expected, sensitivity to splicing disruption is heavily concentrated near the canonical splice sites, in particular the alternate KTS+ and KTS- donors (**Figure 3-4A**). Overall, of the 518 measured SNVs, only 57 (11.0%)

altered splicing with an additional 16 (3.1%) having an intermediate effect on splicing[136](**Table 3-1**). Of the disruptive variants, all but one were near (+/- 15 bp) either the splice acceptor or one of the donors, consistent with disruption of those sites' consensus motifs. The primary disruptive effect for most variants (43/57; 75.4%) was to alter the KTS+/KTS- ratio, roughly evenly split between shifting the balance towards KTS- and KTS+ (24 variants and 19 variants, respectively). A minority of variants led to complete exon skipping (n=14) or activated a cryptic acceptor 17 bp downstream of the native one (n=2), each of which would yield frameshifted transcript predicted to undergo nonsense mediated decay.

3.4.2 Identification of known and novel variants disrupting KTS+ usage

We focused first on the eight known FS/FSGS variants as described in the ClinVar database or published case reports[74,76,172-175]. All eight dramatically lowered the KTS+:KTS- balance, as quantified by log2(KTS+/KTS-) scores of -2.21 or lower **(Figure 3-4B, Figure 3-7; Table 3-1**). By contrast, our assay scored as neutral all but one of the 19 variants listed in ClinVar with an interpretation of Likely Benign, as well as an additional 17 variants observed in the population database gnomAD[136] (log2 ratio scores between -0.11 and 0.53; **Figure 3-7**). The lone exception was c.1447+7A>G, listed in ClinVar as Likely Benign, for which we noted a very subtle shift towards KTS- (score: -0.52) for which the *in vivo* impact is unclear. Thus, pooled minigene assays effectively discriminate between known pathogenic splice disruptive variants and neutral polymorphisms.

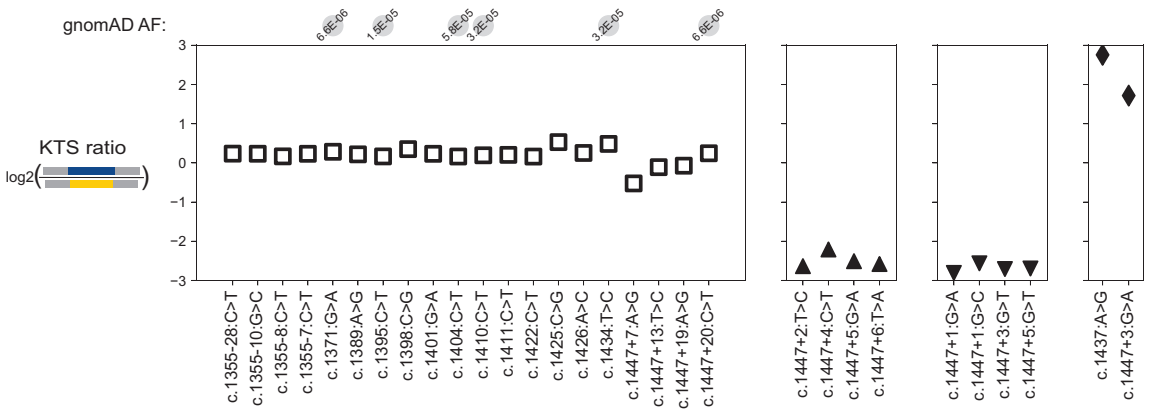


Figure 3-7: Splicing variants in ClinVar and gnomAD.

KTS+/KTS- ratios for variants reported in the literature or in ClinVar, grouped by interpretation, with population allele frequency shown above the plot for variants present in gnomAD.

We next asked whether this map could prospectively identify as-yet unreported variants which disrupt KTS+. We identified sixteen additional SDVs which disrupted KTS+ comparably to the known FS/FSGS variants[136] (median log2 ratio score: -2.14), with all but two corroborated by MaxEntScan splice site strength predictions[78] (median MaxEntScan score=-2.40; **Figure 3-8**). Specifically, eleven of the KTS+ disruptive variants were predicted by MaxEntScan to weaken the KTS+ donor (maximum MaxEntScan score=-1.03), and one variant was predicted to strengthen the KTS- donor allowing the KTS- splice site to outcompete the KTS+ donor (MaxEntScan score=2.18). Two measured splicing effects were discordant with MaxEntScan scores: one variant was predicted to mildly increase KTS+ strength (c.1447:A>C, MaxEntScan score=0.71) and MaxEntScan anticipated another KTS+ disruptive variant to have little impact on the strength of the KTS- donor (c.1443:A>T, MaxEntScan score=-0.04; KTS+ donor was out of MaxEntScan scoring range for this variant). The final two novel variants impacting KTS+ donor use were outside the MaxEntScan scoring range (c.1447+7:A>C and A>T). Among the addition KTS+ disruptive variants, six were located within the codon region

specific to the KTS+ isoform; four of these were synonymous variants, which as a class may be overlooked during classification. None of these is yet deposited in ClinVar nor in published reports, but based upon their disruptiveness in this assay, they represent potential novel pathogenic variants.

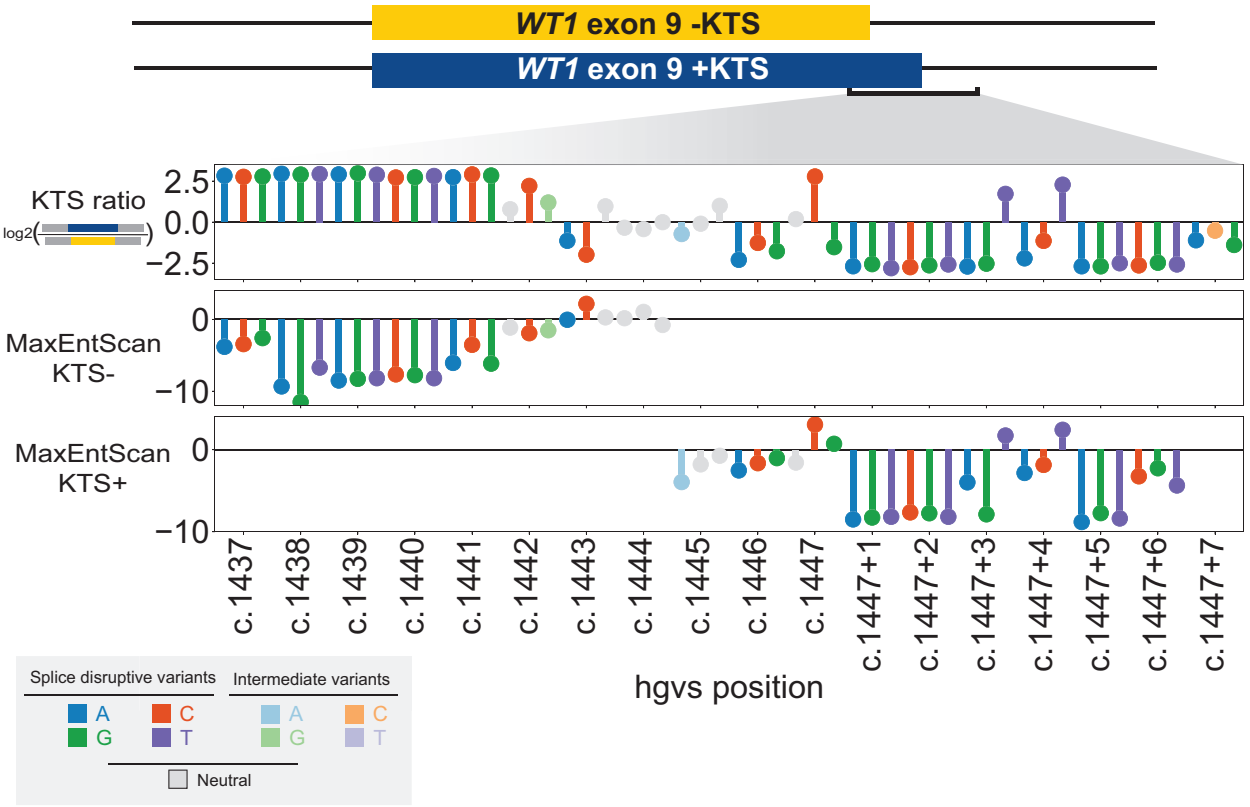


Figure 3-8: MaxEntScan predictions of splice site strength.

Measured splicing scores for variants surrounding the two KTS donors (top) and MaxEntScan predictions of splice site strength for the KTS- (middle) and KTS+ donors. Each lollipop represents a single variant plotted by its cDNA position (x-axis), with dark shading for variants altering the KTS ratio, light shading for intermediate ones, and gray for variants with no effect upon the KTS donors. Variants outside of the MaxEntScan scoring range have no associated lollipop within the MaxEntScan tracks.

3.4.3 Other splice disruptive outcomes

Finally, we searched this map for variants that disrupt splicing in other ways and identified nineteen variants that increased the KTS+/- ratio, opposite the effect associated with FS (**Figure 3-4B, Figure 3-7**). Notably, two have been observed in

patients with 46,XX ovotesticular differences in sexual development (46,XX OTDSD). The first, c.1437A>G, is a synonymous variant recently reported[179] as a *de novo* mutation in a patient with 46,XX OTDSD. Consistent with the strong shift towards KTS+ in our map (score: 2.76), it is predicted by MaxEntScan to weaken the KTS- donor two bases downstream (MaxEntScan score = -3.43; **Figure 3-8**). We observed similar effects from an additional 17 variants near the KTS- donor (median log2 ratio score: 2.83, range: 2.20-2.98). Another variant, c.1447+3G>A (log2 ratio score: 1.72), downstream of the KTS+ donor, was observed during clinical exome sequencing in a 12-year old proband with 46,XX OTDSD. Notably, no renal abnormalities or a history of Wilms' tumor was reported for either of these two individuals. In contrast to the variants near the KTS- donor, this and one other variant not yet observed clinically (c.1447+4C>A) are predicted to strengthen the KTS+ donor (MaxEntScan scores = 1.72 and 2.42 respectively; **Figure 3-8**), possibly leading it to outcompete its upstream counterpart. Finally, we identified a cluster of 14 variants within 26 bp of the *WT1* acceptor which led to complete skipping of exon 9, or use of an alternate cryptic acceptors, in each case leading a frameshift and premature truncation[136]. None of those variants are yet reported in ClinVar or population databases.

3.5 Discussion

Here, we applied a massively parallel splicing assay to systematically test the effects of every single-nucleotide variant in and near *WT1* exon 9, a hotspot of pathogenic variants for multiple genetic forms of nephrotic syndrome including Frasier Syndrome. The resulting splicing effect map correctly identified all seven known FS variants[74,76,172,173,175] as associated with reduced KTS+/KTS- ratio, along with another

variant reported to cause isolated steroid resistant NS and FSGS[174]. By contrast, all but one of the 36 tested variants which appear in the ClinVar database with a Likely benign interpretation, or the general healthy population database gnomAD, scored as splice-neutral, indicating this assay is highly sensitive and specific for identification of pathogenic *WT1* exon 9 splice disruption.

FS is extremely rare: in all, fewer than 200 cases have been reported, represented by only seven distinct FS variants[74,76,172,173,175]. Two of these seven variants (at the +4 and +5 positions) account for most of the FS case reports to date: taking the count of ClinVar submissions as a proxy for frequency, c.1447+4C>T and c.1447+5G>A together had 24 records, compared with only two records combined across the five other known FS variants. The two recurrent variants overlap the only CpG dinucleotide in the KTS+ region, and their frequency is likely explained by the ~10-fold higher *de novo* mutation rate at germline-methylated CpGs[180]. Thus, it is reasonable to expect that there may be a tail of additional variants which are rare even within the context of this rare disorder. Indeed, our results implicate an additional sixteen SNVs as similarly decreasing the KTS+/KTS- ratio, and nominate these as new, as yet-unreported variants with the potential to cause FS/FSGS.

Our map also identified nineteen variants which increase the KTS+/KTS- ratio, either by weakening the KTS- donor or strengthening the KTS+ donor. One of these variants was previously reported[179] in an individual with 46,XX OTDSD, and we report an additional, unrelated patient with a similar presentation carrying a different variant. Our results are consistent with these variants acting to shift the balance of WT1

isoforms toward KTS+, which is known to activate *SRY*, the master regulator of male sex determination[181].

In some Mendelian disorders, missense or synonymous variants may alter splicing by disrupting regulatory elements beyond the canonical splice sites, termed exonic splice enhancers and silencers[182]. Such effects have been observed by other systematic splice assays[55] including at other *WT1* exons[60]. Here, though, we observed variants to the interior of *WT1* exon 9 had little impact upon its splicing. This suggests that *WT1* exon 9 either does not depend on exonic splice regulatory elements for its definition, or that any such elements may be robust to perturbation by single nucleotide variants.

These results may be useful in interpreting variants found in individuals who do not display every feature of FS. For instance, variants disrupting KTS+ in karyotypically female individuals (46,XX) may lead to progressive glomerulopathy, but due to the lack of gonadal dysgenesis, FS may not be suspected and *WT1* genetic testing might not be pursued[183]. In conclusion, our systematic screen provides a lookup table of splice disruptive variants in *WT1* exon 9, circumventing the need for single variant minigene studies. The availability of functional evidence for newly observed rare variants can facilitate their resolution, lessening the burden of variant interpretation upon clinicians, and shortening the diagnostic odyssey for NS patients.

3.6 Acknowledgments

The results presented in this chapter are available as a pre-print[56] and are under review at *Kidney International Reports*. I'd like to thank the other authors of this manuscript for their contributions especially Jennifer Lai Yee who spearheaded this

work. I'd also like to thank Bala Burugula and Ian Dunn who both completed the wet lab experiments presented here and Swaroop Aradyha and Ana Morales at Invitae for their clinical contributions. Finally, I'd like to thank my mentor Jacob Kitzman for his guidance and assistance with analysis throughout the project. This work was supported by the National Institutes of Health (R01GM129123 to JOK) and the Clinical Scientist Institutional Career Development Program Award (5K12HD028820-28 to JLY).

Chapter 4 Benchmarking Splice Variant Prediction Algorithms Using Massively Parallel Splice Assays

4.1 Abstract

4.1.1 Background

Variants that disrupt mRNA splicing account for a sizable fraction of the pathogenic burden in many genetic disorders but identifying splice-disruptive variants (SDVs) beyond the canonical donor and acceptor dinucleotides remains difficult. Computational predictors are often discordant, compounding the challenge of variant interpretation. Because these tools are primarily validated using clinical variants, which are heavily biased to canonical splice site mutations, it remains unclear how well their performance generalizes to other variants.

4.1.2 Results

We benchmarked eight widely used splicing effect prediction algorithms, leveraging massively parallel splicing assays (MPSAs) as a source of experimentally determined ground-truth. MPSAs simultaneously assay many variants to nominate candidate SDVs. Across MPSAs of five genes, we compared experimentally measured splicing outcomes with bioinformatic predictions at 3,616 variants. Algorithms' concordance with MPSA measurements, and with each other, was lower for exonic vs intronic variants, underscoring the difficulty of identifying missense or synonymous SDVs. Deep learning-based predictors (SpliceAI, Pangolin) trained on gene model annotations

achieved the best overall performance at distinguishing disruptive and neutral variants (area under precision recall curve of .845 and .855 respectively). Controlling for overall call rate genome-wide, SpliceAI and Pangolin also showed superior overall sensitivity for identifying SDVs. Finally, our results highlight two practical considerations when scoring variants genome-wide: finding an optimal score cutoff, and the substantial variability introduced by differences in gene model annotation, and we suggest strategies for optimal splice effect prediction in the face of these issues.

4.1.3 Conclusion

SpliceAI and Pangolin showed the best overall performance among predictors tested, however, improvements in splice effect prediction are still needed especially within exons.

4.2 Introduction

Splicing is the process by which introns are removed during mRNA maturation using sequence information encoded in the primary transcript. Sequence variants which disrupt splicing contribute to the allelic spectrum of many human genetic disorders, and it is estimated that overall as many as 1 in 3 disease-associated single-nucleotide variants are splice-disruptive[36,37,184-187]. Splice-disruptive variants (SDVs) are most readily recognized at the essential splice site dinucleotides (GU/AG for U2-type introns), with many examples across Mendelian disorders[41,43-46]. SDVs can also occur at several so-called flanking noncanonical positions[164], which by some estimates outnumber essential splice mutations by several-fold[61,185].

Variants beyond the splice-site motifs may be similarly disruptive but are more challenging to recognize[188]. For instance, some SDVs disrupt splicing enhancers or silencers, short motifs bound by splicing factors to stimulate or suppress nearby splice sites, to confer additional specificity and to provide for regulated alternative splicing[26]. These elements are widespread[189] and maintained by purifying selection[27], but their grammar is often unclear as they feature partial redundancy and tolerate some mutations. Nevertheless, variants which disrupt splicing regulatory elements have been implicated in a number of disorders. A prominent example is in spinal muscular atrophy, in which loss of *SMN1* cannot be fully complemented by its nearly identical paralog *SMN2* due to the loss of an ESE in exon 7 of the latter gene[154,190], a defect which can be therapeutically targeted by antisense oligonucleotides nearby[191]. Synonymous variants, which as a class may be overlooked, may also disrupt existing splice regulatory elements or introduce new ones, as in the case of the X-linked parkinsonism gene, *ATP6PA2*[192].

RNA analysis from patient specimens can provide strong evidence for splice-disruptive variants, and its inclusion in clinical genetic testing can improve diagnostic yield[163,185,193-195]. However, advance knowledge of the affected gene is necessary for targeted RT-PCR analysis and RNA-seq-based tests are not yet widespread[196,197], and both rely upon sufficient expression in the blood or other clinically accessible tissues for detection. Therefore, a need remains for reliable *in silico* prediction of SDVs during genetic testing, and a diverse array of algorithms have been developed to this end. For instance, S-Cap[98] and SQUIRLS[91] implement classifiers that use features such as motif models of splice sites, kmer scores for splice regulatory elements, and evolutionary

sequence conservation trained on sets of benign and pathogenic clinical variants. Numerous recent algorithms use deep learning approaches to predict splice sites' probabilities directly from the primary sequence; SDVs can then be detected by comparing predictions for wild-type and mutant sequence. Rather than training with clinical variant sets, SpliceAI[100] and Pangolin[101] are trained using gene model annotations to label each genomic position as true/false based on whether it appears as an acceptor or donor in a known transcript. SPANR[99] uses the primary sequence to predict percent spliced in (PSI) measurements with training data provided by RNA-seq data. HAL[90] takes a distinct approach by training on a library of randomized sequences and their experimentally observed splicing patterns, while MMSplice[97] combines the training data from HAL with features derived from primary sequence with additional modules trained on annotated splice sites and clinical variants. Finally, ConSpliceML[105] is a metaclassifier that combines SQUIRLS and SpliceAI scores with a population-based constraint metric which measures the regional depletion of predicted splice-disruptive variants among apparently healthy adults in population databases.

Given the proliferation of splicing predictors and their utility in variant interpretation, it is important to understand their performance characteristics. Previous comparisons have suggested that overall SpliceAI represents the state-of-the art with several other algorithms including Pangolin, MMSplice, SQUIRLS, and ConSpliceML showing competitive or in some cases better performance[91,101-108,198]. However, benchmarking efforts to date primarily relied upon curated sets of clinical variants[91,102-107,198], which are strongly enriched for canonical splice site mutations[99,102,107,199-201], likely reflecting the relative ease of their classification. This leaves open the question of

how these tools' performance may generalize, as well as whether certain tools may excel in particular contexts (e.g., for exonic cryptic splice activating mutations). A further challenge is that some of these tools' training data may partially overlap with these benchmarking validation sets, which risks circularity if overlapping variants are not carefully identified and removed.

Massively parallel splicing assays (MPSAs) provide an opportunity to benchmark splicing effect predictors entirely orthogonally to clinical and population variant sets. MPSAs measure up to thousands of variants' splicing effects in a pooled fashion: cells are transfected with a library of variants cloned into a minigene construct with deep RNA sequencing as a quantitative readout of variants' splicing outcomes. MPSAs come in several different flavors: broad MPSA screens assess many exons and measure one or a few variants' effects at each[36,61,62,202], while saturation screens focus on individual exons[19,48,55-60,63] or motifs[17,90] and measure the effects of every possible point variant within each target. MPSAs of short motifs and broad MPSAs have been used to train algorithms[90,97,99] and as features in other bioinformatic splice prediction tools[91]. However, since MPSAs of short motifs could be dependent on the surrounding, fixed exonic and intronic sequence they may not reflect the actions of the motifs within different exons or across variant types so they could be problematic as training sets or machine learning inputs. Two broad MPSA datasets, Vex-seq[62] and MaPSy[36] were recently used to benchmark splicing effect predictors as part of the Critical Assessment of Genome Interpretation (CAGI) competition[203], and another, MFASS[61] has been used to validate a recent meta-predictor[108]. However, a limitation of benchmarking with broad MPSAs is that they may reflect an exon's overall properties while lacking the finer

resolution to assess different variants within it. For instance, an algorithm could perform well by predicting SDVs within exons with weak splice sites, or with evolutionarily conserved sequence, while failing to distinguish between truly disruptive and neutral variants within each.

Here we leverage saturation MPSAs as a complementary, high-resolution source of benchmarking data to evaluate eight recent and widely used splice predictors. Algorithms using deep learning to model splicing impacts using a long window of sequence context, SpliceAI and Pangolin, consistently agreed with measured splicing effects across the various performance metrics, while other tools performed well on specific exons or variant types. Even for the best performing tools, predictions were less concordant with measured effects for exonic variants versus intronic ones, indicating a key area of improvement for future algorithms.

4.3 Methods

4.3.1 Saturation mutagenesis datasets

Splice effect measurements were obtained for a total of 3,616 variants in *POU1F1* (exon 2), *RON* (exon 11), *FAS* (exon 6), *WT1* (exon 9), *BRCA1* (11 exons) from the respective studies' supplementary materials[48,55-57,59]. Variants were labeled as splice disruptive (SDV), intermediate, or neutral according to the classification made by each study; intermediate effect variants (*n*=121) were removed. Intronic variants more than 100 bp from either end of the selected exon were also discarded (*n*=129). The *RON* MPSA used a minigene spanning exons 10, 11, and 12, but as that assay did not measure skipping of exons 10 or 12, we only included variants most likely to influence exon 11 inclusion (i.e., within exon 11 and proximal halves of its flanking introns).

BRCA1 SGE measurements reflect both protein loss of function and mRNA effects, so we retained only synonymous and intronic variants to remove variants for which effects were independent of splicing, and further restricted to internal coding exons. For MPSAs in *POU1F1*, *RON*, and *WT1* that reported effects upon usage of multiple isoforms, we used for each variant the isoform score that was most different than baseline (that is, maximum absolute z-score across isoforms per variant). For consistency in direction of effect (higher measured scores denoting greater disruptiveness), *BRCA1* RNA and function scores' signs were reversed. *FAS* enrichment scores were used without modification.

4.3.2 Manual curation of clinical MLH1 variants

A literature search for variants assayed for splicing effects in the tumor suppressor gene *MLH1* yielded 77 publications (publication years 1995-2021; **Appendix A**). For inclusion, we considered only single-base substitutions, and required each variant's splicing effects be supported either by RT-PCR and sequencing from patient blood-derived RNA, or by mini-gene analysis. One exception is that essential splice site dinucleotide variants from Lynch Syndrome patients were included without molecular evidence, as loss of the native site would be considered strong evidence of pathogenicity by ACMG guidelines[195]. Any splicing outcome other than full exon inclusion was considered pathogenic[204]. Previously reported splice disruptive variants which are too common to be compatible with Lynch Syndrome prevalence (gnomAD MAF>0.5%) or were seen in a homozygous state were removed. Nine variants had conflicting reports (i.e., both pathogenic and benign) and were resolved with a majority vote among the reporting publications, with ties being considered pathogenic. The final

dataset included 296 variants (mean: 1.8 references per variant), of which 160 were splice disruptive.

4.3.3 Random background variant set

We randomly drew 500,000 SNVs from within and near protein-coding genes to serve as a background set of exonic and proximal intronic variants with the potential to effect splicing. We used MANESelect canonical gene model annotations (version 1.0)[205] restricting to protein coding transcripts with at least three coding exons. We discarded transcripts that had exons overlapping or within 100 bp of exon(s) of another transcript (on either strand), so that the variants' classification (intronic vs exonic, proximity to splice site) would not depend upon the choice of gene model; this left 79.6% of all MANESelect transcripts (n=14,618/17,631). SNVs were selected at random from internal coding exons (padded by +/- 100 bp), and then these background SNVs were scored by splice effect predictors.

4.3.4 Scoring with eight splice effect predictors

Pangolin version 1.0.2 was run with masking enabled and a distance equal to the length of the scored exon (for *MLH1* and *BRCA1*: the longest exon for each gene; for background set SNVs, 300 bp), and the reported Pangolin_max scores were used. SpliceAI version 1.3.1 was run via the python interface using a custom wrapper, with masking enabled and distance setting following the same process as for Pangolin. For the transcriptomic background set, due to the high computational time to run SpliceAI, we downloaded version 1.3 precomputed scores from Illumina BaseSpace. SQUIRLS version 1.0.0 and MMSplice version 2.2.0 were both run on the command line with

default settings to compute SQUIRLS score and delta logit PSI values, respectively. HAL was run via the web interface (http://splicing.cs.washington.edu/SE) to predict exon skipping effects. HAL requires a baseline percent spliced in (PSI) value for the wildtype sequence (a parameter which has some predictive value on its own[19,108]). For this parameter, we used the following values: 90% for *MLH1*, 90% for *POU1F1*, 50% for *FAS*, 60% for *RON*, 80% for *BRCA1*, and 60% for *WT1,* based upon WT PSI values from the single exon MSPA original publications which were based on either expert knowledge or measured from WT mini-genes and then rounded to the nearest 10%. For *BRCA1* and *MLH1*, we selected high WT PSI values since alternative splicing in MMR genes rarely creates a functional protein[204], and thus exon skipping is likely incompatible with healthy individuals in the case of dominantly inherited, highly penetrant reproductive cancers (*BRCA1*) and Lynch syndrome (*MLH1*). For exons in the random transcriptomic background, we used 50% to allow for HAL to predict a full range of exon skipping (negative values) and increased exon inclusion (positive values), and to reflect the practical challenge in determining exon specific WT PSI values genome wide. Results with HAL were generally robust to the choices of WT PSI. For SPANR, S-Cap, and ConSpliceML, we obtained precomputed scores (SPIDEX zdelta PSI scores for SPANR; sens scores for S-Cap; ConSpliceML scores for ConSpliceML) from publicly accessible databases provided by the tools' authors. For essential splice site dinucleotide scores, S-Cap provides two models (dominant and recessive), and we selected the lowest score (most severe predicted impact). We then transformed the S-Cap scores (taking y=1-x, for input scores x in [0,1]) to match the direction of effect for other tools with higher values indicating greater likelihood of splice disruption.

We selected the MANESelect transcript model for each gene tested in the benchmarking set: ENST00000350375.7 (*POU1F1;* corresponding to the predominant isoform alpha), ENST00000452863.10 (*WT1* KTS+ isoform), ENST00000296474.8 (*RON*), ENST00000231790.8 (*MLH1*), ENST00000652046.1 (*FAS*), and ENST00000357654.9 (*BRCA1*). MMSplice, SQUIRLS, SpliceAI, and Pangolin all require an accompanying annotation file, and for a fair comparison, we provided each tool an identical annotation in which only the canonical transcript within the region of interest were included. Pre-computed ConSpliceML scores were selected by matching to the genomic position and relevant gene name. SQUIRLS' annotation file was not readily customizable, so we used the default hg19 ENSEMBL annotation files that it supplies. We verified that at or near the tested exons, there were no differences between the selected gene models provided to other tools and the gene models within SQUIRLS' annotations (ENST00000350375.2 for *POU1F1* alpha, ENST00000452863.3 for *WT1* KTS+, ENST00000296474.3 for *RON*, ENST00000231790.2 for *MLH1*, ENST00000355740.2 for *FAS*, ENST00000357654.3 for *BRCA1*). Substantive results did not change for MMSplice or SQUIRLS when they were scored using the most severe predicted impact from both alternative isoforms within *POU1F1* (beta isoform: ENST00000344265.8 for MMSplice and ENST00000344265.3 for SQUIRLS) and *WT1* (KTS- isoform: ENST00000332351.9 for MMSplice and ENST00000332351.3 for SQUIRLS). For the transcriptomic background set, some variants either did not have a precomputed score for some tools, or the precomputed score record mismatched the gene name or accession; this led to a small amount of missingness for some tools expected to score every SNV (SPANR: 1.6% of background variants excluded,

SQUIRLS: 9.4%, SpliceAI: 4.1%). Pangolin and MMSplice each scored every background SDV. HAL only scores exonic variants, so all intronic variants were missing (56.5% of the background set), and S-Cap scores only some synonymous variants and variants within 50 bp of the splice sites, so had missing values for 61.0% of background variants.

