Single-cell analysis of alternative splicing in normal and malignant stem/progenitor cells

SEAN WEN

University College Oxford (Univ)

Medical Research Council (MRC) Weatherall Institute of Molecular Medicine (WIMM)

Thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2022

Abstract

Alternative splicing is the process of utilising a different combination of exons to generate different isoforms from the same gene. Therefore, alternative splicing represents an additional and underappreciated layer of complexity underlying gene expression. To date, high-throughput single-cell RNA-sequencing (scRNA-seq) analyses have focused on characterising gene expression programme, whereas alternative splicing remains challenging to investigate. One possible reason for the sparsity of single-cell alternative splicing studies is the lack of single-cell alternative splicing analytical frameworks. To this end, we have developed analytical frameworks to comprehensively capture the alternative splicing landscape in health and disease models, and to prioritise actionable spliced genes for downstream experimental studies.

The developed frameworks consist of MARVEL, VALERIE, and IMPACT. MARVEL is an R package that provides comprehensive functionalities for the detection and quantification of alternative splicing events to enable dimension reduction analysis, differential splicing analysis, and functional annotation of differentially spliced genes. Functional annotation features include biological pathway enrichment analysis and nonsense-mediated decay prediction. VALERIE is an R package for visual-based validation of differentially spliced genes identified from MARVEL. IMPACT is an integrated in-house database consisting of a collection of pre-processed publicly available myeloid neoplasm and cancer cell line data for prioritising clinically relevant and druggable spliced genes validated by VALERIE.

We validated and demonstrated the application of our analytical frameworks on scRNA-seq data generated from both plate- (e.g., Smart-seq2) and droplet- (e.g., 10x Genomics) based library preparation methods derived from homogeneous cell lines and heterogeneous haematopoietic stem and progenitor cells in health and disease states. We believe our analytical frameworks will be advantageous to biologists to reveal novel biological insights from scRNA-seq data.

Acknowledgment

I have been mentally drafting the acknowledgment section since December 2021 when I first started to write my thesis. And now, sitting in the very same booth in my college library that I frequent during my first year as a DPhil student, I can finally put pen to paper. This section has been purposely saved for the last because I knew my project and collaboration with others would continue to evolve throughout the writing of my thesis, and so would my framing of my years leading to my enrolment into Oxford. Therefore, quite naturally, all these would shape my acknowledgment.

Traditionally, a student will thank their supervisors at the outset of acknowledgment. But I am not a man of traditions and formalities, and hence, when I begin the acknowledgment with my supervisors, I do so out of sincerity and genuineness.

I would like to thank my supervisor Adam Mead for his uncanny ability to nudge me in the right direction, and whose wits and intuition I can only hope to match one day. I would like to thank my co-supervisor Supat Thongjuea who responsibly ensure I do not stray, and who taught me never to gloss over the small details.

To Jim Hughes who has consistently provided honest feedback on my progress by chairing my thesis, transfer, and confirmation committee.

To my colleagues who I have worked with collaboratively on their research projects, and therefore have contributed to my professional growth, on top of the lessons learned from my own project. From Adam Mead's team: Alba Rodriguez Meira, Christina Simoglou Karali, Eleni Louka, Giulia Orlando, Guanlin Wang, Jennifer O'Sullivan, Lucy Field, Nikolaos Sousos, Onima Chowdhury, Richard Salisbury, and Ruggiero Norfo. From Beth Psaila's team: Abdullah Khan, Emily Thomas, and Qian Cheng. From Sten Eirik Jacobsen's team: Affaf Aliouat. From Jacqueline Boultwood's team: Andrea Pellagatti and Juseong Lee. From Douglas Higgs' team: Yuqi Shen. From Tatjana Sauka-Spengler's team: Muhammad Hanifi.