4.3.5 Variant classes

To address performance within different gene regions, we categorized variants as follows: (i) essential splice site dinucleotides, (ii) intron near junction (3-10 bp from nearest exonic base), (iii) proximal intron (11-100 bp from nearest exonic base), (iv) exon near junction (<10 bp from nearest intronic base), and (v) deep exon (≥ 10 bp from nearest intronic base). For variants in multiple transcripts, the category with the most severe consequence was chosen (order: essential splice > exon near junction > intron near junction > deep exon > proximal intron). We assessed the abundance of each variant class within previously curated clinical variant sets. In the case of the SPIP training set[198], we excluded the neutral background set variants; for the S-Cap training set we combined the proportions of 5' core, 5' core extended, and 3' core variants listed from their clinically derived pathogenic set in Figure 1C[98], and for SQUIRLS we tallied variant classes across their training data without alterations[91]. When computing the proportion correct within variant regions in **Figure 4-11**, we defined regions by both splicing impacts and location. For this analysis, we included 3 bp of the exon within the donor and acceptor regions as variation in those areas is also prone to alter splicing[78]. The ESS region in *POU1F1* was defined as the cluster of beta-promoting variants, and the remainder of the beta region was so named for the observed sensitivity to variants

creating cryptic splice sites – including the alpha acceptor area which when disrupted promotes the use of a cryptic acceptor nearby[55]. Similarly for **Figure 4-19B**, 3 bp of each exon within the acceptor and both donor regions were included as part of the splice sites. For the competing donor regions in *WT1*, we also included 6 bp of each respective intron as is commonly defined as the 5' splice site area[78].

4.3.6 Nominating annotation-sensitive alternatively spliced genes

To identify genes with alternative splice forms for which choice of annotation could influence splicing effect predictions, we obtained exon-exon junction read counts from GTEx portal (version 8). We restricted to protein coding genes (n=19,817) and computed, for each of 54 tissues, the median junction read counts per million junction reads (junction CPM) across samples of that tissue for junctions that fell within coding portions of their respective genes. Junctions with a junction CPM ≥ 0.1 were considered expressed (n=16,877 genes had at least one expressed junction in at least one tissue). Next, we identified 12,124 genes where at least one splice site was alternatively used in multiple junctions meeting this expression criterion. Within each group of alternative junctions at a given splice site (e.g., two junctions corresponding to one donor paired with either of two different acceptors corresponding to skipping or inclusion of a cassette exon), we computed the fractional proportions of each junction's use and determined which alternate junctions were included in SpliceAI's default annotations. Fractional proportions were computed separately for each tissue. We deemed 'moderately used unannotated' splice sites as any group of alternative junctions with at least one unannotated expressed junction which had ≥20% fractional usage in a given tissue.

4.3.7 Statistical methods

Area under the curve metrics were calculated using sklearn in python. For **Figure 4-11** and **Figure 4-19B,** a single cutoff was selected for each tool that maximized Youden's J for identifying SDVs across the full *POU1F1* and *WT1* MPSA variant sets respectively. Tools with fewer than ten scored variants in each defined region were excluded. To compute transcriptome-adjusted sensitivity for each algorithm, we first computed the score threshold $t(x)$ at which that algorithm called a given fraction x of the transcriptomic background set as disruptive (for all values of x in [0,1]). Transcriptome-adjusted sensitivity was then the sensitivity for benchmark SDV detection at this threshold: (# benchmark SDVs with score≥t(x)) / (# benchmark SDVs). Area under the curve was then taken for transcriptome-adjusted sensitivity as a function of the background set fraction x, and was computed using the sklearn auc function.

To analyze correlation between algorithms and MPSA measurements, the absolute value of each algorithm's score was taken, to accommodate HAL, MMSplice, SPANR, and Pangolin, for which signed scores indicates exon skipping vs inclusion. *FAS* was one exception to the rule: since *FAS* enrichment scores directly measured exon skipping (negative values) and exon inclusion (positive values), for signed scoring tools (HAL, MMSplice, SPANR, and Pangolin) we compared signed *FAS* scores with tools' signed values, and for the rest (SpliceAI, SQUIRLS, ConSpliceML, S-CAP), compared absolute values of the measured scores with the tools' scores. For classification performance analyses (prAUC, transcriptome-adjusted sensitivity), absolute values of tools' predicted scores were used.

4.3.8 Data availability

Scored datasets are available on Zenodo[136].

4.4 Results

4.4.1 A validation set of variants and splice effects.

We aggregated splicing effect measurements for 2,230 variants from four massively parallel splice assay (MPSA) studies, focusing on saturation screens targeting all single nucleotide variants (SNVs) in and around selected exons[55-57,59] (**Figure 4-1A**). We also included 1,386 variants in *BRCA1* from a recent saturation genome editing (SGE) study, in which mutations were introduced to the endogenous locus by CRISPR/Cas9-mediated genome editing, with splicing outcomes similarly measured by RNA sequencing[48]. For contrast with these saturation-scale datasets, we also prepared a more conventional, gene-focused benchmarking dataset by manually curating a set of 296 variants in the tumor suppressor gene *MLH1* from clinical variant databases and literature reports. In sum, this benchmarking dataset contained 3,912 SNVs across 33 exons spanning six genes (**Figures 4-2** through **4-6**).

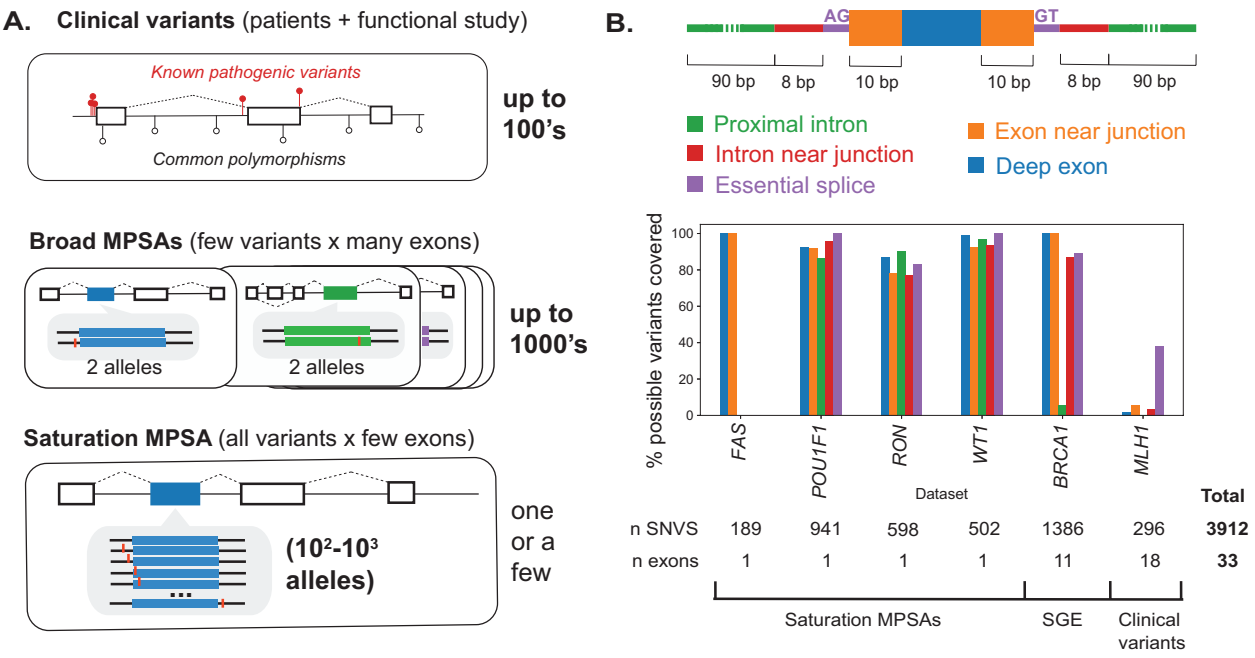


Figure 4-1: Variant sets used for splice effect predictor benchmarking.

A. Potential sources include (top panel) clinical variants including known pathogenic variants and common polymorphisms in frequently screened disease genes, (middle panel) broad massively parallel splice assays (MPSAs) targeting many different exons with one or a few variants each, and (bottom panel) saturation MPSAs in which all possible variants are created for a few target exons. **B.** Variant classes defined by exon/intron region and proximity to splice sites (upper), with the percent coverage of the possible SNVs within each variant class (shaded by variant class), for each dataset in the benchmark set (for BRCA1, missense and stop-gain variants were excluded are not counted in the denominator).

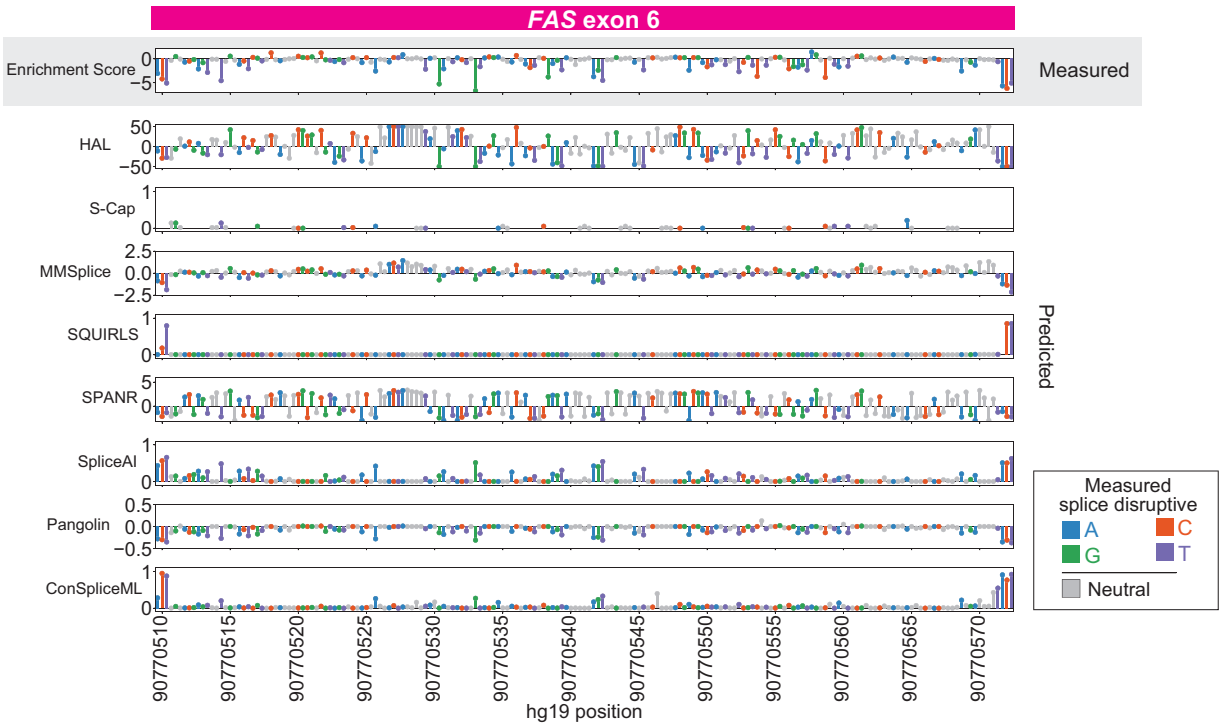


Figure 4-2: Splicing effect map and bioinformatic predictions for *FAS* exon 6.

MPSA measured enrichment score of *FAS* exon 6 (gray, top panel; increased skipping – negative values, increased inclusion – positive values), along with bioinformatic predictions (subsequent panels), with splice effects/predictions plotted by variant position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base).

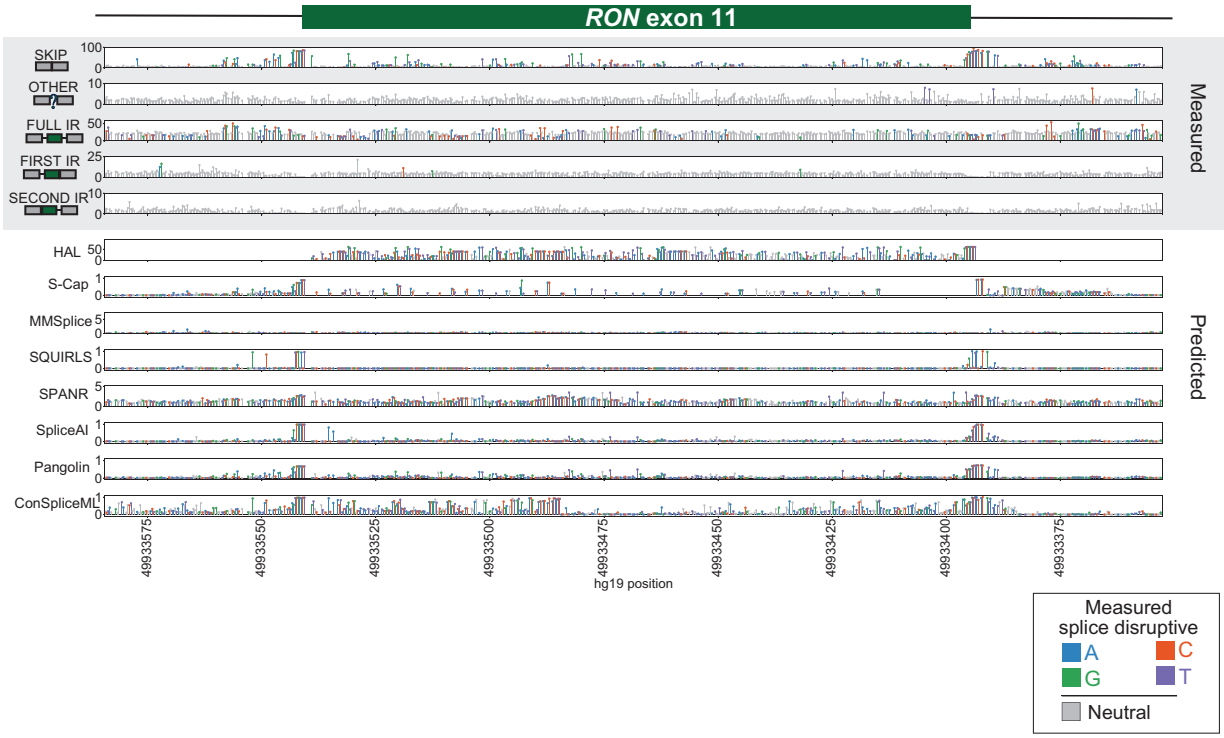


Figure 4-3: Splicing effect map and bioinformatic predictions for *RON* exon 11.

MPSA measured percent usage for different splicing outcomes at *RON* exon 11 (gray, top panel): skipping, other isoforms, full intron retention ("FULL IR"), first intron retention ("FIRST IR"), and second intron retention ("SECOND IR"), along with bioinformatic predictions (subsequent panels), with splice effects/predictions plotted by variant position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base).

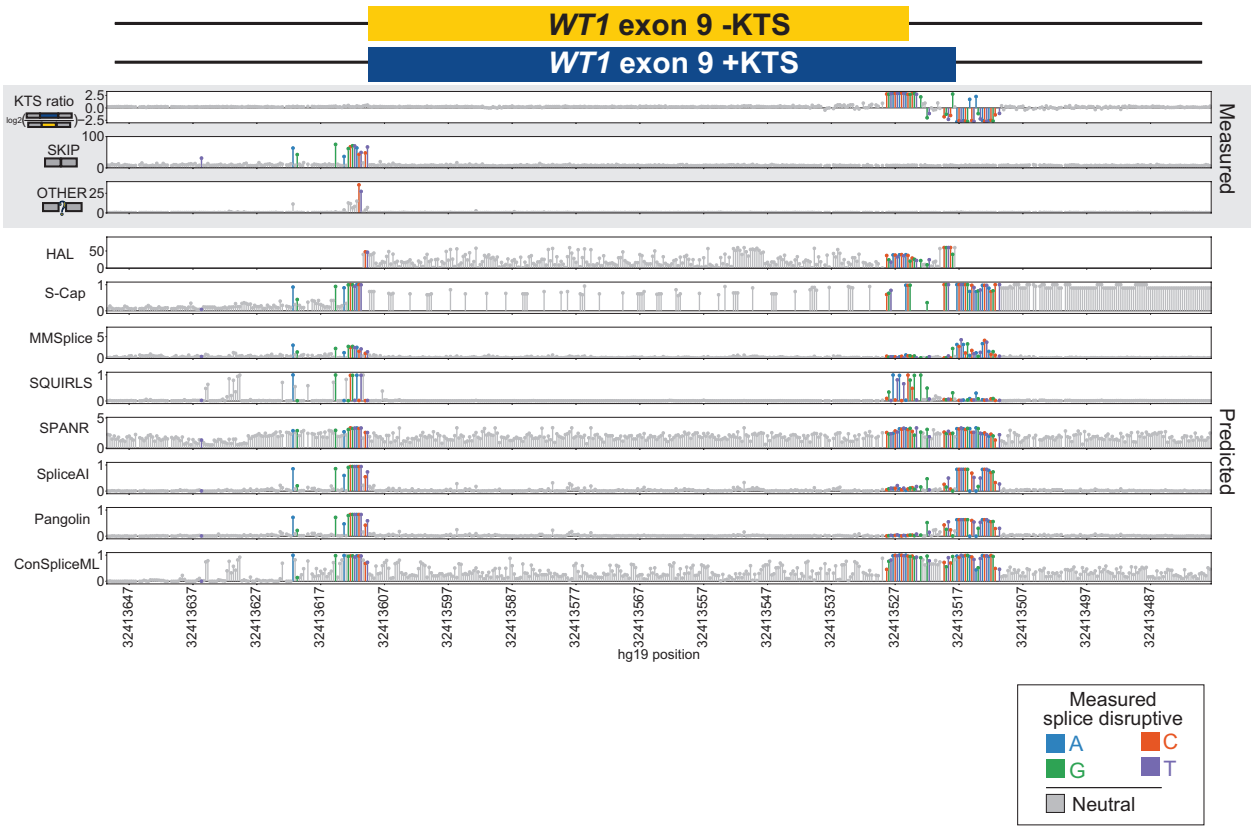


Figure 4-4: Splicing effect map and bioinformatic predictions for *WT1* exon 9.

MPSA measurements for different splicing outcomes at *WT1* exon 9 (gray, top panel): log2ratio(%KTS+/%KTS-), percent exon skipping, and percent other isoforms, along with bioinformatic predictions (subsequent panels), with splice effects/predictions plotted by variant position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base).

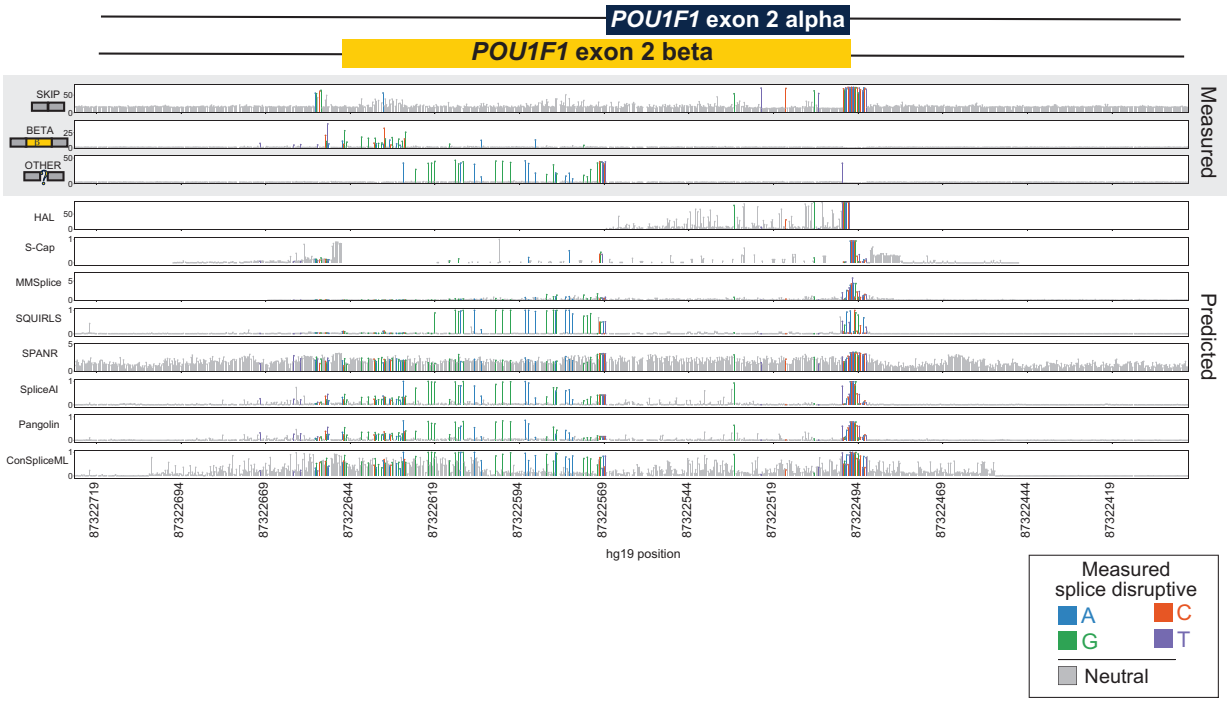


Figure 4-5: Splicing effect map and bioinformatic predictions for *POU1F1* exon 2.

MPSA measured percent usage for different splicing outcomes at *POU1F1* exon 2 (gray, top panel): exon skipping, exon 2 beta, and other isoforms, along with bioinformatic predictions (subsequent panels), with splice effects/predictions plotted by variant position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base).

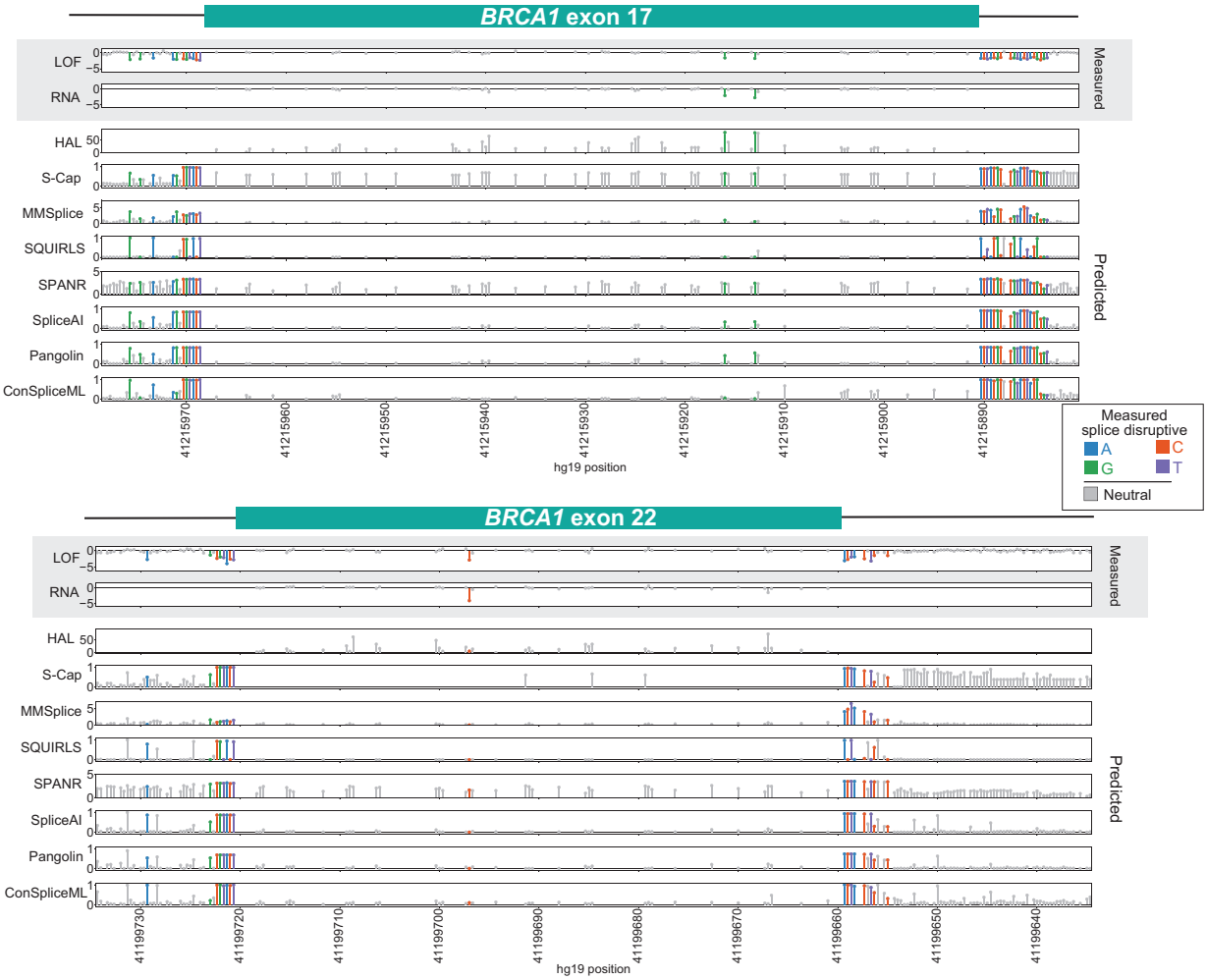


Figure 4-6: Splicing effect map and bioinformatic predictions for select *BRCA1* exons.

SGE measurements for *BRCA1* exons 17 and 22 (gray, top panel): log2-ratio function score and log2ratio RNA score, along with bioinformatic predictions (subsequent panels), with splice effects/predictions plotted by variant position. Each lollipop denotes one variant, shaded by effect in SGE (gray: neutral, colors: SDVs, shaded by mutant base).

As expected, MPSAs measured most of the possible single-nucleotide variants at each target (93.3% of SNVs), with relatively uniform coverage by exon/intron region (**Figure 4-1B**). From the *BRCA1* SGE study, we retained only intronic or synonymous variants, because missense variants' effects could be mediated via protein alteration, splicing impacts, or both. Targeted exons varied in their robustness to splicing disruption from *POU1F1* exon 2 (10.2% SDV) to *RON* exon 11 (68.4% SDV; **Figure 4-**

7) reflecting both intrinsic differences between exons as well as different procedures for calling SDVs between MPSA studies. In contrast to the high coverage of the mutational space from MPSA and SGE datasets, reported clinical variants only sparsely covered the mutational space (1.6% of the possible SNVs in *MLH1* exons +/- 100 bp), and were heavily biased towards variants near splice sites (59.5% of reported variants were within +/-10 bp of a splice site; **Figure 4-8**). Larger clinical variant sets used to train classifiers showed a similar skew: 94.6% of the SQUIRLS training variants[91] and 88.9% of the pathogenic S-Cap training set[98] were within 10 bp of a splice site. Thus, MPSAs offer high coverage without the variant class biases present among clinical variant sets.

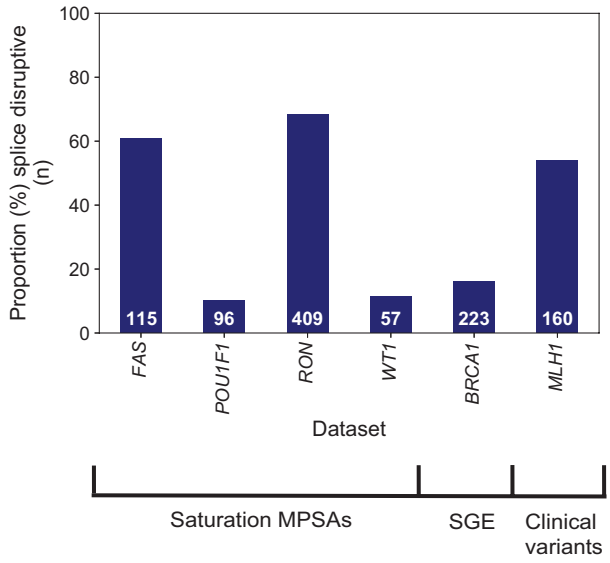


Figure 4-7: Proportion of splice disruptive variants (SDVs) within benchmarked datasets.

Bar plot showing the proportion of SDVs out of all measured variants (y-axis) within the saturation MPSAs (*FAS*, *POU1F1*, *RON*, *WT1*), SGE (*BRCA1*), and clinically curated variant set (*MLH1*) (x-axis). Numbers on each bar display the count of splice disruptive SNVs per dataset.

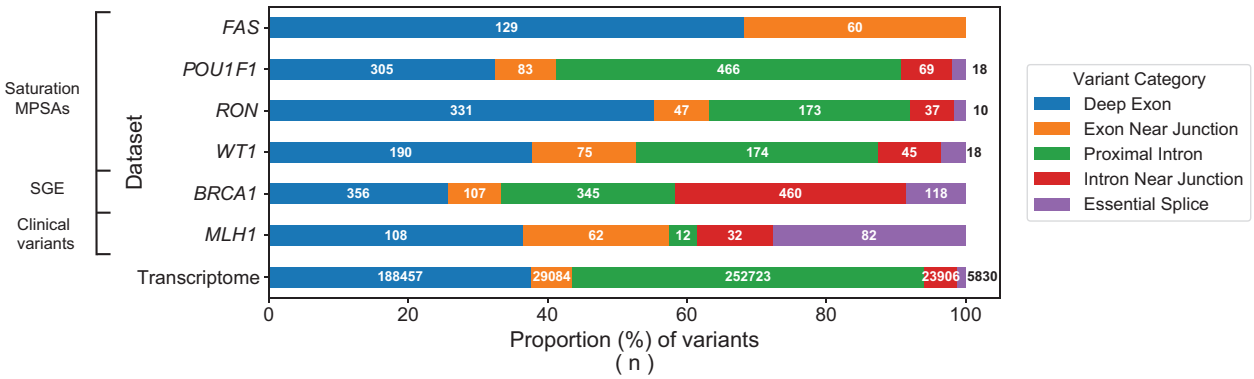


Figure 4-8: Breakdown of benchmark and background variant sets by variant class.

Proportions (x-axis) of variant category (color) deep exon (blue) in each benchmark variant dataset. Datasets are grouped by study type (MPSAs, SGE, and clinical variants), and 'transcriptome' denotes the random background set of variants. Variant categories are defined and shaded as in **Figure 4-1B**. Numbers of each bar indicate count of each type of variant per dataset.

4.4.2 Comparing bioinformatic predictions with MPSA measured effects

We selected eight recent and widely used predictors to evaluate: HAL[90], S-Cap[98], MMSplice[97], SQUIRLS[91], SPANR[99], SpliceAI[100], Pangolin[101], and ConSpliceML[105]. Most variants (93.1%) were scored by all tools except HAL and S-Cap (which focus on exonic variants only, and synonymous/proximal intronic variants, respectively). Algorithms' predictions were only modestly correlated with each other (median pairwise Pearson *r* between absolute values of tools' scores = .58, range: .08 to .97; **Figure 4-9**). One exception was Pangolin and SpliceAI which share similar model architectures and training sets and were almost perfectly correlated with each other (*r* = .97). These two were also strongly correlated with MMSplice (*r* = .81 and .80 respectively). The pattern of modest agreement across tools was not specific to exons and variants tested by MPSAs: we observed a similar degree of correlation between tools' scores across a background set of randomly sampled exonic SNVs (median pairwise *r* = .60; range: .07 to .93; **Figure 4-9** and **Methods**). Concordance between algorithms was notably lower

within exons, both in the MPSA benchmarking set variants (median pairwise *r* = .43) and random background set variants (median *r* = .39).

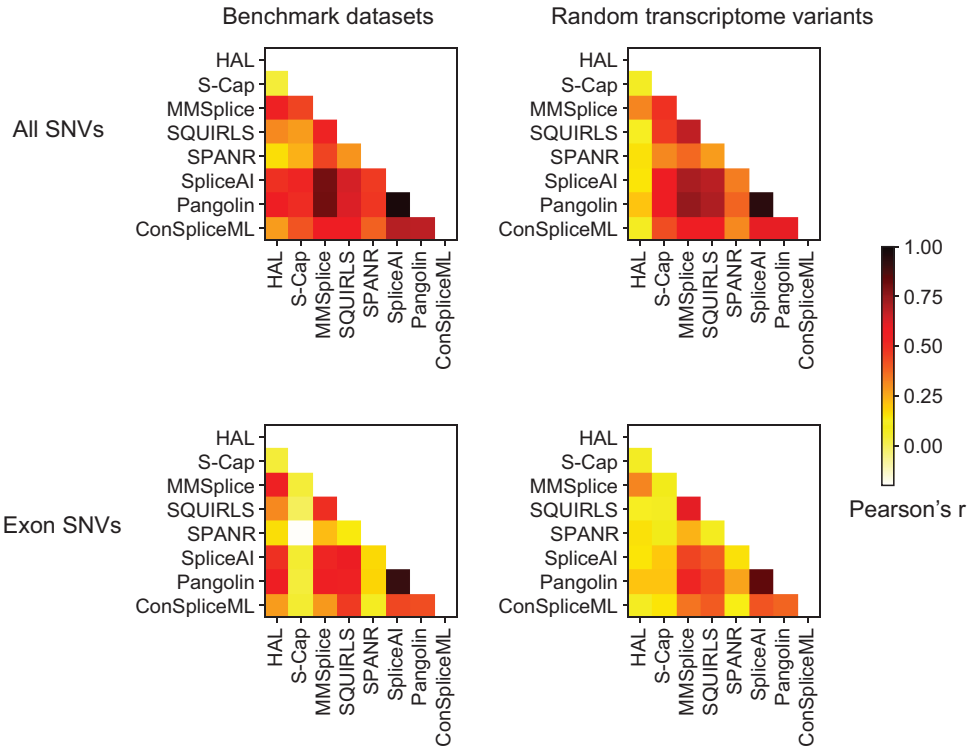


Figure 4-9: Correlation among bioinformatic algorithms.

Heatmaps of Pearson correlations between scores from eight bioinformatic algorithms across benchmark set variants (left column) and randomly selected 'background set' variants (right column). Top row shows correlations across all variants; bottom row shows correlations over only exonic variants.

Agreement between predictors' scores and experimentally measured effects also varied widely by algorithm and MPSA dataset, but were similarly modest, with a median Pearson's *r* of .52 (range of -.06 to .85; **Figure 4-10**). Even for a single algorithm and exon, substantial regional variability in concordance is evident (**Figure 4-11**). For instance, at *POU1F1* exon 2, every tool other than S-Cap recapitulated the strong constraint observed by MPSA at the donor region. By contrast, at a putative exonic splicing silencer (ESS) near the beta alternative 3' splice site, algorithmic and measured

effects were much less concordant, reflecting the difficulty of modeling variant effects outside canonical splicing motifs.

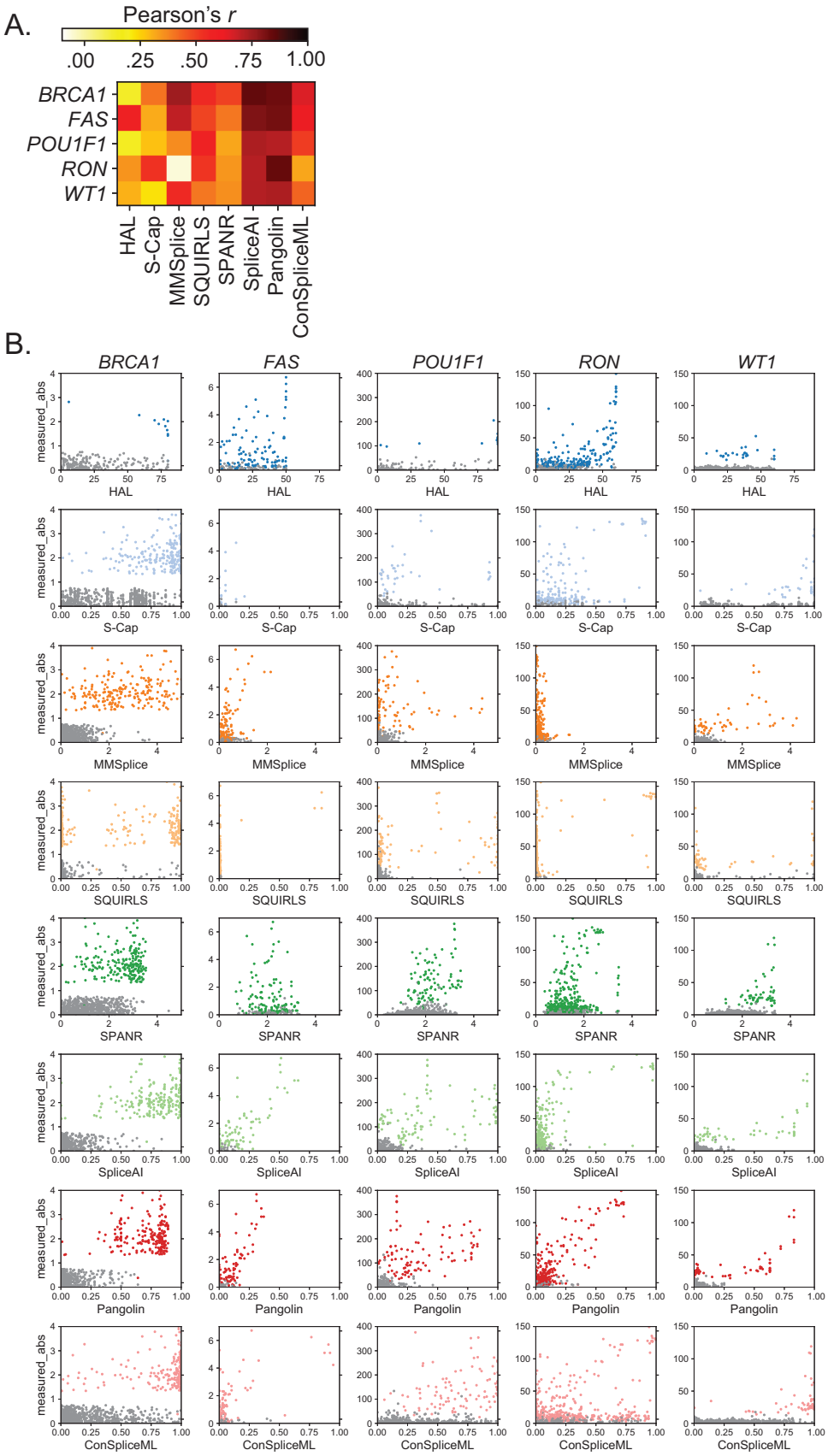

A.
Pearson's r
.00
.25
.50
.75
1.00
BRCA1
FAS
POU1F1
RON
WT1
HAL
S-Cap
MMSplice
SQUIRLS
SPANR
SpliceAI
Pangolin
ConSpliceML
B.
BRCA1
FAS
POU1F1
RON
WT1
measured_abs
HAL
S-Cap
MMSplice
SQUIRLS
SPANR
SpliceAI
Pangolin
ConSpliceML

Figure 4-10: Correlations between bioinformatic algorithms' scores and MPSA measurements.

A. Heatmap showing the Pearson's correlations between bioinformatic algorithms (x-axis) and MPSA-measured effects. *MLH1* SNVs are omitted as they were curated across many different studies and do not have measurements beyond classification as deleterious/neutral. **B.** Scatterplots of absolute measured splicing effects (y-axis) within each benchmarked MPSA dataset (columns) and computational predictions (x-axis) for each bioinformatic algorithm (rows). Shaded points were measured as splice disruptive and measured splice neutral variants are gray. As in A, since *MLH1* SNVs only have binary outcomes, they are omitted.

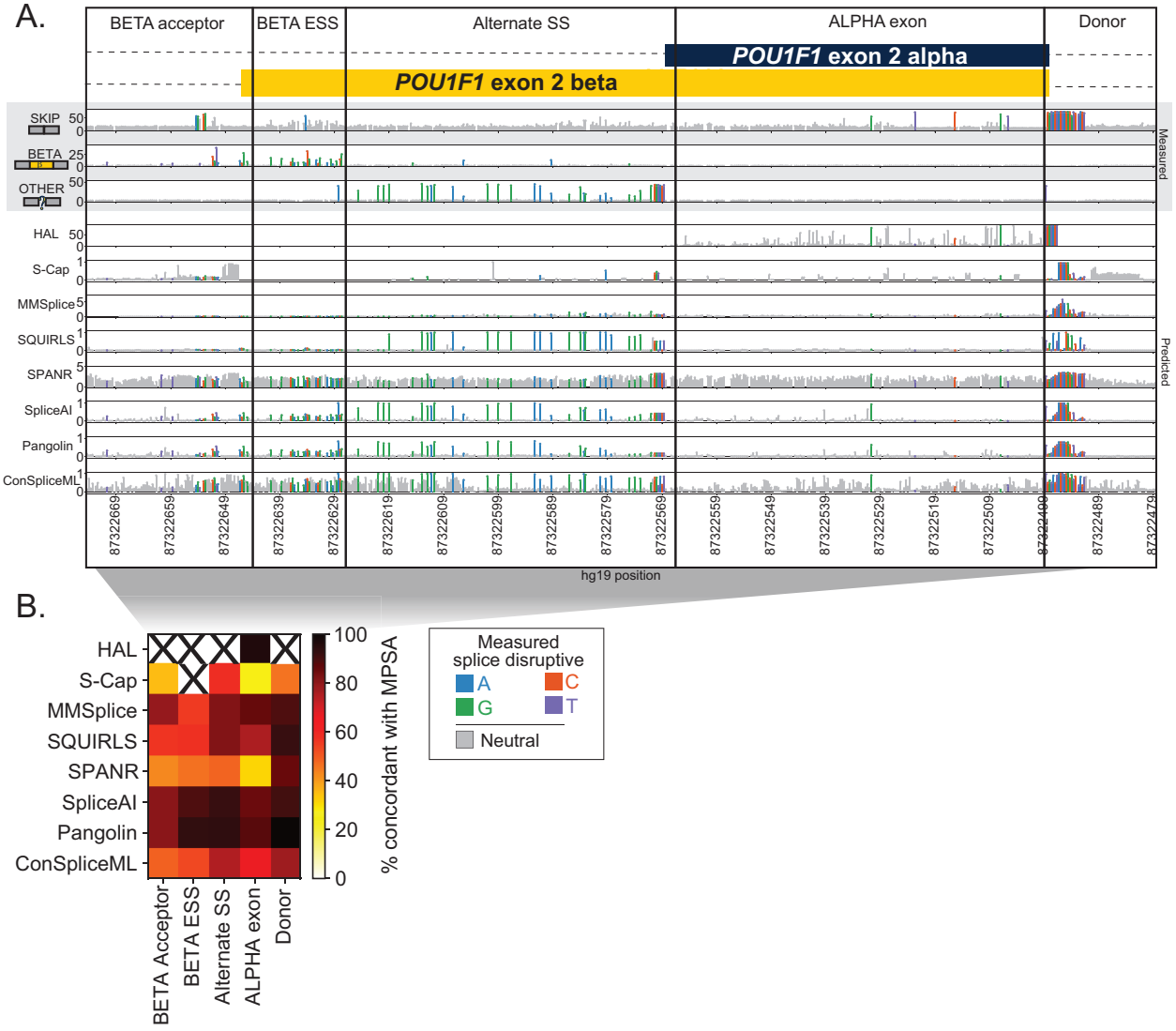


Figure 4-11: Agreement between predictors and experiments varies by gene region.

A. Splicing effect tracks at *POU1F1* exon 2 (alternate isoforms beta and alpha). Upper panel tracks (gray background) show MPSA-measured percentages of exon skipping,

beta exon inclusion, and other (non-alpha/beta/skip) isoform usage; bottom tracks show scores from bioinformatic predictors by position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base). The exon and flanking introns are split by region. Full *POU1F1* track without truncated introns shown in **Figure 4-5**. **B.** Heatmap showing of concordance between each algorithm's binary classification of variants as SDV/neutral versus those of the MPSAs, using the score threshold for each algorithm that maximized its concordance with the MPSA across *POU1F1* exon 2. Concordance is shown per algorithm (row) for each region (column). Regions with <10 scored variants are omitted (black 'X' symbols).

To systematically benchmark each predictor, we treated the splicing status from the experimental assays and curated clinical variant set as ground truth. We quantified the ability of each predictor to distinguish between the splice disruptive (*n*=1,060) and neutral (*n*=2,852) variants in the benchmark set by taking the area under the precision-recall curve (prAUC) per classifier/gene (**Figure 4-12A**). We next asked if classifiers' performance differed by variant type and location. Algorithms consistently performed better for intronic than exonic variants (median prAUC for introns: .773; for exons: .419; **Figure 4-12B**), despite a similar density of SDVs in exons and introns (28.4% and 25.9% SDV, respectively). This difference persisted even when removing canonical splice dinucleotides (**Figure 4-13**). More finely subdividing the benchmark variant set by regions (defined as in **Figure 4-1B)** demonstrated that performance suffers further from splice sites, where the overall load of SDVs is lower (**Figure 4-14**). To summarize overall performance, we counted the number of instances in which each predictor either had the highest prAUC or was within the 95% confidence interval of the winning tool's prAUC (**Figure 4-12C)**. Every tool scored well for at least one dataset or variant class, but Pangolin and SpliceAI had the best performance most frequently (7 and 3 datasets/variant classes, respectively).

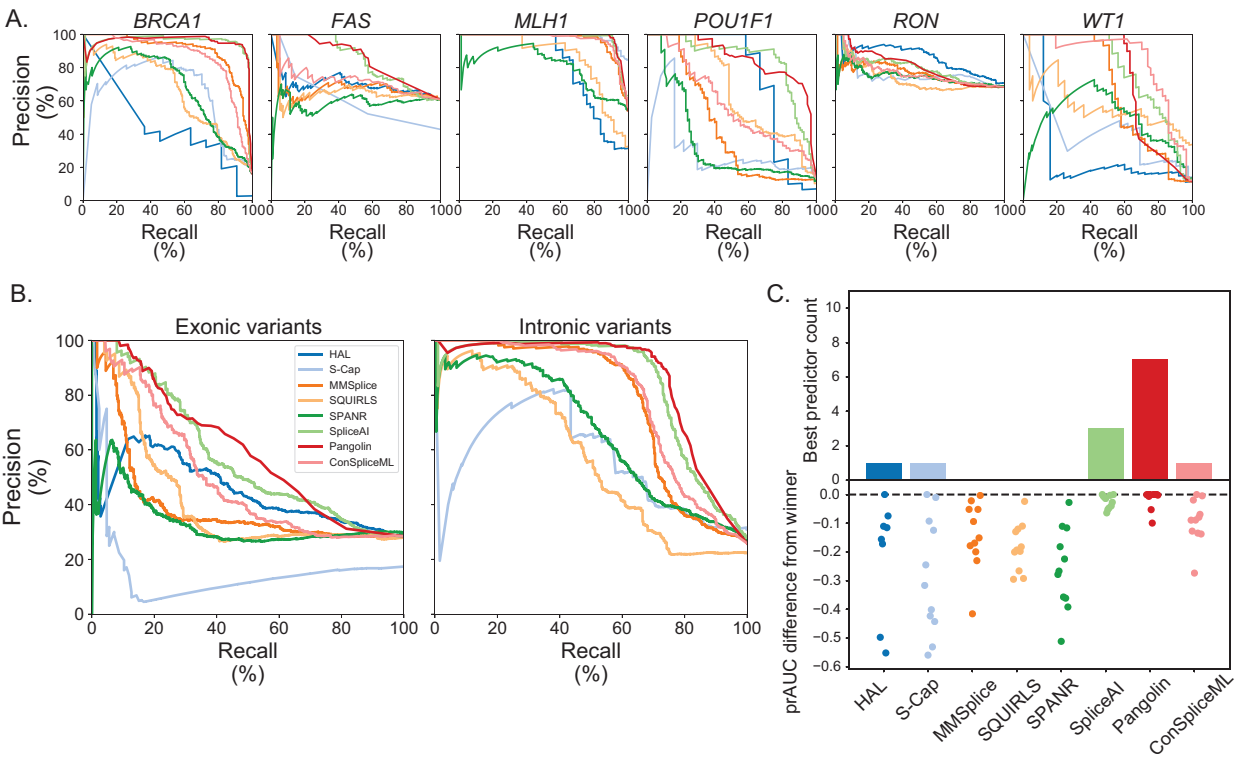


Figure 4-12: Splice effect predictors' classification performance on benchmark variants.

A. Precision-recall curves showing algorithms' performance distinguishing SDVs and splicing-neutral variants in each dataset. **B.** Precision-recall curves displaying the precision (y-axis) and recall (x-axis) of bioinformatic algorithms (colored as in A) to predict splice disruptive variants within exonic (left) and intronic (right) SNVs. **C.** Top panel: tally, for each algorithm, of the number of datasets and variant classes (defined as in **Figure 4-1B**), for which that algorithm had the highest prAUC or was within the 95% confidence interval of the winning tool. Bottom panel: signed difference between the winning tool's prAUC and a given tool's prAUC, each dot corresponds to a single dataset or variant class.

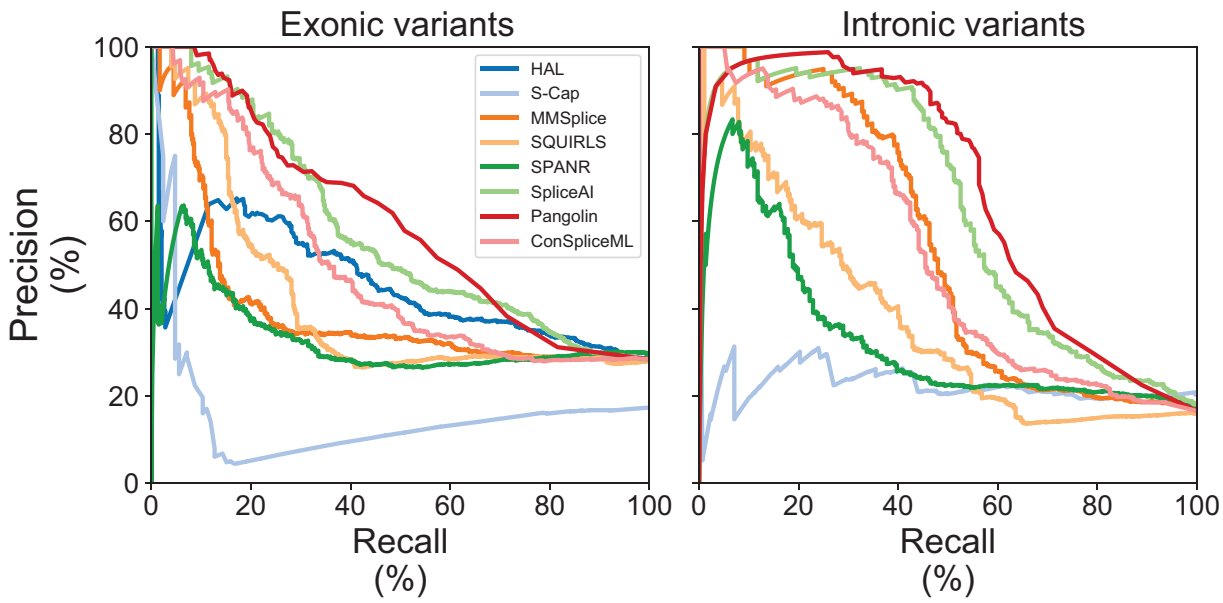


Figure 4-13: Classification performance without essential splice site variants.

Precision-recall curves showing algorithms' performance at distinguishing SDVs from splicing-neutral variants in each dataset, for exonic variants (identical to **Figure 4-12C**) and intronic variants, after removing variants at essential splice sites.

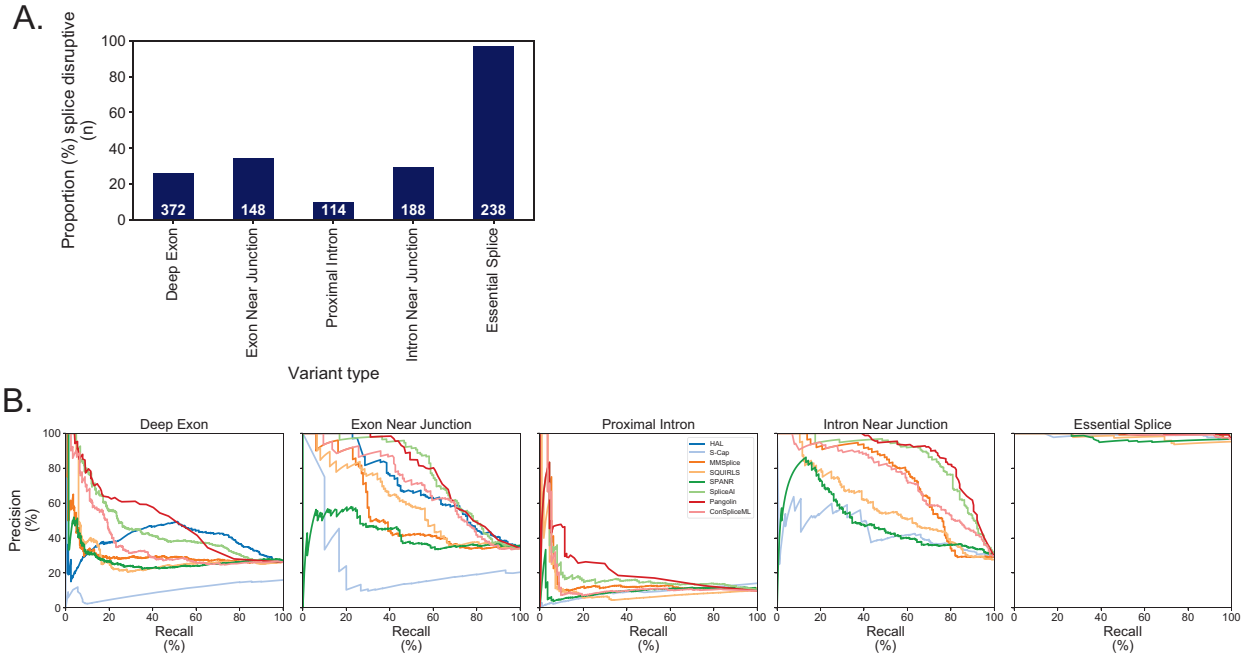


Figure 4-14: Classification performance by variant category.

A. Proportion of variants which are splice disruptive by variant category. Counts of splice disruptive variants in each category are inset. Variant categories are as defined in **Figure 4-1B**. **B**. Precision-recall curves showing algorithms' performance at

distinguishing SDVs from splicing-neutral variants in each dataset, separated by variant category.

4.4.3 Benchmarking in the context of genome-wide prediction

In practice, a splicing effect predictor must sensitively identify SDVs while maintaining a low false positive rate across the thousands of variants identified in an individual genome. We therefore evaluated each tool's sensitivity for SDVs within our benchmark set as a function of its genome-wide SDV call rate. We used a background set of 500,000 simulated SNVs drawn at random from internal protein-coding exons (+/- 100 bp) **(Figure 4-15)**. We scored these background SNVs with each tool and computed the fraction of the background set called as SDV as a function of the tool-specific score threshold. Although the true splice-disruptive fraction of these background variants is unknown, we normalized algorithms to each other by taking, for each algorithm, the score threshold at which it called an equal fraction (e.g., 10%) of the genomic background set as SDV. We then computed the sensitivity across the benchmark-set SDVs using this score threshold and termed this the 'transcriptome-normalized sensitivity'. Taking SpliceAI as an example, at a threshold of SpliceAI≥0.06, 10% of the background set is called as SDV. Applying the same threshold (SpliceAI≥0.06) to *BRCA1* SGE variants in the benchmark set, SpliceAI reaches 98.2% sensitivity and 80.7% specificity (**Figure 4-16A**).

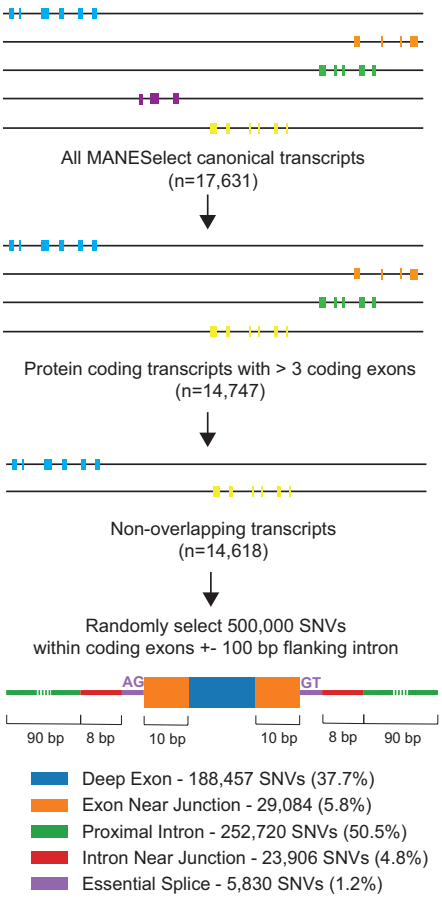


Figure 4-15: Background set of random exonic and near-exonic variants.