To my friends who have enriched my experience during my time at Oxford. Adib Abdullah, Alvin Hung, Amber Madden-Nadeau, Angie Carter, Anli Ouyang, Chia-Hsia Tsai, Claudia Feng, Isaac Woodhouse, Ivo Maffei, James Ashford, Kang Zhu, Kairi Masuda, Liezel Tamon, Maksim Chan, Malhar Khushu, Matthew Kemp, Nancy Young, Nooshin Vincent, Nay-Yan Oo, Patrick Pflughaupt, Rong Li, Sofia Vaz Pinton Simoes

Coelho, Swee-Swee Aung, Tarun Gupta, Thomas Yuen, Wei Wu, William Prescott, Xin Liu, Yusuf Bahasoan, and Zhijia Zhang.

To those who encouraged and supported me during my search of a PhD position during my 2017-2018 gap year. Your vote of confidence meant the world to me: Chee-Onn Leong, Chooi-Ling Lim, Daniel Jonathan Park, Fleur Hammet, Jingmei Li, Joanna Lim, Melissa Southey, and Tu Nguyen-Dumont.

Last but absolutely not least, Peter Pook. Every contribution I make to science hereafter is me paying it forward from you.

This thesis is done, but the work has just begun. In the words of Gil Grissom, I will “follow the evidence.”

July 2022

Contents

1 Introduction

1.1 To splice or not to splice?

To understand the dysregulation of the splicing machinery in diseases, such as cancer, we must first have an understanding of the splicing mechanism and the different proteins and non-coding RNAs involved. This is because genetic alteration or dysregulated expression of these splicing factors can contribute to disease initiation and progression (Grosso, Martins, & Carmo-Fonseca, 2008).

1.1.1 The splicing process

Pre-mRNA molecules are generated by DNA transcription. Splicing is the process of removing the non-coding intronic sequence of the pre-mRNA molecules and subsequently ligating the adjoining exons to yield the mature mRNA molecules for downstream protein translation (Bonnal, Lopez-Oreja, & Valcarcel, 2020; Z. Liu & Rabadan, 2021; E. Wang & Aifantis, 2020).

The splicing process primarily involves the interaction between the major spliceosome and several *cis*-acting RNA sequences (Bonnal et al., 2020; Z. Liu & Rabadan, 2021; E. Wang & Aifantis, 2020). The major spliceosome consists of five small nuclear ribonucleoproteins (snRNPs), namely U1, U2, U4, U5, and U6 snRNPs, several small nuclear RNAs (snRNAs) and many other interacting proteins. The *cis*-acting RNA sequences are namely the 5' splice site, branchpoint, polypyrimidine tract, and 3' splice site (Figure 1.1).

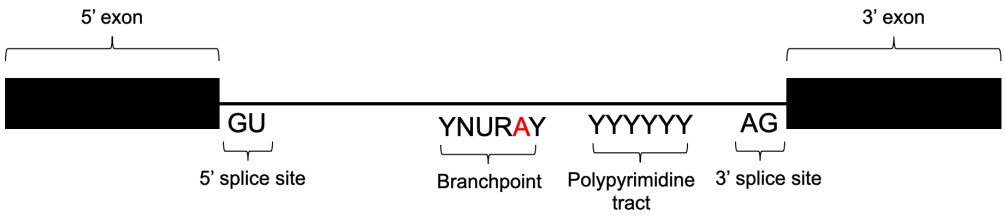


Figure 1.1: *cis*-acting RNA sequences involved in the splicing process. The GU and AG dinucleotides represent the consensus (conserved) splice site sequences at the 5' and 3' exon-intron junctions, respectively. The polypyrimidine tract is usually 15-20 base pairs long and is located ~5-40 base pairs before the 3' splice site. The most important nucleotide within the branchpoint is the adenine base (red) because this

nucleotide interacts with the 5’ splice site to form the intron lariat during the splicing process.

First, U1 snRNP binds to the 5’ splice site of the intron through base-pairing between U1 snRNA and 5’ splice site (Bonnal et al., 2020). Next, SF1, U2AF2, and U2AF1 bind to the branchpoint, polypyrimidine tract, and 3’ splice site sequences, respectively, after which, the U2 snRNP is recruited and bound to the branchpoint through base-pairing between U2 snRNA and branchpoint sequence. SF1 is then replaced by SF3B1, a component of U2 snRNP, at the branchpoint. U4/U6-U5 tri-snRNP complex is then recruited, and this leads to the formation of the catalytically active conformation of the major spliceosome. This catalytically active spliceosome will initiate the splicing out of introns.