Schematic shows criteria used to select gene models and counts of MANESelect transcripts remaining at each step. At bottom, counts and proportions of background set variants by category.

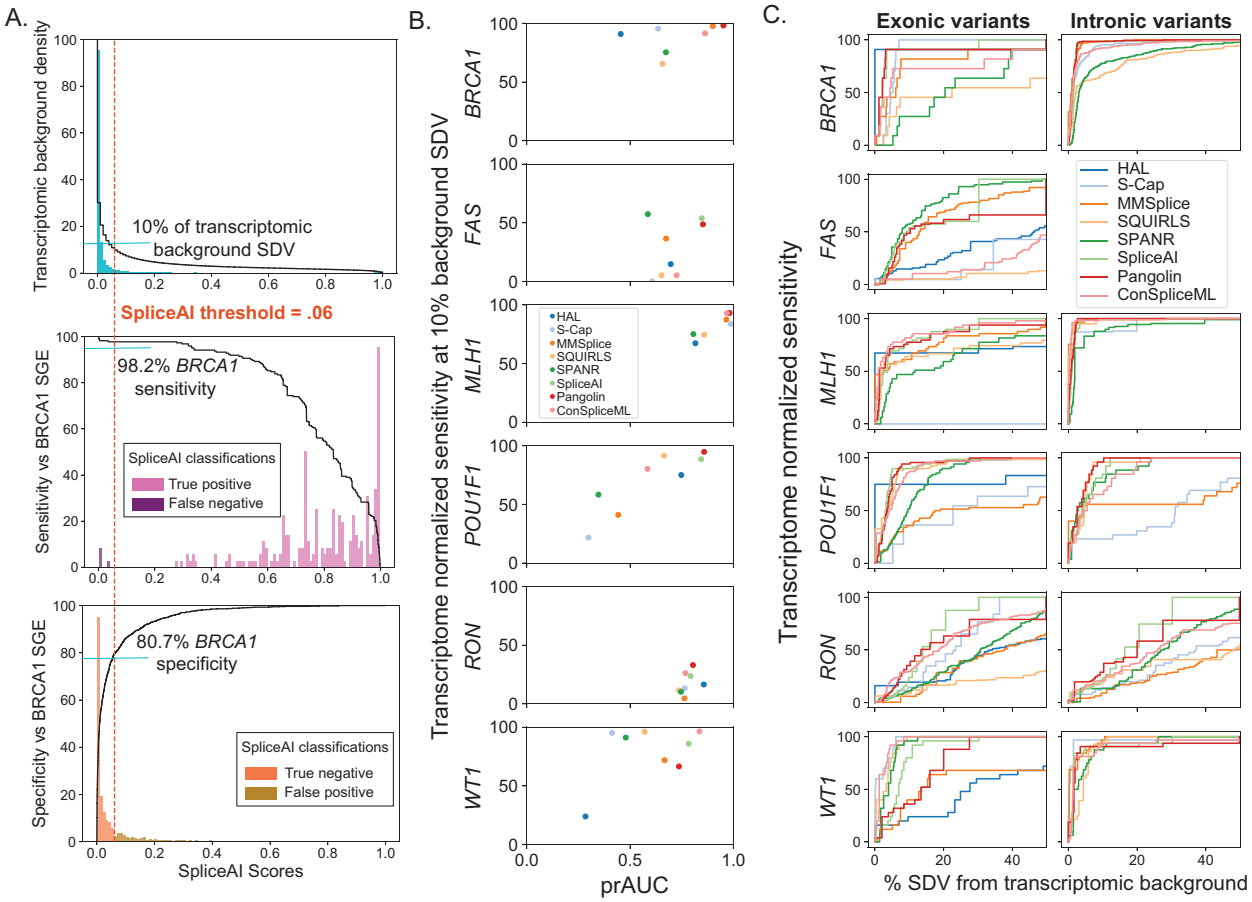


Figure 4-16: Transcriptome normalized sensitivity

A. Example shown for SpliceAI. Upper panel shows SpliceAI scores for the 500,000 background set variants (teal histogram) and the cumulative fraction (black line) of variants above a given score threshold (SpliceAI score≥0.06). Below, histograms of SpliceAI scores for *BRCA1* SGE benchmark variants, either SDVs (middle) or splicing-neutral variants (bottom) and the resulting transcriptome-normalized sensitivity and specificity at a SpliceAI cutoff of 0.06. **B.** Transcriptome-normalized sensitivity (at 10% background set SDV) versus within-benchmark variant set prAUC (**Figure 4-12A)**, by exon. **C.** Transcriptome-normalized sensitivity on benchmark variants plotted as a function of the percent of the background variant set called SDV.

We repeated this process, using for each algorithm the score threshold at which 10% of the background set was called as SDV, and applying this threshold to the benchmark set. Transcriptome-normalized sensitivity varied widely between algorithms, but SpliceAI, ConSpliceML, and Pangolin (87.3%, 85.8%, 79.9% median sensitivity respectively) emerged as consistent leaders. Mirroring the results from the precision-

recall analyses (**Figure 4-12**), median transcriptome-normalized sensitivity was lower for exonic vs intronic introns for all tools examined, by an average of 36.9%. These results were not specific to the transcriptome-wide threshold of 10%: the same three algorithms scored highly for thresholds at which 5% or 20% of the background set scored as SDV. Performance also varied by exon target (**Figure 4-16B**); for example, many of the SDVs in *FAS* exon 6 and *RON* exon 11 were not detected by any algorithm at a threshold which would classify 10% of the background set as SDV. The effects measured by MPSAs in these specific exons may be particularly subtle, creating difficult targets for prediction and suggesting that existing tools may need scoring thresholds tuned to specific exons or variant regions. Finally, we quantified the transcriptome-normalized sensitivity as a function of percent of the background set called SDV and calculated the area under the resulting curve (analogous to the prAUC statistic), showing the tradeoff between benchmark SDV recall and genome-wide SDV rate, and again highlighting consistently lower performance within exons across algorithms and datasets (**Figure 4-16C**).

4.4.4 Determining optimal score cutoffs

Integrating splice effect predictors into variant interpretation pipelines requires a pre-determined score threshold beyond which variants are deemed disruptive. We explored whether our benchmarking could inform this by identifying the score threshold that maximized the Youden's J statistic (*J*=sensitivity+specificity-1). For each algorithm, we first identified optimal score thresholds on each dataset individually to explore differences across genes and exons. For most tools we evaluated, ideal thresholds varied considerably across exons, regions, and variant classes, such that a threshold

derived from one was suboptimal for others (**Figure 4-17**). For some tools, including HAL and ConSpliceML, thresholds optimized on individual datasets spanned nearly the tools' entire range of scores, while for others such as SQUIRLS, SpliceAI, and Pangolin, the optimal thresholds were somewhat less variable. For the tools that showed the consistently highest classification performance and transcriptome-normalized sensitivity, SpliceAI, Pangolin, and ConSpliceML (**Figures 4-11 and 4-16**) - we found that the optimal thresholds were usually lower (72.2% of the time) than the threshold recommended by the tools' authors, largely consistent with conclusions of other previous benchmarking efforts[103,104,107,199]. Optimal thresholds also differed by variant class, suggesting that tuning cutoffs by variants' annotated effects, like those implemented by S-Cap, may offer some improvement for classification accuracy on variants genome-wide.

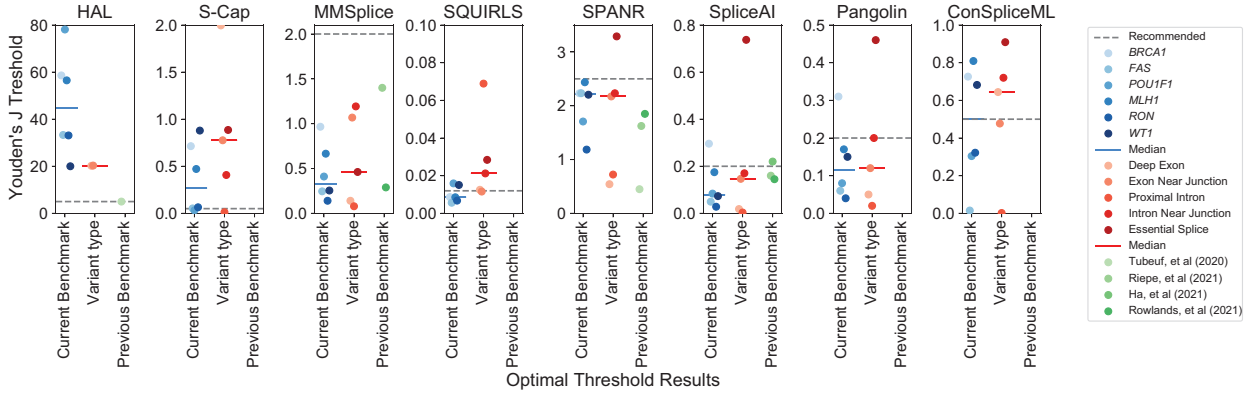


Figure 4-17: Optimal thresholds to classify splice disruptive variants.

Optimal score thresholds (y-axis) for each algorithm, across each benchmark variant datasets (blue points), by variant type (red points), and compared to previous reports (green points). Dashed black line shows tools' recommended thresholds. Solid lines indicate medians.

4.4.5 Variant effects at alternative splice sites

Alternative splicing can present challenges for prediction of nearby variants' impacts. Many splicing effect predictors require gene model annotation as an input, including four of the eight tested here (MMSplice, SQUIRLS, SpliceAI, and Pangolin). Effect predictions for individual variants may be influenced by the inclusion or exclusion of alternative isoforms in these annotations. For instance, SpliceAI and Pangolin by default apply a mask which suppresses scores from variants that either strengthen known splice sites or weaken unannotated splice sites, under the assumption that neither would be deleterious. Although masking may be a useful step to reduce the number of high-scoring variants genome-wide, it requires the provided annotation to be complete, and assumes there is no functional sensitivity to the relative balance among alternative splice forms.

We examined the effects of annotation choices and masking options at two alternatively spliced exons in our benchmark variant set. In the first, *POU1F1*, two functionally distinct isoforms (alpha and beta) result from a pair of competing acceptors at exon 2. Alpha encodes a robust transactivator and normally accounts for ≥97% of *POU1F1* expression in the human pituitary[69,109,115,117]. Beta exhibits dominant negative activity, and SDVs that increase its expression cause combined pituitary hormone deficiency[55,70]. We focused on SpliceAI in which the default annotation file only provides the alpha transcript. Predictions were broadly similar after updating annotations to include only the beta isoform or both: 13.8% (*n*=130/941) and 10.5% (*n*=95/941) of the variants, respectively, changed classifications compared to SpliceAI run with default annotations (each at an SDV cutoff of SpliceAI≥.08 which was optimal across the dataset; **Figure 4-18**) Among these were several pathogenic SDVs including c.143-

5A>G[145] which is associated with combined pituitary hormone deficiency (CPHD), scored as highly disruptive by MPSA[55], and was validated *in vivo* by a mouse model[206]. With the default annotations (alpha isoform only) and when including both isoforms, SpliceAI scores c.143-3A>G as disruptive (SpliceAI=.21 and .16 respectively). However, when only the beta isoform is considered, this variant is predicted neutral (SpliceAI=0). A similar pattern emerged at a cluster of six pathogenic SDVs which disrupt a putative beta suppressing exonic splicing silencer[55]. Therefore, counterintuitively, pathogenic SDVs which act by increasing beta isoform usage go undetected when using an annotation specific to that isoform.

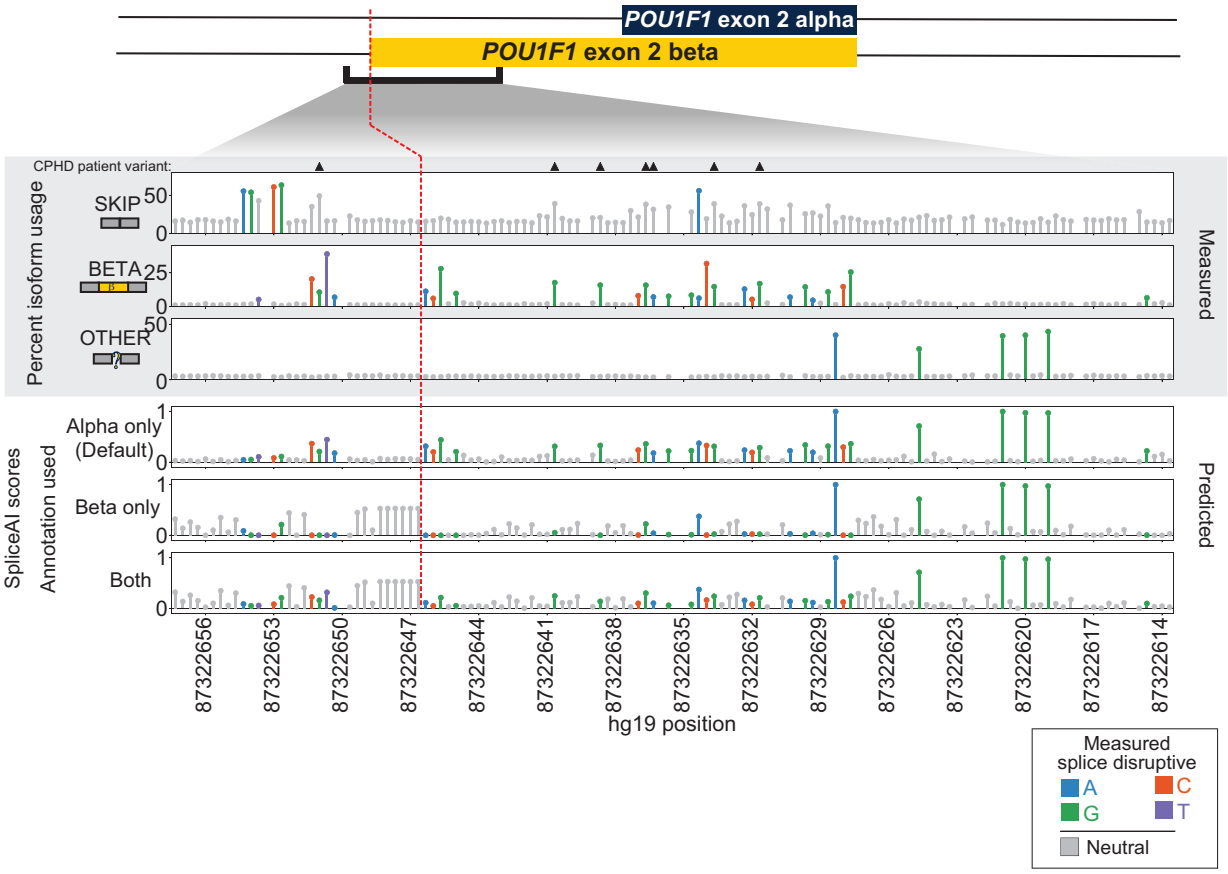


Figure 4-18: Effects of SpliceAI annotations within *POU1F1* exon 2 beta acceptor.

MPSA measured percent usage of *POU1F1* isoforms are shown in the upper tracks (gray background), with variants called SDVs shaded with color and denoted as in

Figure 4-5. SpliceAI scores are shown in the bottom three tracks, obtained using default annotation (alpha isoform only; top), beta isoform only (middle), or both isoforms (bottom). Combined pituitary hormone deficiency (CPHD) patient variants are marked with black triangles.

The choice of canonical transcript may be less clear when alternative isoforms' expression are more evenly balanced, as is the case for *WT1,* a key kidney and urogenital transcription factor gene[207] covered by our benchmarking set. Exon 9 of *WT1* has two isoforms, KTS+ and KTS-, named for the additional three amino acids included when the downstream donor is used[72,75]. In the healthy kidney, KTS+ and KTS- are expressed at a 2:1 ratio[73,76]. Decreases in this ratio cause the rare glomerulopathy Frasier's Syndrome[73,74,76], while increases are associated with differences in a sexual development[179]. We ran SpliceAI using annotations including KTS+ alone (its default), KTS- alone, and with both isoforms (**Figure 4-19A**). A cluster of variants, including one (c.1437A>G) associated with DSD[179] near the unannotated KTS- donor, appear to weaken it but are masked because that donor is absent from the default annotations. Conversely, another variant (c.1447+3G>A) also associated with DSD appears to increase the KTS+/KTS- ratio, but is also masked because it strengthens the annotated KTS+ donor (SpliceAI=0 with default annotation), and similarly scores as neutral when the annotation is updated to include both isoforms (SpliceAI=.02). The same variant scores somewhat more highly (SpliceAI=.12) when only the KTS- annotation is used, but using the KTS- annotation in turn results in failure to capture several known Frasier's Syndrome pathogenic variants near the KTS+ donor[73,76,172-175]. This case illustrates that predictors can fail even when all functionally relevant isoforms are included, because masking may suppress SDVs that are pathogenic due to strengthening an annotated splice site, resulting in imbalanced expression. This

challenge was not specific to SpliceAI; for instance, Pangolin also showed poor recovery of KTS- SDVs (only 25% correctly predicted) due to masking these losses of an unannotated donor (**Figure 4-19B**).

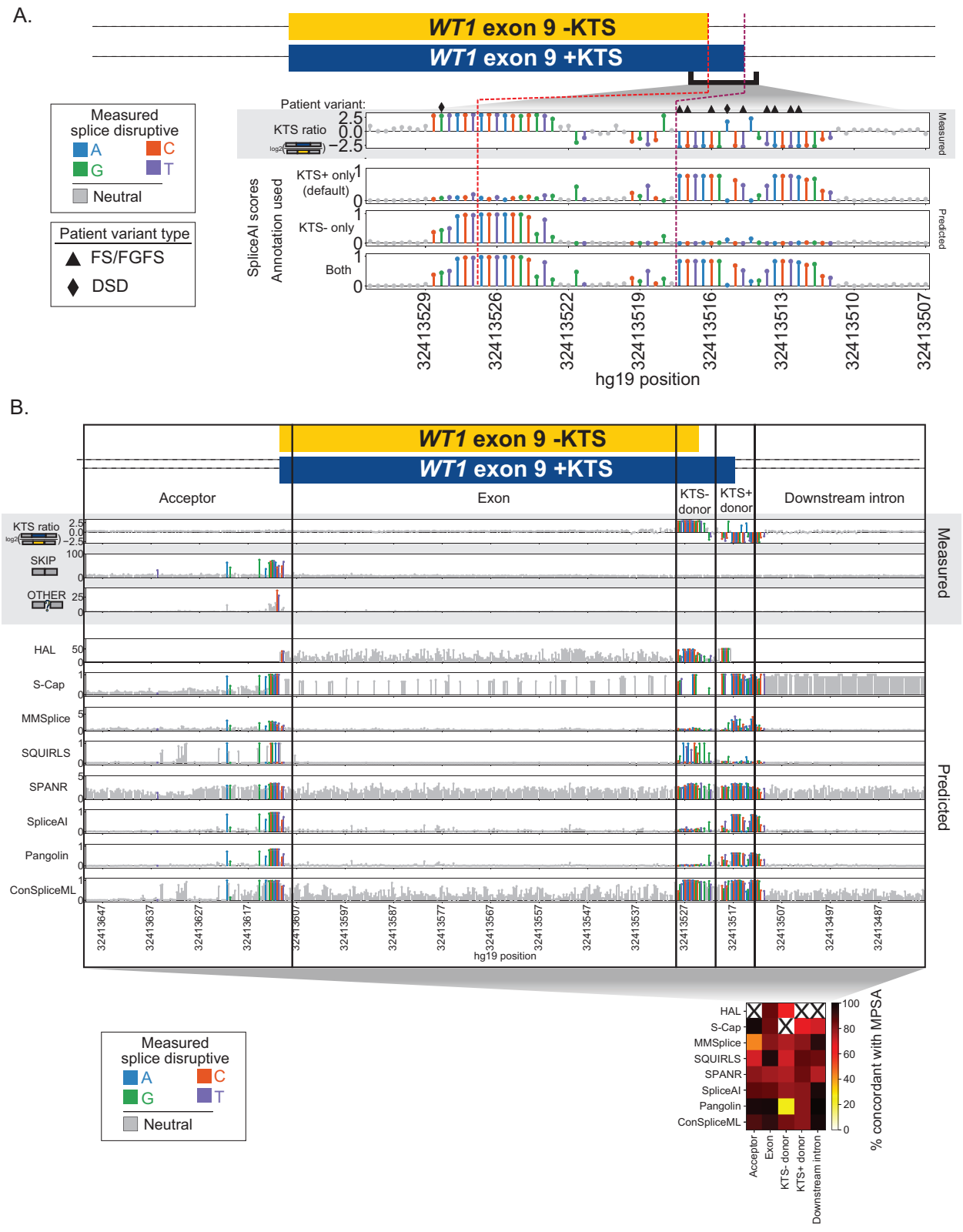


Figure 4-19: Effects of SpliceAI annotations at *WT1* exon 9 donors.

A. MPSA measured percent usage of *WT1* KTS ratio isoforms is shown in the top tracks (gray background), with variants called SDVs shaded with color and denoted as in **Figure 4-4**. SpliceAI scores are shown in the bottom three tracks, obtained using default annotation (KTS+ isoform only; top), KTS- isoform only (middle), or both isoforms (bottom). Frasier's syndrome (FS) and differences in sexual development (DSD) patient variants are marked with black triangles and diamonds respectively. Splicing effect tracks at *POU1F1* exon 2 (alternate isoforms beta and alpha). **B.** Upper panel tracks (gray background) show MPSA measurements of KTS ratio, exon skipping, and other (non-KTS+/KTS-/skip) isoform usage; bottom tracks show scores from bioinformatic predictors by position. Each lollipop denotes one variant, shaded by effect in MPSA (gray: neutral, colors: SDVs, shaded by mutant base). The exon and flanking introns are split by region. *WT1* track without regions shown in **Figure 4-4**. Heatmap showing of concordance between each algorithm's binary classification of variants as SDV/neutral versus those of the MPSAs, using the score threshold for each algorithm that maximized its concordance with the MPSA across *WT1* exon 9. Concordance is shown per algorithm (row) for each region (column). Regions with <10 scored variants are omitted (black 'X' symbols).

POU1F1 and *WT1* do not represent exceptional cases: using RNA-seq junction usage data from the GTEx Consortium[117], we estimate 18.0% of all protein coding genes (*n*=3,571/19,817 genes) have at least one alternate splice site that is expressed and at least with modestly used (≥20% PSI) in at least one tissue, yet absent from SpliceAI default annotations (**Figure 4-20**). One of these is *FGFR2*, a tyrosine kinase gene with key roles in craniofacial development[208-210]. Mutually exclusive inclusion of its exons IIIb and IIIc results in two isoforms (FGFR2b and FGFR2c) with different ligand specificites[208,209,211], and disruption of exon IIIc splicing causes Crouzon, Apert, and Pfeiffer Syndromes, which share overlapping features including craniosynostosis (premature cranial suture fusion)[212-215]. Pathogenic variants cluster near exon IIIc splice sites and at a synonymous site that activates cryptic donor usage within the exon[212,214-234] (**Figures 4-21** and **4-22**). The default annotation excludes exon IIIc, causing all four pathogenic variants at its acceptor to be scored splice neutral, but when IIIc is included in the annotation, all four are predicted with high confidence (all ≥.99). Disabling masking in order to capture cases such as these is not a viable option, as it greatly

reduces overall performance, and drastically increases the number of high-scoring variants which must be reviewed[108].

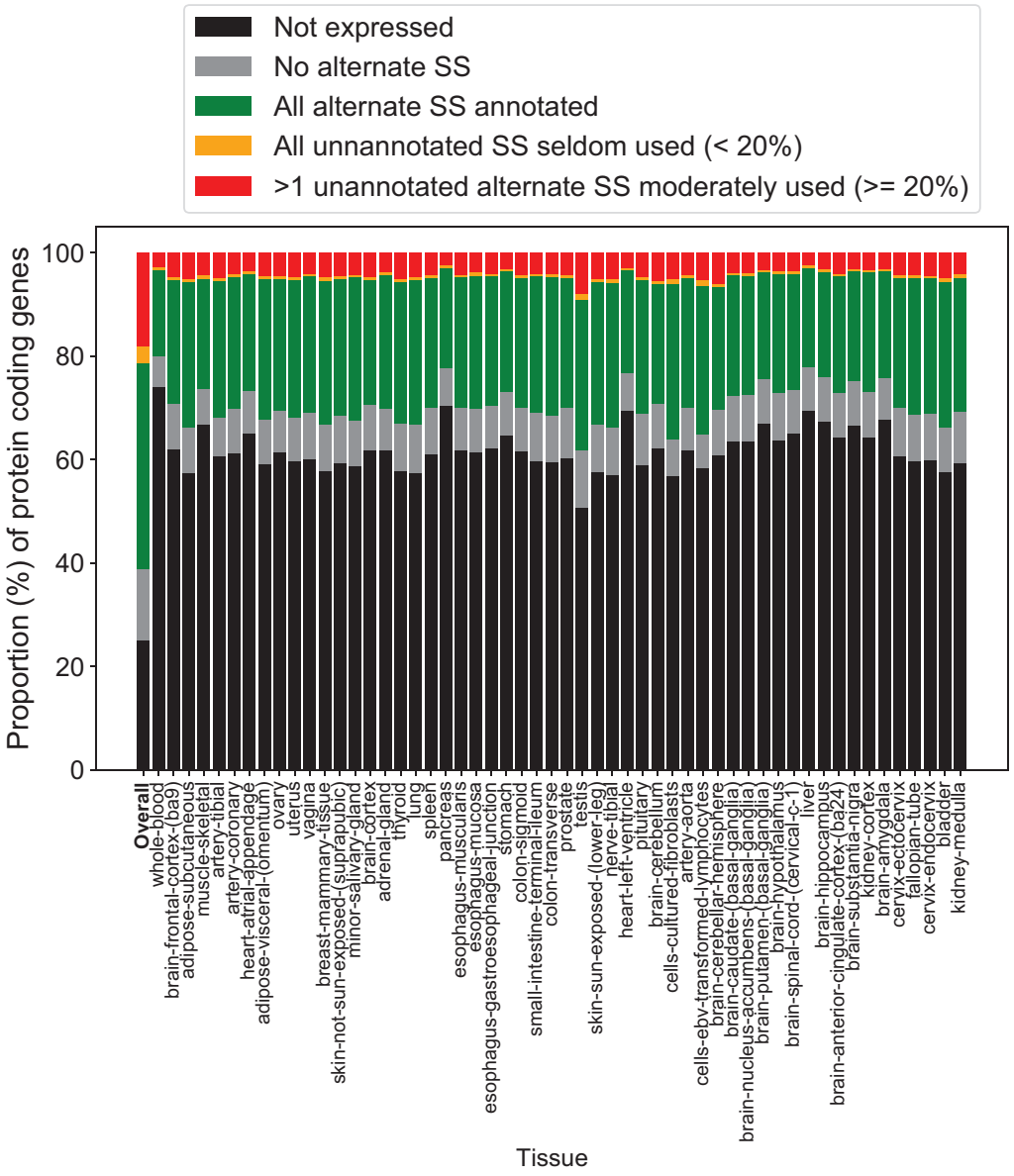


Figure 4-20: Annotation sensitive alternatively spliced genes in GTEx

Barplot showing the proportion of protein coding genes within GTEx (y-axis) that are not expressed (CPM < .1) (black), have no expressed alternate splice sites (SS) (gray), have all alternative sites annotated by SpliceAI (green, have only seldom used unannotated alternate splice sites (orange), and have at least one unannotated alternate splice site with modest use (≥20% PSI; red) both across all tissues (first bar – Overall) and within tissues (x-axis).

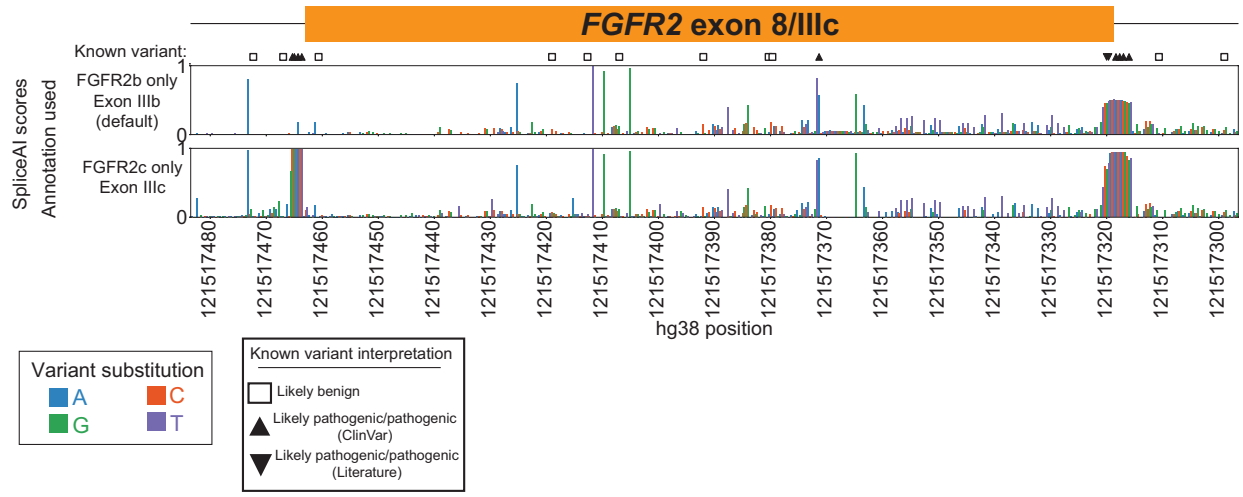


Figure 4-21: Effects of annotation choices in *FGFR2* exon IIIc.

SpliceAI scored against the FGFR2b isoform without exon IIIc (default) and the FGFR2c isoform with exon IIIc. Bars are individual variants shaded by nucleotide substitution. Synonymous and intronic variants classified in ClinVar as likely benign (open square) or likely pathogenic/pathogenic (shaded triangles) are indicated.

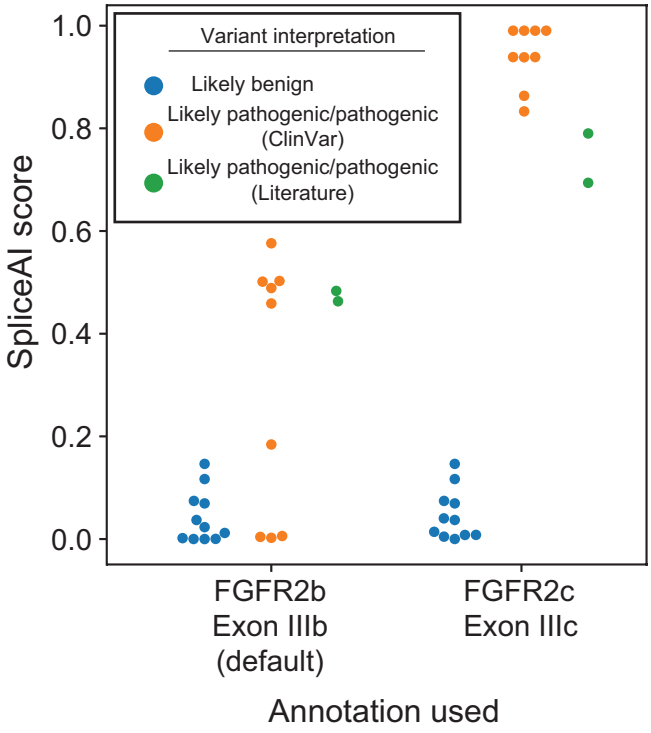


Figure 4-22: Effects of annotation choices on known variants in *FGFR2* exon IIIc.

SpliceAI scores (y-axis) using the FGFR2b isoform without exon IIIc (default; left) and the FGFR2c isoform with exon IIIc (right) annotations (x-axis). Dots shaded by variant classification.

4.5 Discussion

We evaluated the performance of eight splice effect predictors using a benchmark set of variants from saturation-level massively parallel splicing assays (MPSAs) across fifteen exons. By holding the sequence context constant for hundreds of variants per exon, these MPSAs afforded an opportunity to systematically evaluate how well each tool could distinguish individual variants' effects without confounding by differences in exons' overall characteristics. Compared to traditional validation sources such as clinical variant databases, which are enriched for essential splice site mutations, these MPSA datasets had more uniform representation of variant types including those for which classification is currently challenging.

Across most exons tested, the deep learning-based tools Pangolin and SpliceAI had the best overall performance. These two were not uniformly superior, however, and other tools excelled on certain datasets. ConSpliceML was comparably sensitive at identifying SDVs within the benchmarking set, while normalizing for genome-wide call rate, and MMSplice performed well for intronic SDVs. Even for the best performing tools, SDVs were more difficult to identify within exons compared to introns, highlighting an area of focus for future splice prediction algorithms. These results are consistent with other recent splice predictor benchmarks using broad MPSAs and clinical variants, which also noted low concordance among tools[104,108], particularly for exonic variants[107], and poorer classification performance in exons and with greater distance from splice sites[61,91,100,198,199]. As in our study, in these past comparisons, SpliceAI was often but not always the top performer[102-104,106-108,198]. Together, these our results suggest

opportunities for metaclassifiers to better calibrate existing predictors and to leverage each within its strongest domain[105,108].

A key issue this benchmarking study highlights is the challenge of selecting a scoring threshold for splicing predictors. This may reflect differences in exons' and genes' intrinsic vulnerability to SDVs, as a function of factors such as splice site strength[235] and WT exon inclusion rates[19]. For instance, most predictors fared poorly on *FAS* exon 6 and *RON* exon 11, both of which are intermediately included at baseline, and so may be more sensitive to splice disruption[19]. For moderately included exons such as these, more lenient thresholds may be required.

Another consideration is that these predictors do not directly model differences in dosage sensitivity between genes. Recessive disease genes may tolerate SDVs that reduce the abundance of properly spliced mRNA by ~50% whereas in more highly dosage sensitive genes an equally disruptive SDV would be highly deleterious. At the extreme, SDVs that lead to expression of protein isoforms with dominant negative effects may be deleterious even at a low level of expression as in the case of *POU1F1* exon 2 beta-promoting SDVs. The nature of the aberrant splice form is also important to consider – for instance, while *DNM1* loss of function can result in developmental and epileptic encephalopathies, specific SDVs yield in-frame insertions which act in a dominant negative fashion and cause particularly severe presentation[236]. Interpreting the results of bioinformatic splice effect predictions may therefore depend upon knowledge of the individual genes' dosage sensitivity, which potentially limits the utility of readily computed genome-wide scores. Methods such as ConSpliceML offer a means

of inferring such thresholds by modeling on a per-gene or per-exon basis the constraint against SDVs among healthy individuals[105].

Our results also highlight the major influence of gene model annotation, a required input for many splice effect predictors. For two of the MPSA-tested exons in our benchmarking set (*POU1F1* and *WT1*), inclusion of alternate splice forms changed SpliceAI predictions across >10% of variants. Using RNA-seq data from GTEx, we conservatively project that this challenge may impact nearly one in every five genes in the human genome. Such annotation changes are inconvenient for end users and are not readily accommodated by some tools. Moreover, they may not be possible when the functionally relevant isoforms are not known in advance. Using the most comprehensive annotation set is not a universal fix, as illustrated by *POU1F1*, where it resulted in poorer concordance with MPSA measurements, and lower specificity in recovering pathogenic variants. Some tools, including MMSplice and SQUIRLS, provide splicing effect predictions specific to all overlapping transcripts, and could permit investigation of isoform specific effects at the cost of reviewing many additional variant scores.

One limitation of our study is that the splicing assays we drew from made certain tradeoffs in exchange for scale. Minigene-based MPSAs necessarily include only minimal sequence context, and cannot capture effects from transcription elongation rate or nucleosome positioning each of which can influence splicing[237]. MPSA and SGE experiments typically use immortalized cancer cells, in which the splicing factor milieu may differ from that of the relevant tissues *in vivo*. Nevertheless, minigene assays are often well correlated across cell lines[36,55,56,61,184] and have a sufficient track record of concordance with blood RNA analysis that they are often deemed acceptable as

functional evidence during clinical variant interpretation[199,238,239]. Moreover, even when minigene assays misidentify a variant's aberrant splicing outcome(s), they may still correctly flag the variant itself as splice disruptive[202,240]. In the future, improved splice effect benchmarking data could result from MPSAs with longer sequence contexts[165] or delivered to more relevant tissues[241], and from emerging approaches for in situ genome engineering[48,242]. Additionally, the growing usage of RNA-seq in genetic testing[163,193,194] provides an opportunity to contribute both SDVs and neutral variants to the training and validation of future splice predictors.

4.6 Conclusion

Here we have shown that saturation MPSAs provide an opportunity to critically evaluate the performance of computational splice effect predictors. Our results complement past benchmarking efforts using clinical variants and more broadly targeted MPSAs, by testing algorithms' ability to distinguish individual variants' effects within the context of a single exon. This classification task resembles that faced by clinicians during variant interpretation, as even in disease gene exons which are vulnerable to splice disruption, there are many rare variants which do not impact splicing. We nominated SpliceAI and Pangolin as the top-performing tools, noting shortcomings including in exonic variant performance, and identifying practical challenges that end-users may encounter including selection of thresholds and the need for careful attention to gene model annotations. The continued growth of MPSA screens will present an opportunity to further improve splice effect predictors to assist in the interpretation of variants' splicing impacts.

4.7 Acknowledgments

The results presented in this chapter is under review at *Genome Biology* and is currently available on bioRxiv[243]. I'd like to thank my mentor Jacob Kitzman for helping to steer and guide what started as a pandemic project while wet lab work was paused into a fully realized publication. This work was supported by the National Institutes of Health (R01GM129123 to JOK).

Chapter 5 Conclusions and Future Directions

The results presented in this dissertation emphasize the importance of considering the functional effects at the level of RNA splicing when conducting whole genome sequencing. My studies highlight a continued need for improved *in silico* splice effect prediction algorithms. Chapters 2 and 3 demonstrate the use of high throughput, experimental screens of splicing effect, which systematically nominated splice altering variants across whole exons within one experiment. Chapter 2 identified 96 splice altering SNVs in and around exon 2 of the pituitary specific transcription factor *POU1F1* (*n*=96/1,070; 9% of all measured SNVs), and importantly 14 of those splice disruptive variants were synonymous substitutions (*n*=14/108 synonymous substitutions; 13%) which would be considered low priority during clinical sequencing[55]. Although most variants were not splice disruptive, the massively parallel splicing screen detected a putative exonic splice silencer (ESS) region which represses the use of the nearby beta acceptor. This acceptor activates the expression of the normally lowly expressed dominant negative beta isoform which interferes with the function of the predominant alpha isoform[69-71]. Six families with hypopituitarism had variants within the identified ESS region – two of which were synonymous variants, and all variants associated with hypopituitary patients within this study increased the use of the repressive beta isoform by breaking the regulatory motif. Some of the patient variants also intermediately increased exon 2 skipping which also encodes an isoform with a dominant negative effect on the normally predominant alpha isoform[144]. Thus, the high throughput splicing

screen simultaneously exposed clinically relevant, splice disruptive variants and provided some mechanistic insights into the molecular cause of the patients' phenotype.

In Chapter 3, exon 9 of another clinically relevant transcription factor in kidney and reproductive tissue, *WT1,* was likewise systemically tested for splice-disruptive variants[56]. Unlike *POU1F1* exon 2 which has competing alternate acceptors, *WT1* has a set of alternate donors which express the KTS- and KTS+ isoforms. Both isoforms are normally expressed at a ratio of ~2:1 (KTS+:KTS-) in healthy individuals[73,76], but lowered expression of KTS- can lead to a sexual differentiation phenotype[179] and, on the other hand, reduced expression of KTS+ leads to Frasier's syndrome which impacts kidney function and sexual differentiation[73,74,76,172-175]. Using a high throughput splicing assay, we identified 57 splice disruptive SNVs in and around *WT1* exon 9 (*n*=57/518; 11.0%), and the proportion of disruptive SNVs was similar to that observed in *POU1F1* exon 2. Of the splice altering variants, the majority altered the KTS ratio (n=43/57; 75.4%) including 8 variants which were previously observed to reduce KTS+ usage in Frasier's syndrome and a related renal phenotype, focal segmental glomerulosclerosis (FSGS)[74,76,172-175], and two SNVs which lowered KTS- use and were observed in patients with 46,XX ovotesticular differences in sexual development (46,XX OTDSD)[179]. The pooled minigene screen revealed an additional 16 KTS+ reducing SNVs expected to cause Frasier's syndrome and 17 SNVs lowering the usage of KTS- implicating a sexual differentiation phenotype. Four of the variants lowering the KTS ratio were synonymous so may not be immediately identified as potential causal variants during whole genome sequencing. Unlike *POU1F1* exon 2, almost all variants altering splicing clustered near the splice sites – so, either the splicing fidelity of *WT1* exon 9 is not

reliant on intronic or exonic splicing regulatory motifs, or alternatively, those motifs may be robust to perturbation by individual SNVs. Both Chapters 2 and 3 employed massively parallel splicing screens to expedite the discovery of splice altering variants within clinically relevant exons, which have not only contributed to the interpretation of standing clinical variants, but will also serve as a reference of functional evidence to guide the interpretation of future variants potentially shortening patients' diagnostic odysseys.

Although high throughput functional assays allow the systematic measurement of hundreds of variants simultaneously, the vast scale of variants without a clear interpretation makes the task of experimentally determining splice altering variants daunting. Thus, accurate and reliable computational prediction of splice disruptive variants may be the most feasible path to splicing assessments genome-wide. In Chapter 4, I benchmarked eight contemporary bioinformatic algorithms[90,91,97-101,105] to establish the state of the art in splice prediction, and to determine any specific areas for improvement. Since clinically derived datasets typically have an overabundance of variants at canonical splice sites, which are straightforward to predict, I employed data from four massively parallel splicing assays[55-57,59] and one saturation genome editing experiment[48] to evaluate the splice prediction tools. Datasets derived from such saturation screens represent a proportion of variant types and locations similar to that seen across the transcriptome making them ideal to evaluate algorithms' performance within different variant classes. I also tested the tools against one literature curated dataset of *MLH1* variants to underscore the differences seen when benchmarking against clinical variants instead of data from saturation screens. In concordance with

other recent benchmarking studies[102-104,107], I found SpliceAI[100] and the more recently developed tool Pangolin[101] – both of which are deep learning tools trained only on annotated sequence - to outperform other algorithms within most of my evaluation metrics. Other algorithms functioned well at specific tasks – for instance, ConSpliceML[105] showed high sensitivity to select splice altering variants and MMSplice[97] detected intronic splice disruptive variants when they were both tested using a transcriptome normalized threshold. However, all algorithms exhibited lower performance within exons compared to introns, highlighting the ongoing challenge of computationally identifying splice disruptive missense and synonymous variants. Similar trends have also been observed by other groups[61,198,199]. Thus, although Pangolin and SpliceAI appear to be the most sophisticated of the benchmarked tools, the motivation to continue to improve on computational splice effect predictive tools remains. As more massively parallel splicing assays become available, experimentally measured splicing effects could be used to bolster splice prediction when used as features within machine learning frameworks and to train and evaluate the algorithms.

Chapter 4 also emphasized the difficulties encountered within four tools – MMSplice[97], SQUIRLS[91], SpliceAI[100], and Pangolin[101] – which all rely on user input annotation files. Both MMSplice and Pangolin allow more flexibility in customizing annotations – either by accepting standard formats or by providing a script to create the required file type from a standard format respectively. Since SQUIRLS required file format is not easily reproducible, the versatility of the software may degrade over time as updated gene annotations become available. Although creating a custom annotation file for SpliceAI is not cumbersome, computing splice predictions for copious variants

using SpliceAI is computationally intensive leading many users to rely on the table of pre-computed scores. I highlighted some specific examples of where SpliceAI's pre-computed scores are problematic and there are indications that these difficulties may impact a non-trivial proportion of protein coding genes genome-wide (18.0%). Pangolin is similarly computationally intensive to run, and does not yet provide pre-computed splice predictions, however some of the computational burden can be abated by using the given GPU installation.

As a path forward, the next generation of splice prediction tools could apply knowledge of regional conservation to prioritize splice sites of functional import instead relying on user supplied annotation files. For instance, the beta region of *POU1F1* exon 2 and the KTS- region of *WT1* exon 9 are both highly conserved, so variants altering the usage of nearby splice sites could be assigned a stronger prior probability of pathogenicity. ConSpliceML[105], which does not require an annotation input, recently attempted to define regional constraints using SpliceAI predictions and healthy population variants from the gnomAD database. They endeavored to use their constraint metric to infer exons which might be more or less tolerant splice disruption, similar to constraint scores that compare, for each gene, the counts of protein-truncating variants observed in a population versus the counts expected in the absence of selection[95,244]. Within my benchmarking efforts, ConSpliceML was consistently ranked in the top three algorithms for sensitivity at various transcriptomic normalized thresholds (5%, 10%, 20%, and the area under the curve across thresholds) both across all variants and within exons. However, when examining the specificity across the same transcriptomic normalized thresholds and regions, ConSpliceML median scores rank

within the bottom three of the benchmarked tools (specificity AUC = .32 overall and .57 within exons; **Figure 5-1**). Although there is necessarily a compromise between sensitivity and specificity, these results suggest that ConSpliceML may be deeming too many variants as splice disruptive. So, there is room for improvement for the next generation of splice prediction algorithms to use regional constraints to define functional splice sites and exons.

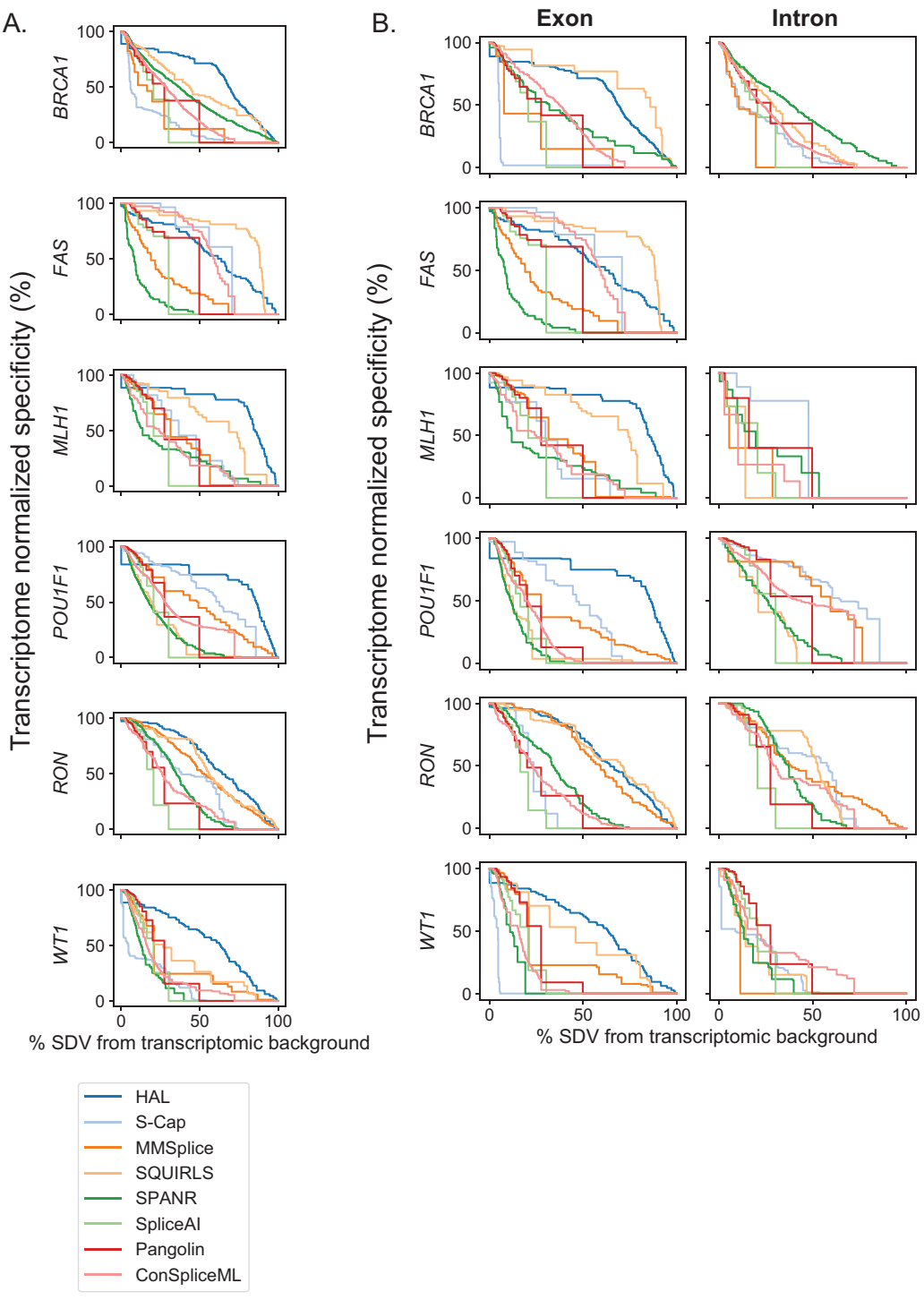


Figure 5-1: Specificity by transcriptomic proportion deemed splice disruptive.

A. Transcriptomic normalized specificity versus % of the genomic background set deemed SDV at varying splice predictor cutoff scores, for each of the benchmark datasets (rows). Lines are colored by bioinformatic algorithm. Transcriptome normalized specificity (y axis value) at *x*% background set SDV is defined as the specificity observed vs the given benchmark dataset, at the algorithm-specific score

cutoff at which x% of the background set of random exonic/near-exonic variants are deemed SDV. (Specificity=TN/(FP+TN)). **B**. Same as A., but split by variant position. Plots for variants within exons (left) and variants within introns (right) are shown.

The lowered prediction accuracy within exons compared to introns could be due to a lack of understanding of the splicing code and in particular the influence of RBP motifs on splicing within exons. There have been copious computational and experimental attempts to enumerate splicing regulatory elements outside of the splice site and branchpoint regions[17,24,25,27-29], but the resulting lists of critical motifs have little overlap and encompass nearly every possible short regulatory motif[30]. Recently SQUIRLS[91] used experimental measurements of a saturation set of 6-mers placed within different regions of two exons as part of a high throughput splicing assay[17]. Although SQUIRLS fared no better than the other algorithms at splice prediction within exons[243], the addition of CLIP-Seq derived binding motifs for known splicing factors could potentially improve exonic splice predictions for the next generation of splicing tools.

Although this dissertation focuses on cis sequence elements that alter splicing, many other factors including secondary RNA structure, nucleosome positioning, and transcriptional elongation to name a few can influence splicing decisions in vivo. Many research projects including this dissertation have sought to define the splicing code – a term coined by Barash, et al[245] – but the number of complex and often competing factors which determine and maintain splicing may be too vast to comprehend outside the context of artificial intelligence. Although SpliceAI and Pangolin were trained only with annotated sequence, the long context used to predict splicing effects may be able to decipher more complex events than simply the effects of sequence alteration. For instance, SpliceAI's performance was markedly lower when given shorter sequence

context, and the algorithm's predictions were correlated with nucleosome occupancy implying that SpliceAI can predict nucleosome positioning from primary sequence[100]. Thus, at the risk of overfitting models, supplying increasingly complex and highly layered artificial intelligence algorithms with our ever expanding databases like genome-wide CLIP-Seq peaks for instance may be the future of finally unraveling the splicing code. Although improvements within in silico splice prediction are warranted, the current tools can immediately be used to prioritize variants for mini-gene experiments within clinical datasets.

Massively parallel splicing assays and splice prediction algorithms can be used in tandem to not only expedite the classification of clinical variants of unknown significance, but they can also be paired to explore interactions, to target high priority exons, and to identify likely variants to create a desired phenotype in animal models. In the context of variant effects, epistasis occurs when variants' effects are non-additive – that is, when the combination of two variants' individual effects is not the same as the effects of those variants in combination. Splicing epistatic effects can occur via numerous mechanisms, including interactions between cis and trans-acting factors (inter-genic epistasis)[246], dual amino acid substitutions within individual trans-acting splicing factors[247], and between two cis-acting sequence variants[57-60,248]. In keeping with the theme of this dissertation, I focus here on the latter: pairs of cis-acting sequence variants which alter splicing of the exon in which they reside.

Cis-acting epistatic splicing effects imply a non-linear interaction among variants' splicing impacts – for instance, two splice neutral variants could have a large perturbation on splicing when paired together on the same haplotype. It has been

previously shown that compared to alternatively spliced exons, constitutively included exons have a higher density of exonic splice enhancing (ESE) regulatory motifs[17,19,89] and that those motifs are more robust to perturbation by SNVs[19] suggesting that some regulatory motifs may require disruption from two variants before splicing efficiency is impacted and representing a possible source of splicing epistasis. Due to purifying selection, common SNPs are unlikely to alter splicing individually, but common SNPs in cis with other individually splice neutral SNVs could combine epistatically in rare instances.

Although the extent of splicing epistasis is as yet undetermined, massively parallel splicing assays have been enlisted to directly measure splicing interactions among SNVs. Within the *FAS* exon I benchmarked in Chapter 4, the splicing effects every possible pair of SNVs as well as the SNVs themselves were measured[57]. They observed a broader distribution of splicing effects for the double nucleotide variants (DNVs) and a larger impact on splicing when two splice neutral SNVs were paired compared to their individual splicing effects. Across all measured SNVs, 40% were found to interact with a neutral SNV and, although these interactions were both distal and proximal, the set was enriched for proximal pairs suggesting that pairs of variants may be breaking the same splicing regulatory motif and thus altering its usage. However, the same group later explored all the possible combinations of twelve variants within the same *FAS* exon that differentiates the human exon from that of primates and found a much lower proportion of epistasis among DNVs (12%)[58]. They also only observed interaction for DNVs within six bp of each other again implying the epistasis could be a result of two hits within the same exonic regulatory motif, and another group

similarly found only proximal interactions within their set of DNVs measured within *WT1* exon 5[60]. In contrast, a linear model of additive mutation effects which ignored any possible interactions among SNVs had an excellent fit to the measured data within *RON* exons 10-12[59]. So, sensitivity to epistatic splicing effects could be exon specific or, like sensitivity to splice disruption by SNVs, could be related to the wild-type inclusion levels of the exon[19].

Recently, the authors of Pangolin demonstrated the tool's ability to model epistatic interactions computationally[101]. Their splicing predictions for DNVs and higher order combinations of variants from the study of the *FAS* ancestral exon[58] had a higher correlation with the measured values than a linear combination of the predicted scores for each of the SNVs individually suggesting that *in silico* modeling of splicing epistasis is possible. To my knowledge, SpliceAI has not been benchmarked against any measured sets of DNVs or higher order combinations of variants, but the software has the capability to score multiple variants on the same haplotype. I wrote a custom wrapper to the SpliceAI software allowing me to easily evaluate *MLH1* – one of the most highly studied genes implicated in the colon cancer predisposition disorder of Lynch syndrome[249] – for possible epistatic splicing effects. I computationally predicted the splicing effects of every possible SNV individually, and all possible pairs of SNVs within short (<300 bp), coding exons of *MLH1.* I focused in particular on exons which would create an out of frame transcript if skipped (*n*=14 exons). Since Lynch syndrome has autosomal dominant inheritance and out of frame transcripts would presumably be subject to nonsense mediated decay[204,249,250], any variants promoting exon skipping in those exons would likely be pathogenic.

Within the pairs of in cis SNVs from out of frame skip exons, the DNVs had higher odds of causing predicted splice disruption (SpliceAI threshold >= 20%) compared to SNVs in 67% (*n*=8/12) of the selected exons (median OR=2.09) as has been previously observed[57]. Of the predicted splice disruptive DNVs in selected exons, only 16.0% (*n*=8,582/53,605) of the SNVs were within 10 bp of each other so a high proportion of the variants predicted to act epistatically did so at a greater distance than has been observed in recent high throughput assays[58,60].

Computationally predicted maps of SNV splicing effects can nominate exons with regulatory regions sensitive to disruption; these exons could then be prioritized for future MPSAs. For instance, within *MLH1* exon 6, SpliceAI predictions suggest the existence of ESE region(s) on the 5' end of the exon (**Figure 5-2**). Since skipping of exon 6 would create an out of frame transcript, variants promoting exon skipping within the putative ESE regions would be pathogenic. Within exon 6, four synonymous variants predicted to be splice disruptive are currently classified as likely benign/benign (LBB) or with an unclear designation within ClinVar (**Figure 5-2**). Exons 9 and 15 also show evidence of predicted exonic splice regulatory regions, and both exons have a handful of predicted splice altering synonymous variants with LBB or uncertain/conflicting interpretations within ClinVar (*n*=4 and 5, respectively). Since synonymous variants may not be suspected to cause loss of function a priori, computational predictions can be used to prioritize future massively parallel splicing

screens to identify clinically relevant pathogenic variants that may otherwise be inscrutable.

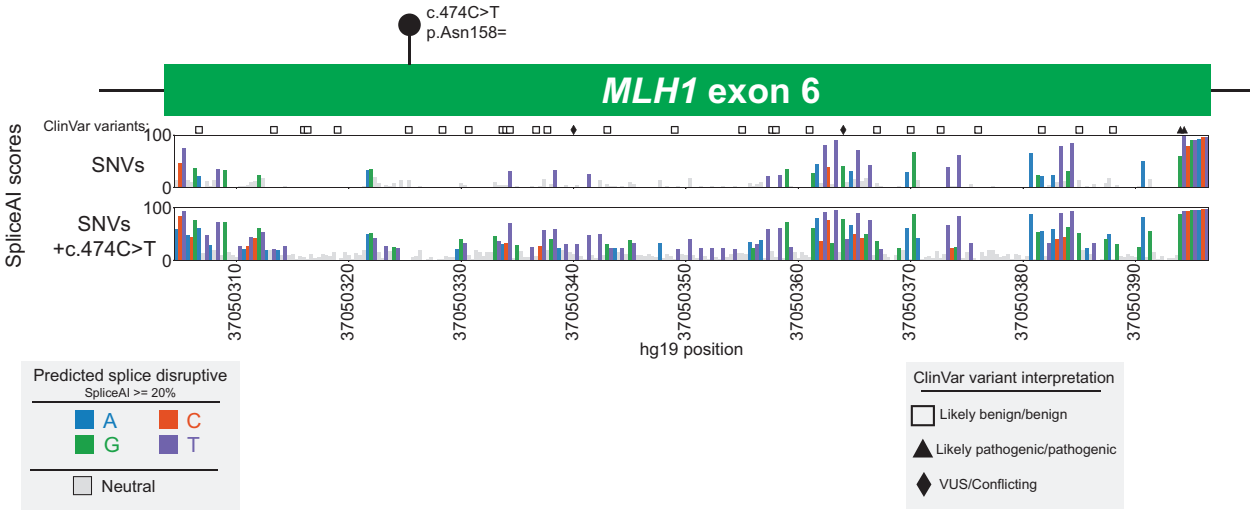


Figure 5-2 SpliceAI predicted splice disruption in *MLH1* exon 6.

Predicted splicing map showing SpliceAI's probability of splice disruption (y-axis) for every exonic *MLH1* exon 6 SNV (top) and every SNV in cis with the common variant c.474C>T (bottom) by transcript position (x-axis). Bars colored by the nucleotide substitution are predicted to be splice disruptive variants (SDV) at a SpliceAI threshold >= 20% and gray variants are predicted to be splice neutral. Synonymous ClinVar variants are indicated above the plot. Black lollipop shows the common gnomAD variant (gnomAD allele frequency > 1%) at c.474C>T.

I next overlaid allele frequencies from gnomAD to explore splicing epistasis between common variants (gnomAD allele frequency (AF) > 1%) and other SNVs within out of frame skip exons. Specifically, I asked whether any common variants might act to sensitize exons to splicing disruption by a second variant *in cis*. As a specific example, the synonymous substitution c.474C>T is listed as likely benign in ClinVar, is common within gnomAD (AF = 1.06%), and is predicted to have no effect on splicing (SpliceAI score = 14.0%; **Figure 5-2**). When paired with the common c.474C>T variant, 77 SNVs that are predicted to have no effect on splicing individually are expected to have at least a 20% probability of altered splicing (28.0% of all possible SNVs) and of those, seven

have a > 50% predicted chance of disrupting splicing (2.5% of all SNVs; **Figures 5-2** and **5-3**). One of the seven high probability neutral-common neutral DNVs, c.534A>G (p.Glu178=), is a synonymous substitution currently classified as likely benign in ClinVar so the possibility of two variants presenting as likely benign individually but pathogenically altering splicing when on the same haplotype would be an interesting result. Several of the neutral SNVs within the high probability neutral-common neutral SNV pairs are located within the interior of the exon (median distance from nearest splice site=7 bp; max=32 bp) and all are located > 10 bp from the common variant (minimum distance=12 bp, median=40 bp) raising questions as to the mechanisms of the epistatic effects. However, these are only computational predictions and will require experimental investigation to verify the effects and further explore possible mechanisms of splicing epistasis. Since epistatic splicing effects with common variants could have broad clinical implications, my computational analysis prioritizes *MLH1* exon 6 as a promising future target for a high throughput splicing screen.

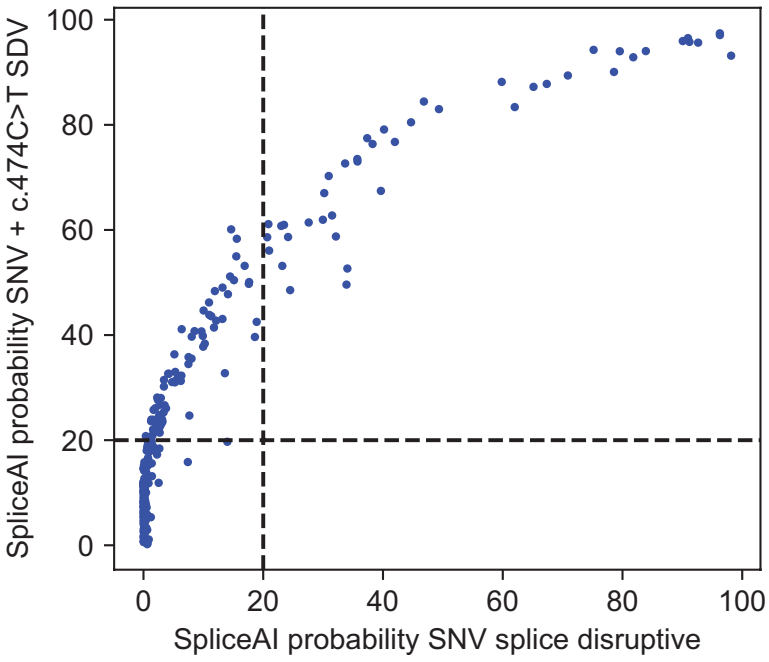


Figure 5-3: SpliceAI predictions of *MLH1* exon 6 SNVs individually and paired with common gnomAD variant.

Scatterplot showing the SpliceAI predictions of splice disruption within *MLH1* exon 6 SNVs (x-axis) and SpliceAI predictions of splice disruption for every SNV in cis with c.474C>T on the same haplotype (y-axis). Dashed horizontal and vertical lines show the SpliceAI threshold used to deem variants as splice disruptive.

Computational splicing predictions can also be used to identify likely variants to create a desired phenotype in animal models. One example is desmoplakin , a cell to cell binding protein which links desmosomes to intermediate filaments[251]. Missense and truncating mutations in *DSP* can cause recessive and dominant forms of cardiomyopathy[252,253]. Most pathogenic *DSP* variants cause protein truncations and lead to left dominant arrhythmogenic cardiomyopathy, early mortality, and specific skin and hair phenotypes with dominant inheritance[254]. The truncating variants present uniformly across the gene and phenotypic severity is correlated with RNA and protein expression levels implicating haploinsufficiency as the disease mechanism[254]. Although *DSP* has a mouse ortholog, there is currently no animal model of the cardiac phenotype caused by variants in *DSP*. One challenge in creating a mice model of *DSP* driven cardiomyopathy has been dosage sensitivity – loss of function truncation variants in the homozygous state are embryonic lethal and heterozygous mice do not display the cardiac phenotype. As a result, available mouse models are limited by the inability to mirror the expression defect that reaches a level of pathogenicity similar to that in human. Judicious selection of splicing variants, including those beyond canonical sites, might provide the ability to fine-tune *DSP* dosage to create models with the desired phenotype. I used my custom SpliceAI wrapper on mouse genomic sequence to identify probable variants promoting skipping of out of frame exons at various predicted rates across the range of SpliceAI predicted probabilities. Thus, we could computationally

hone the dosage in a gene acting through haploinsufficiency to accelerate the development of an animal model of disease. In this way, the mechanisms of the disease and possible treatments can be better understood to inform treatment of cardiomyopathy patients. So, accurate computational predictions of splicing can be used as a guide to hasten scientific discoveries within human diseases.

Altered splicing explains a substantial minority of the burden of pathogenic variants in human genetic disease. By applying massively parallel splicing assays to two clinically relevant exons, we have shortened the diagnostic odyssey for some patients and uncovered some of the mechanisms of splicing perturbations within those exons. Pairing experimentally measured splicing results with computational predictions of splicing effects could be a path forward to better understanding the complete splicing code, and we have employed massively parallel assays to identify the state of the art in computational splicing prediction. Although computational predictions may be the only feasible route to ascertain splicing effects genome-wide, massively parallel assays are critical to hone and validate those predictions. Although only a small part of the work necessary, this dissertation aims to move towards deciphering the splicing code as it relates to the identification of pathogenic variants in clinically relevant disease genes so patients can receive explicit genetic diagnostics leading to a definite course of treatment.

Appendix

Table A-1: *MLH1* literature curated variants

chrom	hgvs variant	hg19 pos	hg38 pos	ref	alt	PubMedID(s)	SDV
3	c.117-34A>T	37038076	36996585	A	T	24090359	FALSE
3	c.117-11T>A	37038099	36996608	T	A	11066084	TRUE
3	c.117-2A>G	37038108	36996617	A	G	32849802	TRUE
3	c.117-2A>T	37038108	36996617	A	T	8521394	TRUE
3	c.117-1G>C	37038109	36996618	G	C	12624141, 19224586.0	TRUE
3	c.121G>C	37038114	36996623	G	C	30233647, 18561205.0	FALSE
3	c.122A>G	37038115	36996624	A	G	15300854, 26247049.0, 32123317.0	TRUE
3	c.146T>A	37038139	36996648	T	A	16395668	FALSE
3	c.191A>G	37038184	36996693	A	G	31332305, 23729658.0	FALSE
3	c.198C>T	37038191	36996700	C	T	19267393, 22949379.0	FALSE
3	c.199G>A	37038192	36996701	G	A	16395668, 16995940.0, 18561205.0, 22949379.0, 9833759.0	FALSE
3	c.199G>T	37038192	36996701	G	T	16395668, 10480359.0	FALSE
3	c.200G>A	37038193	36996702	G	A	16995940	FALSE
3	c.207+1G>T	37038201	36996710	G	T	15849733	TRUE
3	c.207+2T>C	37038202	36996711	T	C	16142001, 21642682.0	TRUE
3	c.207+2T>G	37038202	36996711	T	G	22480969	TRUE
3	c.208-3C>G	37042443	37000952	C	G	19267393	TRUE
3	c.208-2A>G	37042444	37000953	A	G	8521398	TRUE
3	c.208-1G>A	37042445	37000954	G	A	18931482	TRUE

3	c.210A>C	37042448	37000957	A	C	15923275	FALSE
3	c.214G>C	37042452	37000961	G	C	15923275	TRUE
3	c.214G>T	37042452	37000961	G	T	15923275	TRUE
3	c.216T>C	37042454	37000963	T	C	15923275	FALSE
3	c.218T>C	37042456	37000965	T	C	22736432	FALSE
3	c.229T>C	37042467	37000976	T	C	18561205, 16395668.0	FALSE
3	c.230G>A	37042468	37000977	G	A	19669161	FALSE
3	c.238T>G	37042476	37000985	T	G	32849802	FALSE
3	c.244A>G	37042482	37000991	A	G	22736432	FALSE
3	c.277A>G	37042515	37001024	A	G	26247049	FALSE
3	c.292G>A	37042530	37001039	G	A	18561205	FALSE
3	c.299G>C	37042537	37001046	G	C	32849802, 23729658.0	FALSE
3	c.301G>A	37042539	37001048	G	A	18561205, 27629256.0	TRUE
3	c.302G>A	37042540	37001049	G	A	18561205, 30233647.0	FALSE
3	c.303T>G	37042541	37001050	T	G	22949379	FALSE
3	c.304G>A	37042542	37001051	G	A	16395668, 12183410.0, 23729658.0	TRUE
3	c.305A>C	37042543	37001052	A	C	31642931	TRUE
3	c.306+1G>A	37042545	37001054	G	A	15849733, 10471527.0	TRUE
3	c.306+4A>G	37042548	37001057	A	G	18561205, 32634176.0	TRUE
3	c.306+5G>A	37042549	37001058	G	A	16142001, 20858721.0, 30233647.0, 23523604.0	TRUE
3	c.307-29C>A	37045863	37004372	C	A	19267393, 22949379.0	FALSE
3	c.307-19A>G	37045873	37004382	A	G	18561205	FALSE
3	c.307-2A>C	37045890	37004399	A	C	12655568	TRUE
3	c.307-1G>C	37045891	37004400	G	C	14517962	TRUE
3	c.318C>A	37045903	37004412	C	A	16395668	FALSE
3	c.318C>G	37045903	37004412	C	G	29505604	FALSE
3	c.320T>G	37045905	37004414	T	G	16995940, 29505604.0, 8776590.0	FALSE

3	c.326A>C	37045911	37004420	A	C	29505604	FALSE
3	c.331G>C	37045916	37004425	G	C	29505604	FALSE
3	c.332C>A	37045917	37004426	C	A	29505604	FALSE
3	c.332C>T	37045917	37004426	C	T	29505604, 10777691.0	FALSE
3	c.338T>A	37045923	37004432	T	A	16395668, 29505604.0	FALSE
3	c.346A>C	37045931	37004440	A	C	29505604	FALSE
3	c.347C>A	37045932	37004441	C	A	18561205	FALSE
3	c.350C>G	37045935	37004444	C	G	29505604	FALSE
3	c.350C>T	37045935	37004444	C	T	16395668, 18561205.0, 19267393.0, 22949379.0, 29505604.0, 31332305.0, 11139242.0, 8574961.0, 10732761.0, 10480359.0	FALSE
3	c.375A>G	37045960	37004469	A	G	16395668, 9718327.0	FALSE
3	c.376T>A	37045961	37004470	T	A	18561205	FALSE
3	c.380G>A	37045965	37004474	G	A	11112663	TRUE
3	c.380+1G>A	37045966	37004475	G	A	15849733, 12555990.0	TRUE
3	c.380+2T>A	37045967	37004476	T	A	12655568	TRUE
3	c.380+2T>C	37045967	37004476	T	C	18726168	TRUE
3	c.381-2A>G	37048480	37006989	A	G	8971183, 10375096.0, 12624141.0, 21642682.0, 15024732.0	TRUE
3	c.389A>G	37048490	37006999	A	G	22949379	FALSE
3	c.394G>C	37048495	37007004	G	C	29505604, 23729658.0	FALSE
3	c.403C>G	37048504	37007013	C	G	29505604	FALSE
3	c.438A>G	37048539	37007048	A	G	32849802	FALSE
3	c.453G>A	37048554	37007063	G	A	32634176	TRUE
3	c.453+1G>T	37048555	37007064	G	T	18931482, 19224586.0	TRUE
3	c.453+1G>A	37048555	37007064	G	A	24278394	TRUE
3	c.453+2T>C	37048556	37007065	T	C	11920650	TRUE
3	c.454-51T>C	37050254	37008763	T	C	8776590	FALSE

3	c.454-13A>G	37050292	37008801	A	G	23729658, 32849802.0	TRUE
3	c.454-2A>G	37050303	37008812	A	G	16395668	TRUE
3	c.454-1G>A	37050304	37008813	G	A	10200055, 8776590.0, 15235038.0, 15342696.0	TRUE
3	c.454-1G>C	37050304	37008813	G	C	12658575	TRUE
3	c.454-1G>T	37050304	37008813	G	T	9593786	TRUE
3	c.464T>G	37050315	37008824	T	G	16341550	FALSE
3	c.543C>G	37050394	37008903	C	G	26247049	TRUE
3	c.543C>T	37050394	37008903	C	T	28334867	TRUE
3	c.544A>G	37050395	37008904	A	G	16395668, 19459153.0, 10480359.0, 21642682.0	TRUE
3	c.545G>A	37050396	37008905	G	A	26247049	TRUE
3	c.545+1G>A	37050397	37008906	G	A	15849733, 19669161.0	TRUE
3	c.545+2T>A	37050398	37008907	T	A	21590452	TRUE
3	c.545+3A>G	37050399	37008908	A	G	9218993, 31332305.0, 15253764.0, 24278394.0	TRUE
3	c.546-2A>C	37053309	37011818	A	C	16830052	TRUE
3	c.546-2A>G	37053309	37011818	A	G	8521398, 16451135.0, 12052501.0, 10471527.0, 10732761.0, 12658575.0, 24278394.0, 21642682.0	TRUE
3	c.546-1G>A	37053310	37011819	G	A	32849802, 15342696.0	TRUE
3	c.554T>G	37053319	37011828	T	G	8808596	FALSE
3	c.572G>T	37053337	37011846	G	T	29505604	FALSE
3	c.577T>C	37053342	37011851	T	C	29505604	FALSE
3	c.588+1G>T	37053354	37011863	G	T	22949379	TRUE
3	c.588+2T>A	37053355	37011864	T	A	21681552	TRUE
3	c.588+2T>C	37053355	37011864	T	C	19224586	TRUE
3	c.588+5G>A	37053358	37011867	G	A	15713769, 16341550.0, 18561205.0, 24090359.0, 15926618.0	TRUE

3	c.588+11G>C	37053364	37011873	G	C	18561205, 9718327.0	FALSE
3	c.589-10T>A	37053492	37012001	T	A	15926618	TRUE
3	c.589-2A>C	37053500	37012009	A	C	32849802	TRUE
3	c.589-2A>G	37053500	37012009	A	G	19267393, 10882759.0, 10422993.0, 22949379.0	TRUE
3	c.589-1G>T	37053501	37012010	G	T	12658575	TRUE
3	c.595G>C	37053508	37012017	G	C	29505604	TRUE
3	c.637G>A	37053550	37012059	G	A	16395668, 18561205.0, 29505604.0	FALSE
3	c.644A>G	37053557	37012066	A	G	18561205	FALSE
3	c.647T>G	37053560	37012069	T	G	18561205	FALSE
3	c.649C>T	37053562	37012071	C	T	29505604, 11920458.0, 16425354.0	FALSE
3	c.655A>G	37053568	37012077	A	G	8776590, 10777691.0, 10882759.0, 9718327.0	FALSE
3	c.677G>A	37053590	37012099	G	A	16341550, 18561205.0, 10422993.0, 12373605.0, 29505604.0, 16736289.0, 8571956.0	TRUE
3	c.677G>T	37053590	37012099	G	T	16451135, 29505604.0, 12658575.0	TRUE
3	c.677+1G>A	37053591	37012100	G	A	24278394	TRUE
3	c.677+1G>T	37053591	37012100	G	T	12624141, 15342696.0	TRUE
3	c.677+3A>T	37053593	37012102	A	T	24090359	TRUE
3	c.677+3A>C	37053593	37012102	A	C	15365996	TRUE
3	c.677+3A>G	37053593	37012102	A	G	15713769, 15849733.0, 19669161.0	TRUE
3	c.678-1G>C	37055922	37014431	G	C	22949379	TRUE
3	c.678-1G>T	37055922	37014431	G	T	10471527	TRUE
3	c.702G>A	37055947	37014456	G	A	22736432, 22949379.0	FALSE
3	c.731G>A	37055976	37014485	G	A	16995940, 9218993.0	FALSE
3	c.739T>C	37055984	37014493	T	C	16395668	FALSE
3	c.778C>T	37056023	37014532	C	T	18561205	FALSE
3	c.779T>G	37056024	37014533	T	G	27629256, 10882759.0	FALSE

3	c.790+1G>T	37056036	37014545	G	T	15342696	TRUE
3	c.790+1G>A	37056036	37014545	G	A	15713769, 16395668.0, 16142001.0, 15955785.0, 17440950.0, 12624141.0, 20305446.0, 12658575.0, 19224586.0, 21642682.0, 15342696.0	TRUE
3	c.790+2T>A	37056037	37014546	T	A	15365995, 7757073.0, 32849802.0	TRUE
3	c.790+2T>C	37056037	37014546	T	C	15713769, 12624141.0	TRUE
3	c.790+3A>T	37056038	37014547	A	T	12373605, 8808596.0, 19224586.0, 21642682.0	TRUE
3	c.790+4A>G	37056039	37014548	A	G	16341550, 10323887.0	TRUE
3	c.790+4A>T	37056039	37014548	A	T	20717847	TRUE
3	c.790+5G>T	37056040	37014549	G	T	18561205, 22766992.0	TRUE
3	c.790+10A>G	37056045	37014554	A	G	18561205, 18561205.0, 22949379.0	FALSE
3	c.791-7T>A	37058990	37017499	T	A	22736432, 30233647.0	TRUE
3	c.791-5T>G	37058992	37017501	T	G	16395668, 18561205.0	TRUE
3	c.791-3T>G	37058994	37017503	T	G	12624141	TRUE
3	c.791-2A>G	37058995	37017504	A	G	11606497, 12624141.0, 21642682.0	TRUE
3	c.791-1G>C	37058996	37017505	G	C	8571956, 26247049.0, 22949379.0	TRUE
3	c.791-1G>T	37058996	37017505	G	T	10422993	TRUE
3	c.791A>G	37058997	37017506	A	G	26761715	FALSE
3	c.793C>A	37058999	37017508	C	A	26247049, 26761715.0	TRUE
3	c.793C>T	37058999	37017508	C	T	16995940, 18561205.0, 15713769.0, 26247049.0, 26761715.0, 32849802.0, 31332305.0	TRUE
3	c.794G>A	37059000	37017509	G	A	18561205, 26761715.0, 8993976.0	FALSE
3	c.803A>G	37059009	37017518	A	G	26761715	FALSE
3	c.814T>G	37059020	37017529	T	G	26761715	FALSE
3	c.815T>C	37059021	37017530	T	C	26761715	FALSE

3	c.842C>T	37059048	37017557	C	T	16995940, 26761715.0	TRUE
3	c.845C>G	37059051	37017560	C	G	26761715	TRUE
3	c.855C>T	37059061	37017570	C	T	26761715	FALSE
3	c.856A>C	37059062	37017571	A	C	26761715	FALSE
3	c.861C>T	37059067	37017576	C	T	26761715	FALSE
3	c.875T>C	37059081	37017590	T	C	16341550, 26761715.0	FALSE
3	c.882C>G	37059088	37017597	C	G	26761715	TRUE
3	c.882C>T	37059088	37017597	C	T	16395668, 18561205.0, 26247049.0, 16736289.0, 26761715.0	TRUE
3	c.883A>C	37059089	37017598	A	C	16451135, 26761715.0, 16830052.0	TRUE
3	c.883A>G	37059089	37017598	A	G	15713769, 26247049.0, 26761715.0	TRUE
3	c.884G>A	37059090	37017599	G	A	22949379, 26761715.0	TRUE
3	c.884G>C	37059090	37017599	G	C	26761715	TRUE
3	c.884+2T>C	37059092	37017601	T	C	15849733, 15955785.0	TRUE
3	c.884+3A>G	37059093	37017602	A	G	31642931	TRUE
3	c.884+4A>G	37059094	37017603	A	G	18561205, 17653898.0, 21034533.0, 21642682.0	TRUE
3	c.885-24T>A	37061777	37020286	T	A	18561205	FALSE
3	c.885-5G>T	37061796	37020305	G	T	18561205	FALSE
3	c.923A>C	37061839	37020348	A	C	32849802	TRUE
3	c.935A>C	37061851	37020360	A	C	32123317	FALSE
3	c.960G>C	37061876	37020385	G	C	18561205	FALSE
3	c.974G>A	37061890	37020399	G	A	22736432	FALSE
3	c.977T>C	37061893	37020402	T	C	18561205, 8574961.0, 10732761.0	FALSE
3	c.986A>C	37061902	37020411	A	C	31332305, 10323887.0	TRUE
3	c.1003C>T	37061919	37020428	C	T	18561205	FALSE
3	c.1013A>G	37061929	37020438	A	G	18561205, 31332305.0	FALSE
3	c.1037A>G	37061953	37020462	A	G	18561205, 23729658.0, 21642682.0	TRUE

3	c.1038G>A	37061954	37020463	G	A	12183410, 15555211.0	TRUE
3	c.1038G>C	37061954	37020463	G	C	16341550	TRUE
3	c.1038G>T	37061954	37020463	G	T	18561205, 21642682.0	TRUE
3	c.1038+1G>C	37061955	37020464	G	C	17440950, 21642682.0	TRUE
3	c.1039-8T>A	37067120	37025629	T	A	27629256, 24090359.0, 27629256.0, 15713769.0, 19669161.0	FALSE
3	c.1039-2A>G	37067126	37025635	A	G	15849733, 19669161.0	TRUE
3	c.1039-2A>T	37067126	37025635	A	T	31332305	TRUE
3	c.1039-1G>A	37067127	37025636	G	A	10200055	TRUE
3	c.1039-1G>T	37067127	37025636	G	T	21642682	TRUE
3	c.1043T>C	37067132	37025641	T	C	27629256	FALSE
3	c.1098G>T	37067187	37025696	G	T	15173238	FALSE
3	c.1136A>G	37067225	37025734	A	G	32849802, 23729658.0	FALSE
3	c.1139C>T	37067228	37025737	C	T	15173238	FALSE
3	c.1146G>C	37067235	37025744	G	C	15173238	FALSE
3	c.1147A>T	37067236	37025745	A	T	15173238	FALSE
3	c.1151T>A	37067240	37025749	T	A	16425354, 10777691.0, 21034533.0	FALSE
3	c.1166G>A	37067255	37025764	G	A	32849802	TRUE
3	c.1204A>C	37067293	37025802	A	C	15173238	FALSE
3	c.1217G>A	37067306	37025815	G	A	31332305, 27629256.0	FALSE
3	c.1242G>C	37067331	37025840	G	C	15173238	FALSE
3	c.1270G>C	37067359	37025868	G	C	15173238	FALSE
3	c.1283A>T	37067372	37025881	A	T	15173238	FALSE
3	c.1313C>T	37067402	37025911	C	T	15173238	FALSE
3	c.1339T>C	37067428	37025937	T	C	15173238	FALSE
3	c.1360G>C	37067449	37025958	G	C	16395668	FALSE
3	c.1361G>C	37067450	37025959	G	C	15173238	FALSE
3	c.1376C>T	37067465	37025974	C	T	15173238	TRUE

3	c.1383G>T	37067472	37025981	G	T	15173238	FALSE
3	c.1389A>C	37067478	37025987	A	C	15173238	TRUE
3	c.1401C>T	37067490	37025999	C	T	31332305	FALSE
3	c.1409+1G>A	37067499	37026008	G	A	17348456	TRUE
3	c.1409+1G>C	37067499	37026008	G	C	16451135, 10200055.0	TRUE
3	c.1418A>G	37070283	37028792	A	G	31332305	FALSE
3	c.1420C>T	37070285	37028794	C	T	18561205	FALSE
3	c.1421G>A	37070286	37028795	G	A	18561205	FALSE
3	c.1421G>C	37070286	37028795	G	C	27629256	FALSE
3	c.1558+1G>T	37070424	37028933	G	T	12658575, 19224586.0, 21642682.0	TRUE
3	c.1558+1G>A	37070424	37028933	G	A	31332305	TRUE
3	c.1558+11G>A	37070434	37028943	G	A	8863153	FALSE
3	c.1558+14G>A	37070437	37028946	G	A	16395668, 9718327.0, 27629256.0	FALSE
3	c.1559-3C>G	37081674	37040183	C	G	18566915	TRUE
3	c.1559-2A>C	37081675	37040184	A	C	10200055	TRUE
3	c.1559-2A>G	37081675	37040184	A	G	15555211, 12183410.0, 24278394.0, 21642682.0	TRUE
3	c.1559-2A>T	37081675	37040184	A	T	22949379	TRUE
3	c.1559-1G>A	37081676	37040185	G	A	15849733	TRUE
3	c.1559-1G>C	37081676	37040185	G	C	12624141, 31332305.0, 15926618.0	TRUE
3	c.1559-1G>T	37081676	37040185	G	T	8776590, 21642682.0	TRUE
3	c.1559T>G	37081677	37040186	T	G	19669161	FALSE
3	c.1569G>T	37081687	37040196	G	T	16995940	FALSE
3	c.1616C>A	37081734	37040243	C	A	18561205	FALSE
3	c.1633A>G	37081751	37040260	A	G	26247049	FALSE
3	c.1646T>C	37081764	37040273	T	C	18561205	FALSE
3	c.1652A>C	37081770	37040279	A	C	16395668, 18561205.0, 9833759.0	FALSE
3	c.1652A>G	37081770	37040279	A	G	32849802	FALSE

3	c.1653C>T	37081771	37040280	C	T	31332305	FALSE
3	c.1667G>T	37081785	37040294	G	T	15300854	TRUE
3	c.1667+1G>T	37081786	37040295	G	T	21642682	TRUE
3	c.1667+2T>C	37081787	37040296	T	C	21056691	TRUE
3	c.1668-19A>G	37083740	37042249	A	G	16395668, 8776590.0, 9718327.0	FALSE
3	c.1668-3C>T	37083756	37042265	C	T	9419403	TRUE
3	c.1668-2A>G	37083757	37042266	A	G	21056691	TRUE
3	c.1668-1G>A	37083758	37042267	G	A	19267393, 14635101.0	TRUE
3	c.1668-1G>T	37083758	37042267	G	T	15342696	TRUE
3	c.1676T>G	37083767	37042276	T	G	29505604	TRUE
3	c.1681T>C	37083772	37042281	T	C	29505604	FALSE
3	c.1685A>C	37083776	37042285	A	C	19669161, 29505604.0	FALSE
3	c.1731G>A	37083822	37042331	G	A	16451135, 16395668.0, 18561205.0, 19669161.0, 8808596.0, 24278394.0, 11112663.0, 19224586.0, 21642682.0, 14635101.0	TRUE
3	c.1731G>C	37083822	37042331	G	C	18931482	TRUE
3	c.1731+1G>C	37083823	37042332	G	C	16034045	TRUE
3	c.1731+2T>G	37083824	37042333	T	G	15849733	TRUE
3	c.1731+3A>T	37083825	37042334	A	T	20305446, 18769833.0	TRUE
3	c.1731+4A>G	37083826	37042335	A	G	24278394, 23729658.0, 23729658.0	TRUE
3	c.1731+5G>A	37083827	37042336	G	A	18561205, 19685281.0	TRUE
3	c.1731+6T>G	37083828	37042337	T	G	32849802	TRUE
3	c.1732-19T>C	37088991	37047500	T	C	27629256	FALSE
3	c.1732-9T>C	37089001	37047510	T	C	16395668	FALSE
3	c.1732-2A>T	37089008	37047517	A	T	18566915, 24090359.0, 9245993.0, 8571956.0	TRUE
3	c.1732-1G>A	37089009	37047518	G	A	15849733, 19267393.0, 15955785.0	TRUE
3	c.1742C>T	37089020	37047529	C	T	16425354	FALSE

3	c.1743G>A	37089021	37047530	G	A	27629256	FALSE
3	c.1754T>G	37089032	37047541	T	G	18561205	FALSE
3	c.1757C>A	37089035	37047544	C	A	16395668	FALSE
3	c.1763T>C	37089041	37047550	T	C	10777691	FALSE
3	c.1808C>G	37089086	37047595	C	G	18561205	FALSE
3	c.1814A>C	37089092	37047601	A	C	27629256	FALSE
3	c.1820T>A	37089098	37047607	T	A	18561205	FALSE
3	c.1855G>C	37089133	37047642	G	C	18561205	FALSE
3	c.1865T>A	37089143	37047652	T	A	20858721	FALSE
3	c.1896G>A	37089174	37047683	G	A	8571956	TRUE
3	c.1896G>C	37089174	37047683	G	C	17666659	TRUE
3	c.1896G>T	37089174	37047683	G	T	21642682, 15024732.0	TRUE
3	c.1896+1G>A	37089175	37047684	G	A	8993979, 10422993.0	TRUE
3	c.1896+1G>T	37089175	37047684	G	T	15849733, 10471527.0, 21642682.0	TRUE
3	c.1896+2T>C	37089176	37047685	T	C	15849733	TRUE
3	c.1918C>T	37090029	37048538	C	T	18561205, 23729658.0	FALSE
3	c.1919C>T	37090030	37048539	C	T	18561205	FALSE
3	c.1958T>G	37090069	37048578	T	G	16995940	FALSE
3	c.1959G>T	37090070	37048579	G	T	16395668, 18561205.0, 9718327.0	FALSE
3	c.1961C>T	37090072	37048581	C	T	19669161, 16995940.0	FALSE
3	c.1963A>G	37090074	37048583	A	G	18561205, 16995940.0	FALSE
3	c.1964T>C	37090075	37048584	T	C	31332305	FALSE
3	c.1967T>C	37090078	37048587	T	C	18561205	FALSE
3	c.1976G>C	37090087	37048596	G	C	16995940, 10534773.0	TRUE
3	c.1976G>T	37090087	37048596	G	T	10534773	TRUE
3	c.1984A>C	37090095	37048604	A	C	16341550, 31332305.0	TRUE
3	c.1988A>G	37090099	37048608	A	G	10732761, 23729658.0	TRUE
3	c.1989G>T	37090100	37048609	G	T	16395668, 18561205.0, 10480359.0	TRUE

3	c.1989G>A	37090100	37048609	G	A	32849802	TRUE
3	c.1989+1G>A	37090101	37048610	G	A	15849733, 14635101.0	TRUE
3	c.1989+1G>T	37090101	37048610	G	T	15849733, 10323887.0	TRUE
3	c.1989+5G>C	37090105	37048614	G	C	8993976	TRUE
3	c.1989+6T>G	37090106	37048615	T	G	32123317	FALSE
3	c.1990-3C>G	37090392	37048901	C	G	15993273	TRUE
3	c.1990-2A>G	37090393	37048902	A	G	15365996	TRUE
3	c.1990-1G>T	37090394	37048903	G	T	15849733, 19267393.0	TRUE
3	c.1990-1G>A	37090394	37048903	G	A	11920650, 22949379.0	TRUE
3	c.1996T>C	37090401	37048910	T	C	18561205, 23729658.0	FALSE
3	c.2027T>C	37090432	37048941	T	C	19669161	FALSE
3	c.2041G>A	37090446	37048955	G	A	11139242, 19669161.0, 8880570.0, 18561205.0	FALSE
3	c.2059C>T	37090464	37048973	C	T	27629256, 11139242.0, 16395668.0, 19267393.0, 22949379.0, 11920458.0	FALSE
3	c.2066A>G	37090471	37048980	A	G	18561205	FALSE
3	c.2103G>A	37090508	37049017	G	A	26247049	TRUE
3	c.2103G>C	37090508	37049017	G	C	15849733, 16341550.0	TRUE
3	c.2103+1G>A	37090509	37049018	G	A	12658575, 8571956.0, 21642682.0	TRUE
3	c.2103+1G>T	37090509	37049018	G	T	15342696	TRUE
3	c.2103+3A>G	37090511	37049020	A	G	31642931, 23729658.0, 21642682.0	TRUE
3	c.2104-2A>T	37091975	37050484	A	T	10375096	TRUE
3	c.2104-2A>G	37091975	37050484	A	G	32849802, 15365995.0	TRUE
3	c.2104-1G>A	37091976	37050485	G	A	14635101	TRUE

References

1. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. Jan 2016;17(1):19-32. doi:10.1038/nrg.2015.3
2. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem*. 2015;84:291-323. doi:10.1146/annurev-biochem-060614-034316
3. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*. 2013;4(1):61-76. doi:10.1002/wrna.1141
4. Wan R, Bai R, Zhan X, Shi Y. How Is Precursor Messenger RNA Spliced by the Spliceosome? *Annu Rev Biochem*. Jun 20 2020;89:333-358. doi:10.1146/annurev-biochem-013118-111024
5. Jutzi D, Akinyi MV, Mechtersheimer J, Frilander MJ, Ruepp MD. The emerging role of minor intron splicing in neurological disorders. *Cell Stress*. Feb 22 2018;2(3):40-54. doi:10.15698/cst2018.03.126
6. He H, Liyanarachchi S, Akagi K, et al. Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. *Science*. Apr 08 2011;332(6026):238-40. doi:10.1126/science.1200587
7. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res*. Apr 2008;36(7):2257-67. doi:10.1093/nar/gkn073
8. Mercer TR, Clark MB, Andersen SB, et al. Genome-wide discovery of human splicing branchpoints. *Genome Res*. Feb 2015;25(2):290-303. doi:10.1101/gr.182899.114
9. Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev*. Apr 01 2018;32(7-8):577-591. doi:10.1101/gad.312058.118
10. Briese M, Haberman N, Sibley CR, et al. A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nat Struct Mol Biol*. Oct 2019;26(10):930-940. doi:10.1038/s41594-019-0300-4
11. Leman R, Tubeuf H, Raad S, et al. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. *BMC Genomics*. Jan 28 2020;21(1):86. doi:10.1186/s12864-020-6484-5
12. Signal B, Gloss BS, Dinger ME, Mercer TR. Machine learning annotation of human branchpoints. *Bioinformatics*. Mar 15 2018;34(6):920-927. doi:10.1093/bioinformatics/btx688
13. Zhang Q, Fan X, Wang Y, Sun MA, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. *Bioinformatics*. Oct 15 2017;33(20):3166-3172. doi:10.1093/bioinformatics/btx401

14. Coolidge CJ, Seely RJ, Patton JG. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res*. Feb 15 1997;25(4):888-96. doi:10.1093/nar/25.4.888
15. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. Sep 06 2012;489(7414):57-74. doi:10.1038/nature11247
16. Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A*. Sep 2001;98(20):11193-8. doi:10.1073/pnas.201407298
17. Ke S, Shang S, Kalachikov SM, et al. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res*. Aug 2011;21(8):1360-74. doi:10.1101/gr.119628.110
18. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. May 2010;11(5):345-55. doi:10.1038/nrg2776
19. Baeza-Centurion P, Miñana B, Valcárcel J, Lehner B. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife*. Oct 28 2020;9doi:10.7554/eLife.59959
20. Lefave CV, Squatrito M, Vorlova S, et al. Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. *EMBO J*. Sep 13 2011;30(19):4084-97. doi:10.1038/emboj.2011.259
21. Wang J, Steinbacher S, Augustin M, et al. The crystal structure of a constitutively active mutant RON kinase suggests an intramolecular autophosphorylation hypothesis. *Biochemistry*. Sep 21 2010;49(37):7972-4. doi:10.1021/bi100409w
22. Collesi C, Santoro MM, Gaudino G, Comoglio PM. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Mol Cell Biol*. Oct 1996;16(10):5518-26. doi:10.1128/MCB.16.10.5518
23. Ray D, Kazan H, Cook KB, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*. Jul 2013;499(7457):172-7. doi:10.1038/nature12311
24. Goren A, Ram O, Amit M, et al. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell*. Jun 2006;22(6):769-81. doi:10.1016/j.molcel.2006.05.008
25. Zhang C, Li WH, Krainer AR, Zhang MQ. RNA landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A*. Apr 15 2008;105(15):5797-802. doi:10.1073/pnas.0801692105
26. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*. May 2008;14(5):802-13. doi:10.1261/rna.876308
27. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science*. Aug 2002;297(5583):1007-13. doi:10.1126/science.1073774
28. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*. Jul 01 2003;31(13):3568-71. doi:10.1093/nar/gkg616
29. Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*. Nov 2004;101(44):15700-5. doi:10.1073/pnas.0404901101
30. Cáceres EF, Hurst LD. The evolution, impact and properties of exonic splice enhancers. *Genome Biol*. Dec 2013;14(12):R143. doi:10.1186/gb-2013-14-12-r143

31. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. Dec 2008;40(12):1413-5. doi:10.1038/ng.259
32. Artamonova II, Gelfand MS. Comparative genomics and evolution of alternative splicing: the pessimists' science. *Chem Rev*. Aug 2007;107(8):3407-30. doi:10.1021/cr068304c
33. Weyn-Vanhentenryck SM, Feng H, Ustianenko D, et al. Precise temporal regulation of alternative splicing during neural development. *Nat Commun*. Jun 06 2018;9(1):2189. doi:10.1038/s41467-018-04559-0
34. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. Sep 11 2011;478(7367):64-9. doi:10.1038/nature10496
35. Cvačková Z, Matějů D, Staněk D. Retinitis pigmentosa mutations of SNRNP200 enhance cryptic splice-site recognition. *Hum Mutat*. Mar 2014;35(3):308-17. doi:10.1002/humu.22481
36. Soemedi R, Cygan KJ, Rhine CL, et al. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. Jun 2017;49(6):848-855. doi:10.1038/ng.3837
37. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A*. Jul 05 2011;108(27):11093-8. doi:10.1073/pnas.1101135108
38. Park JH, Gail MH, Weinberg CR, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A*. Nov 01 2011;108(44):18026-31. doi:10.1073/pnas.1114759108
39. Busslinger M, Moschonas N, Flavell RA. Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. *Cell*. Dec 1981;27(2 Pt 1):289-98. doi:10.1016/0092-8674(81)90412-8
40. Thein SL. The molecular basis of β-thalassemia. *Cold Spring Harb Perspect Med*. May 01 2013;3(5):a011700. doi:10.1101/cshperspect.a011700
41. Takeshima Y, Yagi M, Okizuka Y, et al. Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. *J Hum Genet*. Jun 2010;55(6):379-88. doi:10.1038/jhg.2010.49
42. Aznarez I, Chan EM, Zielenski J, Blencowe BJ, Tsui LC. Characterization of disease-associated mutations affecting an exonic splicing enhancer and two cryptic splice sites in exon 13 of the cystic fibrosis transmembrane conductance regulator gene. *Hum Mol Genet*. Aug 2003;12(16):2031-40. doi:10.1093/hmg/ddg215
43. Fang LJ, Simard MJ, Vidaud D, et al. A novel mutation in the neurofibromatosis type 1 (NF1) gene promotes skipping of two exons by preventing exon definition. *J Mol Biol*. Apr 13 2001;307(5):1261-70. doi:10.1006/jmbi.2001.4561
44. Ramalho AS, Beck S, Penque D, et al. Transcript analysis of the cystic fibrosis splicing mutation 1525-1G>A shows use of multiple alternative splicing sites and suggests a putative role of exonic splicing enhancers. *J Med Genet*. Jul 2003;40(7):e88. doi:10.1136/jmg.40.7.e88
45. Takahara K, Schwarze U, Imamura Y, et al. Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site

mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *Am J Hum Genet*. Sep 2002;71(3):451-65. doi:10.1086/342099
46. Habara Y, Takeshima Y, Awano H, et al. In vitro splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G-->A mutations in introns of the dystrophin gene. *J Med Genet*. Aug 2009;46(8):542-7. doi:10.1136/jmg.2008.061259
47. Starita LM, Ahituv N, Dunham MJ, et al. Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet*. Sep 07 2017;101(3):315-325. doi:10.1016/j.ajhg.2017.07.014
48. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. 10 2018;562(7726):217-222. doi:10.1038/s41586-018-0461-z
49. Miller DT, Lee K, Abul-Husn NS, et al. ACMG SF v3.1 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*. Jul 2022;24(7):1407-1414. doi:10.1016/j.gim.2022.04.006
50. Tabet D, Parikh V, Mali P, Roth FP, Claussnitzer M. Scalable Functional Assays for the Interpretation of Human Genetic Variation. *Annu Rev Genet*. Nov 30 2022;56:441-465. doi:10.1146/annurev-genet-072920-032107
51. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc*. Oct 2016;11(10):1782-7. doi:10.1038/nprot.2016.135
52. Jia X, Burugula BB, Chen V, et al. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet*. Jan 07 2021;108(1):163-175. doi:10.1016/j.ajhg.2020.12.003
53. Lue NZ, Garcia EM, Ngan KC, Lee C, Doench JG, Liau BB. Base editor scanning charts the DNMT3A activity landscape. *Nat Chem Biol*. Oct 20 2022;doi:10.1038/s41589-022-01167-4
54. Klein JC, Agarwal V, Inoue F, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods*. Nov 2020;17(11):1083-1091. doi:10.1038/s41592-020-0965-y
55. Gergics P, Smith C, Bando H, et al. High-throughput splicing assays identify missense and silent splice-disruptive POU1F1 variants underlying pituitary hormone deficiency. *Am J Hum Genet*. Aug 05 2021;108(8):1526-1539. doi:10.1016/j.ajhg.2021.06.013
56. Smith C, Burugula BB, Dunn I, Aradhya S, Kitzman JO, Lai Yee J. High-throughput splicing assays identify known and novel W*T1* exon 9 variants in nephrotic syndrome. medRxiv2023.
57. Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat Commun*. 05 2016;7:11558. doi:10.1038/ncomms11558
58. Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*. 01 2019;176(3):549-563.e23. doi:10.1016/j.cell.2018.12.010
59. Braun S, Enculescu M, Setty ST, et al. Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat Commun*. 08 2018;9(1):3315. doi:10.1038/s41467-018-05748-7

60. Ke S, Anquetil V, Zamalloa JR, et al. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res*. 01 2018;28(1):11-24. doi:10.1101/gr.219683.116
61. Cheung R, Insigne KD, Yao D, et al. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol Cell*. 01 2019;73(1):183-194.e8. doi:10.1016/j.molcel.2018.10.037
62. Adamson SI, Zhan L, Graveley BR. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol*. 06 2018;19(1):71. doi:10.1186/s13059-018-1437-x
63. Cortés-López M, Schulz L, Enculescu M, et al. High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. *Nat Commun*. Sep 22 2022;13(1):5570. doi:10.1038/s41467-022-31818-y
64. Scott A, Hernandez F, Chamberlin A, Smith C, Karam R, Kitzman JO. Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. *Genome Biol*. Dec 22 2022;23(1):266. doi:10.1186/s13059-022-02839-z
65. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. *Gastroenterology*. Jun 2010;138(6):2044-58. doi:10.1053/j.gastro.2010.01.054
66. Moreira L, Balaguer F, Lindor N, et al. Identification of Lynch syndrome among patients with colorectal cancer. *JAMA*. Oct 17 2012;308(15):1555-65. doi:10.1001/jama.2012.13088
67. Ingraham HA, Flynn SE, Voss JW, et al. The POU-specific domain of Pit-1 is essential for sequence-specific, high affinity DNA binding and DNA-dependent Pit-1-Pit-1 interactions. *Cell*. Jun 15 1990;61(6):1021-33. doi:10.1016/0092-8674(90)90067-o
68. Pfaffle R, Klammt J. Pituitary transcription factors in the aetiology of combined pituitary hormone deficiency. *Best Pract Res Clin Endocrinol Metab*. Feb 2011;25(1):43-60. doi:10.1016/j.beem.2010.10.014
69. Konzak KE, Moore DD. Functional isoforms of Pit-1 generated by alternative messenger RNA splicing. *Mol Endocrinol*. Feb 1992;6(2):241-7. doi:10.1210/mend.6.2.1569967
70. Jonsen MD, Duval DL, Gutierrez-Hartmann A. The 26-amino acid beta-motif of the Pit-1beta transcription factor is a dominant and independent repressor domain. *Mol Endocrinol*. Sep 2009;23(9):1371-84. doi:10.1210/me.2008-0137
71. Haugen BR, Wood WM, Gordon DF, Ridgway EC. A thyrotrope-specific variant of Pit-1 transactivates the thyrotropin beta promoter. *J Biol Chem*. Oct 05 1993;268(28):20818-24.
72. Mrowka C, Schedl A. Wilms' tumor suppressor gene WT1: from structure to renal pathophysiologic features. *J Am Soc Nephrol*. Nov 2000;11 Suppl 16:S106-15.
73. Klamt B, Koziell A, Poulat F, et al. Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms. *Hum Mol Genet*. Apr 1998;7(4):709-14. doi:10.1093/hmg/7.4.709
74. Kikuchi H, Takata A, Akasaka Y, et al. Do intronic mutations affecting splicing of WT1 exon 9 cause Frasier syndrome? *J Med Genet*. Jan 1998;35(1):45-8. doi:10.1136/jmg.35.1.45

75. Larsson SH, Charlieu JP, Miyagawa K, et al. Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. *Cell*. May 05 1995;81(3):391-401. doi:10.1016/0092-8674(95)90392-5
76. Barbaux S, Niaudet P, Gubler MC, et al. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. *Nat Genet*. Dec 1997;17(4):467-70. doi:10.1038/ng1297-467
77. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res*. Sep 1987;15(17):7155-74. doi:10.1093/nar/15.17.7155
78. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-94. doi:10.1089/1066527041410418
79. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*. Sep 2002;3(9):698-709. doi:10.1038/nrg890
80. Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. Jul 2014;16(7):497-503. doi:10.1038/gim.2013.176
81. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol*. 1997;4(3):311-23. doi:10.1089/cmb.1997.4.311
82. Ke S, Chasin LA. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biol*. 2010;11(8):R84. doi:10.1186/gb-2010-11-8-r84
83. Aznarez I, Barash Y, Shai O, et al. A systematic analysis of intronic sequences downstream of 5' splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome Res*. Aug 2008;18(8):1247-58. doi:10.1101/gr.073155.107
84. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*. Jun 2004;18(11):1241-50. doi:10.1101/gad.1195304
85. Voelker RB, Berglund JA. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res*. Jul 2007;17(7):1023-33. doi:10.1101/gr.6017807
86. Tian H, Kole R. Selection of novel exon recognition elements from a pool of random sequences. *Mol Cell Biol*. Nov 1995;15(11):6291-8. doi:10.1128/MCB.15.11.6291
87. Coulter LR, Landree MA, Cooper TA. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol*. Apr 1997;17(4):2143-50. doi:10.1128/MCB.17.4.2143
88. Liu HX, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev*. Jul 01 1998;12(13):1998-2012. doi:10.1101/gad.12.13.1998
89. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. *Cell*. Dec 17 2004;119(6):831-45. doi:10.1016/j.cell.2004.11.010

90. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*. Oct 2015;163(3):698-711. doi:10.1016/j.cell.2015.09.054
91. Danis D, Jacobsen JOB, Carmody LC, et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am J Hum Genet*. Nov 04 2021;108(11):2205. doi:10.1016/j.ajhg.2021.09.014
92. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. May 2009;37(9):e67. doi:10.1093/nar/gkp215
93. Cheng J, Nguyen TYD, Cygan KJ, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol*. Mar 1 2019;20(1):48. doi:10.1186/s13059-019-1653-z
94. Fairbrother WG, Yeo GW, Yeh R, et al. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res*. Jul 2004;32(Web Server issue):W187-90. doi:10.1093/nar/gkh393
95. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. May 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
96. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 01 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
97. Cheng J, Nguyen TYD, Cygan KJ, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol*. 03 2019;20(1):48. doi:10.1186/s13059-019-1653-z
98. Jagadeesh KA, Paggi JM, Ye JS, et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet*. Apr 2019;51(4):755-763. doi:10.1038/s41588-019-0348-4
99. Xiong HY, Alipanahi B, Lee LJ, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. Jan 2015;347(6218):1254806. doi:10.1126/science.1254806
100. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. 01 2019;176(3):535-548.e24. doi:10.1016/j.cell.2018.12.015
101. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol*. Apr 21 2022;23(1):103. doi:10.1186/s13059-022-02664-4
102. Wai HA, Lord J, Lyon M, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med*. Mar 2020;doi:10.1038/s41436-020-0766-9
103. Riepe TV, Khan M, Roosing S, Cremers FPM, 't Hoen PAC. Benchmarking deep learning splice prediction tools using functional splice assays. *Hum Mutat*. Jul 2021;42(7):799-810. doi:10.1002/humu.24212
104. Rowlands C, Thomas HB, Lord J, et al. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Sci Rep*. Oct 18 2021;11(1):20607. doi:10.1038/s41598-021-99747-2
105. Cormier MJ, Pedersen BS, Bayrak-Toydemir P, Quinlan AR. Combining genetic constraint with predictions of alternative splicing to prioritize deleterious splicing in rare

disease studies. *BMC Bioinformatics*. Nov 14 2022;23(1):482. doi:10.1186/s12859-022-05041-x
106. Strauch Y, Lord J, Niranjan M, Baralle D. CI-SpliceAI-Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *PLoS One*. 2022;17(6):e0269159. doi:10.1371/journal.pone.0269159
107. Ha C, Kim JW, Jang JH. Performance Evaluation of SpliceAI for the Prediction of Splicing of. *Genes (Basel)*. Aug 25 2021;12(9)doi:10.3390/genes12091308
108. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med*. Feb 2021;13(1):31. doi:10.1186/s13073-021-00835-9
109. Gordon DF, Haugen BR, Sarapura VD, Nelson AR, Wood WM, Ridgway EC. Analysis of Pit-1 in regulating mouse TSH beta promoter activity in thyrotropes. *Mol Cell Endocrinol*. Oct 1993;96(1-2):75-84. doi:10.1016/0303-7207(93)90097-4
110. Davis SW, Keisler JL, Perez-Millan MI, Schade V, Camper SA. All Hormone-Producing Cell Types of the Pituitary Intermediate and Anterior Lobes Derive From Prop1-Expressing Progenitors. *Endocrinology*. Apr 2016;157(4):1385-96. doi:10.1210/en.2015-1862
111. Li S, Crenshaw EB, 3rd, Rawson EJ, Simmons DM, Swanson LW, Rosenfeld MG. Dwarf locus mutants lacking three pituitary cell types result from mutations in the POU-domain gene pit-1. *Nature*. Oct 11 1990;347(6293):528-33. doi:10.1038/347528a0
112. Camper SA, Saunders TL, Katz RW, Reeves RH. The Pit-1 transcription factor gene is a candidate for the murine Snell dwarf mutation. *Genomics*. Nov 1990;8(3):586-90.
113. Slabaugh MB, Lieberman ME, Rutledge JJ, Gorski J. Growth hormone and prolactin synthesis in normal and homozygous Snell and Ames dwarf mice. *Endocrinology*. Oct 1981;109(4):1040-6. doi:10.1210/endo-109-4-1040
114. Lin SC, Li S, Drolet DW, Rosenfeld MG. Pituitary ontogeny of the Snell dwarf mouse reveals Pit-1-independent and Pit-1-dependent origins of the thyrotrope. *Development*. Mar 1994;120(3):515-22.
115. Wallis M. Evolution of the POU1F1 transcription factor in mammals: Rapid change of the alternatively-spliced beta-domain. *Gen Comp Endocrinol*. May 1 2018;260:100-106. doi:10.1016/j.ygcen.2018.01.005
116. Schanke JT, Conwell CM, Durning M, Fisher JM, Golos TG. Pit-1/growth hormone factor 1 splice variant expression in the rhesus monkey pituitary gland and the rhesus and human placenta. *J Clin Endocrinol Metab*. Mar 1997;82(3):800-7. doi:10.1210/jcem.82.3.3791
117. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. Sep 11 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
118. Tatsumi K, Miyai K, Notomi T, et al. Cretinism with combined hormone deficiency caused by a mutation in the PIT1 gene. *Nat Genet*. Apr 1992;1(1):56-8. doi:10.1038/ng0492-56
119. Fang Q, George AS, Brinkmeier ML, et al. Genetics of Combined Pituitary Hormone Deficiency: Roadmap into the Genome Era. *Endocr Rev*. 12 2016;37(6):636-675. doi:10.1210/er.2016-1101

120. Gergics P. Pituitary Transcription Factor Mutations Leading to Hypopituitarism. *Exp Suppl*. 2019;111:263-298. doi:10.1007/978-3-030-25905-1_13
121. Birla S, Vijayakumar P, Sehgal S, Bhatnagar S, Pallavi K, Sharma A. Characterization of a Novel POU1F1 Mutation Identified on Screening 160 Growth Hormone Deficiency Patients. *Horm Metab Res*. Apr 2019;51(4):248-255. doi:10.1055/a-0867-1026
122. Bas F, Abali ZY, Toksoy G, et al. Precocious or early puberty in patients with combined pituitary hormone deficiency due to POU1F1 gene mutation: case report and review of possible mechanisms. *Hormones (Athens)*. Dec 2018;17(4):581-588. doi:10.1007/s42000-018-0079-4
123. Blum WF, Klammt J, Amselem S, et al. Screening a large pediatric cohort with GH deficiency for mutations in genes regulating pituitary development and GH secretion: Frequencies, phenotypes and growth outcomes. *EBioMedicine*. Oct 2018;36:390-400. doi:10.1016/j.ebiom.2018.09.026
124. Bertko E, Klammt J, Dusatkova P, et al. Combined pituitary hormone deficiency due to gross deletions in the POU1F1 (PIT-1) and PROP1 genes. *J Hum Genet*. Aug 2017;62(8):755-762. doi:10.1038/jhg.2017.34
125. Birla S, Khadgawat R, Jyotsna VP, et al. Identification of Novel PROP1 and POU1F1 Mutations in Patients with Combined Pituitary Hormone Deficiency. *Horm Metab Res*. Dec 2016;48(12):822-827. doi:10.1055/s-0042-117112
126. Jadhav S, Diwaker C, Lila AR, et al. POU1F1 mutations in combined pituitary hormone deficiency: differing spectrum of mutations in a Western-Indian cohort and systematic analysis of world literature. *Pituitary*. Mar 20 2021;doi:10.1007/s11102-021-01140-9
127. Sobrier ML, Tsai YC, Perez C, et al. Functional characterization of a human POU1F1 mutation associated with isolated growth hormone deficiency: a novel etiology for IGHD. *Hum Mol Genet*. Feb 1 2016;25(3):472-83. doi:10.1093/hmg/ddv486
128. Cohen LE, Zanger K, Brue T, Wondisford FE, Radovick S. Defective retinoic acid regulation of the Pit-1 gene enhancer: a novel mechanism of combined pituitary hormone deficiency. *Mol Endocrinol*. Mar 1999;13(3):476-84. doi:10.1210/mend.13.3.0251
129. Skowronska-Krawczyk D, Ma Q, Schwartz M, et al. Required enhancer-matrin-3 network interactions for a homeodomain transcription program. *Nature*. Oct 9 2014;514(7521):257-61. doi:10.1038/nature13573
130. Guo MH, Shen Y, Walvoord EC, et al. Whole exome sequencing to identify genetic causes of short stature. *Horm Res Paediatr*. 2014;82(1):44-52. doi:10.1159/000360857
131. Nisson PE, Ally A, Watkins PC. Protocols for trapping internal and 3'-terminal exons. *PCR Methods Appl*. Aug 1994;4(1):S24-39. doi:10.1101/gr.4.1.s24
132. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*. Feb 2010;7(2):119-22. doi:10.1038/nmeth.1416
133. Zorita E, Cuscó P, Filion GJ. Starcode: sequence clustering based on all-pairs search. *Bioinformatics*. Jun 2015;31(12):1913-9. doi:10.1093/bioinformatics/btv053
134. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXivorg*. 2012;1207.3907v2

135. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. May 1 2005;21(9):1859-75. doi:10.1093/bioinformatics/bti310
136. Smith C. Massively parallel screens to identify splice disruptive variants in human disease genes [Data set]. Zenodo 2023. doi:10.5281/zenodo.7803918
137. Lerario AM, Mohan DR, Montenegro LR, et al. SELAdb: A database of exonic variants in a Brazilian population referred to a quaternary medical center in Sao Paulo. *Clinics (Sao Paulo)*. 2020;75:e1913. doi:10.6061/clinics/2020/e1913
138. Vishnopolska SA, Turjanski AG, Herrera Pinero M, et al. Genetics and genomic medicine in Argentina. *Mol Genet Genomic Med*. Jul 26 2018;doi:10.1002/mgg3.455
139. Ma SL, Vega-Warner V, Gillies C, et al. Whole Exome Sequencing Reveals Novel PHEX Splice Site Mutations in Patients with Hypophosphatemic Rickets. *PLoS One*. 2015;10(6):e0130729. doi:10.1371/journal.pone.0130729
140. Ho Y, Cooke NE, Liebhaber SA. An autoregulatory pathway establishes the definitive chromatin conformation at the pit-1 locus. *Mol Cell Biol*. May 2015;35(9):1523-32. doi:10.1128/MCB.01283-14
141. Rajas F, Delhase M, De La Hoya M, Verdood P, Castrillo JL, Hooghe-Peters EL. Nuclear factor 1 regulates the distal silencer of the human PIT1/GHF1 gene. *Biochem J*. Jul 1 1998;333 (Pt 1):77-84. doi:10.1042/bj3330077
142. DiMattia GE, Rhodes SJ, Krones A, et al. The Pit-1 gene is regulated by distinct early and late pituitary-specific enhancers. *Dev Biol*. Feb 1 1997;182(1):180-90. doi:10.1006/dbio.1996.8472
143. Theill LE, Hattori K, Lazzaro D, Castrillo JL, Karin M. Differential splicing of the GHF1 primary transcript gives rise to two functionally distinct homeodomain proteins. *EMBO J*. Jun 1992;11(6):2261-9.
144. Inoue H, Mukai T, Sakamoto Y, et al. Identification of a novel mutation in the exon 2 splice donor site of the POU1F1/PIT-1 gene in Japanese identical twins with mild combined pituitary hormone deficiency. *Clin Endocrinol (Oxf)*. Jan 2012;76(1):78-87. doi:10.1111/j.1365-2265.2011.04165.x
145. Takagi M, Kamasaki H, Yagi H, Fukuzawa R, Narumi S, Hasegawa T. A novel heterozygous intronic mutation in POU1F1 is associated with combined pituitary hormone deficiency. *Endocr J*. Feb 2017;64(2):229-234. doi:10.1507/endocrj.EJ16-0361
146. Cerbone M, Dattani MT. Progression from isolated growth hormone deficiency to combined pituitary hormone deficiency. *Growth Horm IGF Res*. Dec 2017;37:19-25. doi:10.1016/j.ghir.2017.10.005
147. Turton JP, Reynaud R, Mehta A, et al. Novel mutations within the POU1F1 gene associated with variable combined pituitary hormone deficiency. *J Clin Endocrinol Metab*. Aug 2005;90(8):4762-70. doi:10.1210/jc.2005-0570
148. Solomon BD, Mercier S, Velez JI, et al. Analysis of genotype-phenotype correlations in human holoprosencephaly. *Am J Med Genet C Semin Med Genet*. Feb 15 2010;154C(1):133-41. doi:10.1002/ajmg.c.30240
149. Domene S, Roessler E, El-Jaick KB, et al. Mutations in the human SIX3 gene in holoprosencephaly are loss of function. *Hum Mol Genet*. Dec 15 2008;17(24):3919-28. doi:10.1093/hmg/ddn294
150. Gaston-Massuet C, Andoniadou CL, Signore M, et al. Genetic interaction between the homeobox transcription factors HESX1 and SIX3 is required for normal

pituitary development. *Dev Biol*. Dec 15 2008;324(2):322-33. doi:10.1016/j.ydbio.2008.08.008
151. Holloway JM, Szeto DP, Scully KM, Glass CK, Rosenfeld MG. Pit-1 binding to specific DNA sites as a monomer or dimer determines gene-specific use of a tyrosine-dependent synergy domain. *Genes Dev*. Aug 15 1995;9(16):1992-2006. doi:10.1101/gad.9.16.1992
152. Rhodes SJ, Chen R, DiMattia GE, et al. A tissue-specific enhancer confers Pit-1-dependent morphogen inducibility and autoregulation on the pit-1 gene. *Genes Dev*. Jun 1993;7(6):913-32.
153. Cohen RN, Brue T, Naik K, Houlihan CA, Wondisford FE, Radovick S. The role of CBP/p300 interactions and Pit-1 dimerization in the pathophysiological mechanism of combined pituitary hormone deficiency. *J Clin Endocrinol Metab*. Jan 2006;91(1):239-47. doi:10.1210/jc.2005-1211
154. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*. Apr 2002;3(4):285-98. doi:10.1038/nrg775
155. König J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*. Jan 2012;13(2):77-83. doi:10.1038/nrg3141
156. Ray D, Kazan H, Chan ET, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*. Jul 2009;27(7):667-70. doi:10.1038/nbt.1550
157. Salvatori R, Hayashida CY, Aguiar-Oliveira MH, et al. Familial dwarfism due to a novel mutation of the growth hormone-releasing hormone receptor gene. *J Clin Endocrinol Metab*. Mar 1999;84(3):917-23. doi:10.1210/jcem.84.3.5599
158. Alatzoglou KS, Dattani MT. Phenotype-genotype correlations in congenital isolated growth hormone deficiency (IGHD). *Indian J Pediatr*. Jan 2012;79(1):99-106. doi:10.1007/s12098-011-0614-7
159. Shariat N, Holladay CD, Cleary RK, Phillips JA, 3rd, Patton JG. Isolated growth hormone deficiency type II caused by a point mutation that alters both splice site strength and splicing enhancer function. *Clin Genet*. Dec 2008;74(6):539-45. doi:10.1111/j.1399-0004.2008.01042.x
160. Berg MA, Guevara-Aguirre J, Rosenbloom AL, Rosenfeld RG, Francke U. Mutation creating a new splice site in the growth hormone receptor genes of 37 Ecuadorean patients with Laron syndrome. *Hum Mutat*. 1992;1(1):24-32. doi:10.1002/humu.1380010105
161. Miletta MC, Lochmatter D, Pektovic V, Mullis PE. Isolated growth hormone deficiency type 2: from gene to therapy. *Endocr Dev*. 2012;23:109-20. doi:10.1159/000341766
162. Kuijper EC, Bergsma AJ, Pijnappel W, Aartsma-Rus A. Opportunities and challenges for antisense oligonucleotide therapies. *J Inherit Metab Dis*. May 11 2020;doi:10.1002/jimd.12251
163. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. Apr 19 2017;9(386)doi:10.1126/scitranslmed.aal5209

164. Lord J, Gallone G, Short PJ, et al. Pathogenicity and selective constraint on variation near splice sites. *Genome Res*. Feb 2019;29(2):159-170. doi:10.1101/gr.238444.118
165. Khan M, Cornelis SS, Pozo-Valero MD, et al. Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genet Med*. Jul 2020;22(7):1235-1246. doi:10.1038/s41436-020-0787-4
166. Chen JM, Lin JH, Masson E, et al. The Experimentally Obtained Functional Impact Assessments of 5' Splice Site GT'GC Variants Differ Markedly from Those Predicted. *Curr Genomics*. Jan 2020;21(1):56-66. doi:10.2174/1389202921666200210141701
167. Dionnet E, Defour A, Da Silva N, et al. Splicing impact of deep exonic missense variants in CAPN3 explored systematically by minigene functional assay. *Hum Mutat*. Jul 15 2020;doi:10.1002/humu.24083
168. Bickmore WA, Oghene K, Little MH, Seawright A, van Heyningen V, Hastie ND. Modulation of DNA binding specificity by alternative splicing of the Wilms tumor wt1 gene transcript. *Science*. Jul 10 1992;257(5067):235-7. doi:10.1126/science.1321494
169. Wells J, Rivera MN, Kim WJ, Starbuck K, Haber DA. The predominant WT1 isoform (+KTS) encodes a DNA-binding protein targeting the planar cell polarity gene Scribble in renal podocytes. *Mol Cancer Res*. Jul 2010;8(7):975-85. doi:10.1158/1541-7786.MCR-10-0033
170. Potluri S, Assi SA, Chin PS, et al. Isoform-specific and signaling-dependent propagation of acute myeloid leukemia by Wilms tumor 1. *Cell Rep*. Apr 20 2021;35(3):109010. doi:10.1016/j.celrep.2021.109010
171. Lefebvre J, Clarkson M, Massa F, et al. Alternatively spliced isoforms of WT1 control podocyte-specific gene expression. *Kidney Int*. Aug 2015;88(2):321-31. doi:10.1038/ki.2015.140
172. Tsuji Y, Yamamura T, Nagano C, et al. Systematic Review of Genotype-Phenotype Correlations in Frasier Syndrome. *Kidney Int Rep*. Oct 2021;6(10):2585-2593. doi:10.1016/j.ekir.2021.07.010
173. Miyoshi Y, Santo Y, Tachikawa K, et al. Lack of puberty despite elevated estradiol in a 46,XY phenotypic female with Frasier syndrome. *Endocr J*. Jun 2006;53(3):371-6. doi:10.1507/endocrj.k05-180
174. Gast C, Pengelly RJ, Lyon M, et al. Collagen (COL4A) mutations are the most frequent mutations underlying adult focal segmental glomerulosclerosis. *Nephrol Dial Transplant*. Jun 2016;31(6):961-70. doi:10.1093/ndt/gfv325
175. Bruening W, Bardeesy N, Silverman BL, et al. Germline intronic and exonic mutations in the Wilms' tumour gene (WT1) affecting urogenital development. *Nat Genet*. May 1992;1(2):144-8. doi:10.1038/ng0592-144
176. Zorita E, Cusco P, Filion GJ. Starcode: sequence clustering based on all-pairs search. *Bioinformatics*. Jun 15 2015;31(12):1913-9. doi:10.1093/bioinformatics/btv053
177. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. Jan 01 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
178. Yang C, Romaniuk PJ. The ratio of +/-KTS splice variants of the Wilms' tumour suppressor protein WT1 mRNA is determined by an intronic enhancer. *Biochem Cell Biol*. Aug 2008;86(4):312-21. doi:10.1139/o08-075

179. Sirokha D, Gorodna O, Vitrenko Y, et al. A Novel WT1 Mutation Identified in a 46,XX Testicular/Ovotesticular DSD Patient Results in the Retention of Intron 9. *Biology (Basel)*. Nov 30 2021;10(12)doi:10.3390/biology10121248
180. Kong A, Frigge ML, Masson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. Aug 23 2012;488(7412):471-5. doi:10.1038/nature11396
181. Miyamoto Y, Taniguchi H, Hamel F, Silversides DW, Viger RS. A GATA4/WT1 cooperation regulates transcription of genes required for mammalian sex determination and differentiation. *BMC Mol Biol*. Apr 29 2008;9:44. doi:10.1186/1471-2199-9-44
182. Rossanti R, Horinouchi T, Yamamura T, et al. Evaluation of Suspected Autosomal Alport Syndrome Synonymous Variants. *Kidney360*. Mar 31 2022;3(3):497-505. doi:10.34067/KID.0005252021
183. Demmer L, Primack W, Loik V, Brown R, Therville N, McElreavey K. Frasier syndrome: a cause of focal segmental glomerulosclerosis in a 46,XX female. *J Am Soc Nephrol*. Oct 1999;10(10):2215-8. doi:10.1681/ASN.V10102215
184. Rhine CL, Cygan KJ, Soemedi R, et al. Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet*. Mar 2018;14(3):e1007231. doi:10.1371/journal.pgen.1007231
185. Truty R, Ouyang K, Rojahn S, et al. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *Am J Hum Genet*. Apr 01 2021;108(4):696-708. doi:10.1016/j.ajhg.2021.03.006
186. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. Jun 2017;136(6):665-677. doi:10.1007/s00439-017-1779-6
187. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res*. Oct 2011;21(10):1563-71. doi:10.1101/gr.118638.110
188. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res*. Dec 16 2014;42(22):13534-44. doi:10.1093/nar/gku1206
189. Savisaar R, Hurst LD. Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res*. Oct 2018;28(10):1442-1454. doi:10.1101/gr.233999.117
190. Kashima T, Manley JL. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet*. Aug 2003;34(4):460-3. doi:10.1038/ng1207
191. Hua Y, Vickers TA, Okunola HL, Bennett CF, Krainer AR. Antisense masking of an hnRNP A1/A2 intronic splicing silencer corrects SMN2 splicing in transgenic mice. *Am J Hum Genet*. Apr 2008;82(4):834-48. doi:10.1016/j.ajhg.2008.01.014
192. Korvatska O, Strand NS, Berndt JD, et al. Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS). *Hum Mol Genet*. Aug 15 2013;22(16):3259-68. doi:10.1093/hmg/ddt180
193. Landrith T, Li B, Cass AA, et al. Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. *NPJ Precis Oncol*. 2020;4:4. doi:10.1038/s41698-020-0109-y

194. Horton C, Cass A, Conner BR, et al. Mutational and splicing landscape in a cohort of 43,000 patients tested for hereditary cancer. *NPJ Genom Med*. Aug 25 2022;7(1):49. doi:10.1038/s41525-022-00323-y
195. Richards S, Aziz N, Bale S, et al. 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*. May 2015;17(5):405-24. doi:10.1038/gim.2015.30
196. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*. May 2016;17(5):257-71. doi:10.1038/nrg.2016.10
197. Sheinson DM, Wong WB, Flores C, Ogale S, Gross CP. Association Between Medicare's National Coverage Determination and Utilization of Next-Generation Sequencing. *JCO Oncol Pract*. Nov 2021;17(11):e1774-e1784. doi:10.1200/OP.20.01023
198. Leman R, Parfait B, Vidaud D, et al. SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing. *Hum Mutat*. Dec 2022;43(12):2308-2323. doi:10.1002/humu.24491
199. Tubeuf H, Charbonnier C, Soukarieh O, et al. Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. *Hum Mutat*. Oct 2020;41(10):1811-1829. doi:10.1002/humu.24091
200. Leman R, Gaildrat P, Le Gac G, et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Res*. Sep 06 2018;46(15):7913-7923. doi:10.1093/nar/gky372
201. Moles-Fernández A, Duran-Lozano L, Montalban G, et al. Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations? *Front Genet*. 2018;9:366. doi:10.3389/fgene.2018.00366
202. Rhine CL, Neil C, Wang J, et al. Massively parallel reporter assays discover de novo exonic splicing mutants in paralogs of Autism genes. *PLoS Genet*. Jan 2022;18(1):e1009884. doi:10.1371/journal.pgen.1009884
203. Mount SM, Avsec Ž, Carmel L, et al. Assessing predictions of the impact of variants on splicing in CAGI5. *Hum Mutat*. Sep 2019;40(9):1215-1224. doi:10.1002/humu.23869
204. Thompson BA, Martins A, Spurdle AB. A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. *Clin Genet*. Feb 2015;87(2):100-8. doi:10.1111/cge.12450
205. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. Apr 2022;604(7905):310-315. doi:10.1038/s41586-022-04558-8
206. Akiba K, Hasegawa Y, Katoh-Fukui Y, et al. POU1F1/Pou1f1 c.143-83A > G Variant Disrupts the Branch Site in Pre-mRNA and Leads to Dwarfism. *Endocrinology*. Dec 19 2022;164(2)doi:10.1210/endocr/bqac198
207. Hastie ND. Wilms' tumour 1 (WT1) in development, homeostasis and disease. *Development*. Aug 15 2017;144(16):2862-2872. doi:10.1242/dev.153163
208. Eswarakumar VP, Horowitz MC, Locklin R, Morriss-Kay GM, Lonai P. A gain-of-function mutation of Fgfr2c demonstrates the roles of this receptor variant in

osteogenesis. *Proc Natl Acad Sci U S A*. Aug 24 2004;101(34):12555-60. doi:10.1073/pnas.0405031101
209. Pfaff MJ, Xue K, Li L, Horowitz MC, Steinbacher DM, Eswarakumar JVP. FGFR2c-mediated ERK-MAPK activity regulates coronal suture development. *Dev Biol*. Jul 15 2016;415(2):242-250. doi:10.1016/j.ydbio.2016.03.026
210. Gong SG. Isoforms of receptors of fibroblast growth factors. *J Cell Physiol*. Dec 2014;229(12):1887-95. doi:10.1002/jcp.24649
211. Mohammadi M, Olsen SK, Ibrahimi OA. Structural basis for fibroblast growth factor receptor activation. *Cytokine Growth Factor Rev*. Apr 2005;16(2):107-37. doi:10.1016/j.cytogfr.2005.01.008
212. Roscioli T, Elakis G, Cox TC, et al. Genotype and clinical care correlations in craniosynostosis: findings from a cohort of 630 Australian and New Zealand patients. *Am J Med Genet C Semin Med Genet*. Nov 2013;163C(4):259-70. doi:10.1002/ajmg.c.31378
213. Wilkie AO. Craniosynostosis: genes and mechanisms. *Hum Mol Genet*. 1997;6(10):1647-56. doi:10.1093/hmg/6.10.1647
214. Teebi AS, Kennedy S, Chun K, Ray PN. Severe and mild phenotypes in Pfeiffer syndrome with splice acceptor mutations in exon IIIc of FGFR2. *Am J Med Genet*. Jan 01 2002;107(1):43-7. doi:10.1002/ajmg.10125
215. Reardon W, Winter RM, Rutland P, Pulleyn LJ, Jones BM, Malcolm S. Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. *Nat Genet*. Sep 1994;8(1):98-103. doi:10.1038/ng0994-98
216. Kan R, Twigg SR, Berg J, Wang L, Jin F, Wilkie AO. Expression analysis of an FGFR2 IIIc 5' splice site mutation (1084+3A->G). *J Med Genet*. Aug 2004;41(8):e108. doi:10.1136/jmg.2004.018507
217. Cornejo-Roldan LR, Roessler E, Muenke M. Analysis of the mutational spectrum of the FGFR2 gene in Pfeiffer syndrome. *Hum Genet*. May 1999;104(5):425-31. doi:10.1007/s004390050979
218. Schell U, Hehr A, Feldman GJ, et al. Mutations in FGFR1 and FGFR2 cause familial and sporadic Pfeiffer syndrome. *Hum Mol Genet*. Mar 1995;4(3):323-8. doi:10.1093/hmg/4.3.323
219. Paumard-Hernández B, Berges-Soria J, Barroso E, et al. Expanding the mutation spectrum in 182 Spanish probands with craniosynostosis: identification and characterization of novel TCF12 variants. *Eur J Hum Genet*. Jul 2015;23(7):907-14. doi:10.1038/ejhg.2014.205
220. Fernandes MB, Maximino LP, Perosa GB, Abramides DV, Passos-Bueno MR, Yacubian-Fernandes A. Apert and Crouzon syndromes-Cognitive development, brain abnormalities, and molecular aspects. *Am J Med Genet A*. Jun 2016;170(6):1532-7. doi:10.1002/ajmg.a.37640
221. Oldridge M, Zackai EH, McDonald-McGinn DM, et al. De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am J Hum Genet*. Feb 1999;64(2):446-61. doi:10.1086/302245
222. Hollway GE, Suthers GK, Haan EA, et al. Mutation detection in FGFR2 craniosynostosis syndromes. *Hum Genet*. Feb 1997;99(2):251-5. doi:10.1007/s004390050348

223. Del Gatto F, Breathnach R. A Crouzon syndrome synonymous mutation activates a 5' splice site within the IIIc exon of the FGFR2 gene. *Genomics*. Jun 10 1995;27(3):558-9. doi:10.1006/geno.1995.1095
224. Li X, Park WJ, Pyeritz RE, Jabs EW. Effect on splicing of a silent FGFR2 mutation in Crouzon syndrome. *Nat Genet*. Mar 1995;9(3):232-3. doi:10.1038/ng0395-232
225. Traynis I, Bernstein JA, Gardner P, Schrijver I. Analysis of the alternative splicing of an FGFR2 transcript due to a novel 5' splice site mutation (1084+1G>A): case report. *Cleft Palate Craniofac J*. Jan 2012;49(1):104-8. doi:10.1597/10-217
226. Lin Y, Gao H, Ai S, et al. FGFR2 mutations and associated clinical observations in two Chinese patients with Crouzon syndrome. *Mol Med Rep*. Nov 2017;16(5):5841-5846. doi:10.3892/mmr.2017.7397
227. Wang H, Xiao F, Dong X, et al. Diagnostic and clinical utility of next-generation sequencing in children born with multiple congenital anomalies in the China neonatal genomes project. *Hum Mutat*. Apr 2021;42(4):434-444. doi:10.1002/humu.24170
228. Kress W, Collmann H, Büsse M, Halliger-Keller B, Mueller CR. Clustering of FGFR2 gene mutations inpatients with Pfeiffer and Crouzon syndromes (FGFR2-associated craniosynostoses). *Cytogenet Cell Genet*. 2000;91(1-4):134-7. doi:10.1159/000056833
229. Kan SH, Elanko N, Johnson D, et al. Genomic screening of fibroblast growth-factor receptor 2 reveals a wide spectrum of mutations in patients with syndromic craniosynostosis. *Am J Hum Genet*. Feb 2002;70(2):472-86. doi:10.1086/338758
230. Bessenyei B, Nagy A, Szakszon K, et al. Clinical and genetic characteristics of craniosynostosis in Hungary. *Am J Med Genet A*. Dec 2015;167A(12):2985-91. doi:10.1002/ajmg.a.37298
231. McCann E, Kaye SB, Newman W, Norbury G, Black GC, Ellis IH. Novel phenotype of craniosynostosis and ocular anterior chamber dysgenesis with a fibroblast growth factor receptor 2 mutation. *Am J Med Genet A*. Oct 15 2005;138A(3):278-81. doi:10.1002/ajmg.a.30944
232. Apra C, Collet C, Arnaud E, Di Rocco F. FGFR2 splice site mutations in Crouzon and Pfeiffer syndromes: two novel variants. *Clin Genet*. Jun 2016;89(6):746-8. doi:10.1111/cge.12705
233. Steinberger D, Reinhartz T, Unsöld R, Müller U. FGFR2 mutation in clinically nonclassifiable autosomal dominant craniosynostosis with pronounced phenotypic variation. *Am J Med Genet*. Dec 02 1996;66(1):81-6. doi:10.1002/(SICI)1096-8628(19961202)66:1<81::AID-AJMG19>3.0.CO;2-M
234. Fenwick AL, Goos JA, Rankin J, et al. Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon syndrome. *BMC Med Genet*. Aug 31 2014;15:95. doi:10.1186/s12881-014-0095-4
235. Glidden DT, Buerer JL, Saueressig CF, Fairbrother WG. Hotspot exons are common targets of splicing perturbations. *Nat Commun*. May 12 2021;12(1):2756. doi:10.1038/s41467-021-22780-2
236. Parthasarathy S, Ruggiero SM, Gelot A, et al. A recurrent de novo splice site variant involving DNM1 exon 10a causes developmental and epileptic encephalopathy through a dominant-negative mechanism. *Am J Hum Genet*. Dec 01 2022;109(12):2253-2269. doi:10.1016/j.ajhg.2022.11.002

237. Kornblihtt AR. Promoter usage and alternative splicing. *Curr Opin Cell Biol*. Jun 2005;17(3):262-8. doi:10.1016/j.ceb.2005.04.014
238. van der Klift HM, Jansen AM, van der Steenstraten N, et al. Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. *Mol Genet Genomic Med*. Jul 2015;3(4):327-45. doi:10.1002/mgg3.145
239. Morak M, Pineda M, Martins A, et al. Splicing analyses for variants in MMR genes: best practice recommendations from the European Mismatch Repair Working Group. *Eur J Hum Genet*. Sep 2022;30(9):1051-1059. doi:10.1038/s41431-022-01106-w
240. Dawes R, Bournazos AM, Bryen SJ, et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat Genet*. Feb 2023;55(2):324-332. doi:10.1038/s41588-022-01293-8
241. Becirovic E, Böhm S, Nguyen ON, et al. In Vivo Analysis of Disease-Associated Point Mutations Unveils Profound Differences in mRNA Splicing of Peripherin-2 in Rod and Cone Photoreceptors. *PLoS Genet*. Jan 2016;12(1):e1005811. doi:10.1371/journal.pgen.1005811
242. Erwood S, Bily TMI, Lequyer J, et al. Saturation variant interpretation using CRISPR prime editing. *Nat Biotechnol*. Jun 2022;40(6):885-895. doi:10.1038/s41587-021-01201-1
243. Smith C, Kitzman JO. Benchmarking splice variant prediction algorithms using massively parallel splicing assays. *bioRxiv*. May 07 2023;doi:10.1101/2023.05.04.539398
244. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;9(8):e1003709. doi:10.1371/journal.pgen.1003709
245. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. *Nature*. May 2010;465(7294):53-9. doi:10.1038/nature09000
246. Galarza-Muñoz G, Briggs FBS, Evsyukova I, et al. Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. *Cell*. Mar 23 2017;169(1):72-84.e13. doi:10.1016/j.cell.2017.03.007
247. Taylor J, Mi X, North K, et al. Single-cell genomics reveals the genetic and molecular bases for escape from mutational epistasis in myeloid neoplasms. *Blood*. Sep 24 2020;136(13):1477-1486. doi:10.1182/blood.2020006868
248. Zhou J, Wong MS, Chen WC, Krainer AR, Kinney JB, McCandlish DM. Higher-order epistasis and phenotypic prediction. *Proc Natl Acad Sci U S A*. Sep 27 2022;119(39):e2204233119. doi:10.1073/pnas.2204233119
249. Bronner CE, Baker SM, Morrison PT, et al. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature*. Mar 17 1994;368(6468):258-61. doi:10.1038/368258a0
250. Lynch HT, Krush AJ. Cancer family "G" revisited: 1895-1970. *Cancer*. Jun 1971;27(6):1505-11. doi:10.1002/1097-0142(197106)27:6<1505::aid-cncr2820270635>3.0.co;2-l
251. Smith EA, Fuchs E. Defining the interactions between intermediate filaments and desmosomes. *J Cell Biol*. Jun 01 1998;141(5):1229-41. doi:10.1083/jcb.141.5.1229

252. Rampazzo A, Nava A, Malacrida S, et al. Mutation in human desmoplakin domain binding to plakoglobin causes a dominant form of arrhythmogenic right ventricular cardiomyopathy. *Am J Hum Genet*. Nov 2002;71(5):1200-6. doi:10.1086/344208
253. Bauce B, Basso C, Rampazzo A, et al. Clinical profile of four families with arrhythmogenic right ventricular cardiomyopathy caused by dominant desmoplakin mutations. *Eur Heart J*. Aug 2005;26(16):1666-75. doi:10.1093/eurheartj/ehi341
254. Smith ED, Lakdawala NK, Papoutsidakis N, et al. Desmoplakin Cardiomyopathy, a Fibrotic and Inflammatory Form of Cardiomyopathy Distinct From Typical Dilated or Arrhythmogenic Right Ventricular Cardiomyopathy. *Circulation*. Jun 09 2020;141(23):1872-1884. doi:10.1161/CIRCULATIONAHA.119.044934