The splicing out of introns is an intricate process that can be summarised into a two-step transesterification process (E. Wang & Aifantis, 2020) (Figure 1.2). The first transesterification reaction involves the nucleophilic attack on the 5’ splice site by a 2’-hydroxyl group on the branchpoint sequence. This leads to a cleaved 5’ exon and a lariat structure containing the intron and 3’ exon. The second transesterification reaction involves the attack on the 3’ splice site by a 2’-hydroxyl group on the detached 5’ exon. This leads to removal of intron removal and ligation of 5’ and 3’ exons to yield the mature mRNA molecule. It is noteworthy that the major spliceosome catalyses the removal of the U2-type introns, which constitute majority of all introns, whereas the minor spliceosome catalyses the removal of the U12-type introns, which constitute minority of all introns (Madan et al., 2015). U2-type introns consists of GU and AG dinucleotide sequence at the 5’ and 3’ splice site, respectively, whereas U12-type introns consists of dinucleotide sequence other than GU and AG at the 5’ and 3’ splice site, respectively.

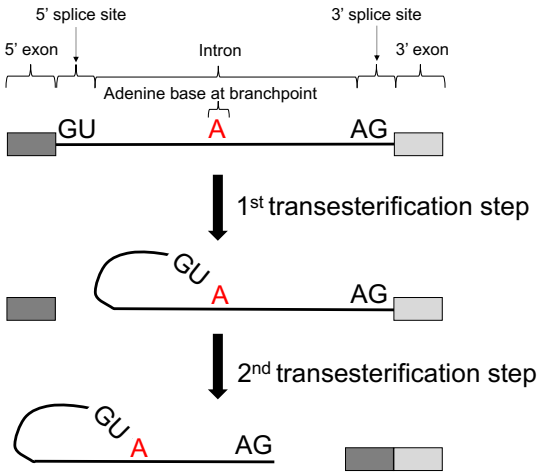


Figure 1.2: Splicing out of intron in a two-step transesterification process. For simplicity, the components of the major spliceosome are not shown. Figure adapted from (Bonnal et al., 2020).

1.1.2 The alternative splicing process

Not all exons are constitutively spliced in. Similarly, not all introns are constitutively spliced out. The process of generating different combination of exons and introns that make up the mature mRNA molecules from the same gene is called alternative splicing (Bonnal et al., 2020; E. Wang & Aifantis, 2020). Alternative splicing enables a diversity of mRNA molecules, and consequently, enables a diversity of proteins to be generated from the same gene which ultimately contribute to shaping the cellular phenotype.

Alternative splicing is coordinated by *cis*-acting RNA sequences and the recruitment of RNA-binding proteins (RBPs) to these sequences (Bonnal et al., 2020; E. Wang & Aifantis, 2020). *cis*-acting RNA sequences are namely exonic splice enhances (ESEs) and silencers (ESSs) and intronic splice enhances (ISEs) and silencers (ISSs). RBPs can be categorised into serine-arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). SR proteins are characterised by containing RNA recognition motif (RRM) and SR dipeptide-rich domains. On the other hand, hnRNPs are structurally more diverse and contain different types of RNA-binding domains and also unstructured domains. It is these domains that mediate protein-protein and protein-RNA interactions.

SR proteins preferentially bind to exonic and intronic splicing enhancers (ESEs and ISEs) whereas hnRNPs preferentially bind to exonic and intronic splicing silencers (ESSs and ISSs) (E. Wang & Aifantis, 2020). In general, SR proteins promote exon inclusion whereas hnRNPs antagonise the SR proteins to promote exon exclusion. The interplay between these RBPs and *cis*-acting RNA sequences yields seven main types of alternative splicing events (Brooks et al., 2014) (Figure 1.3).

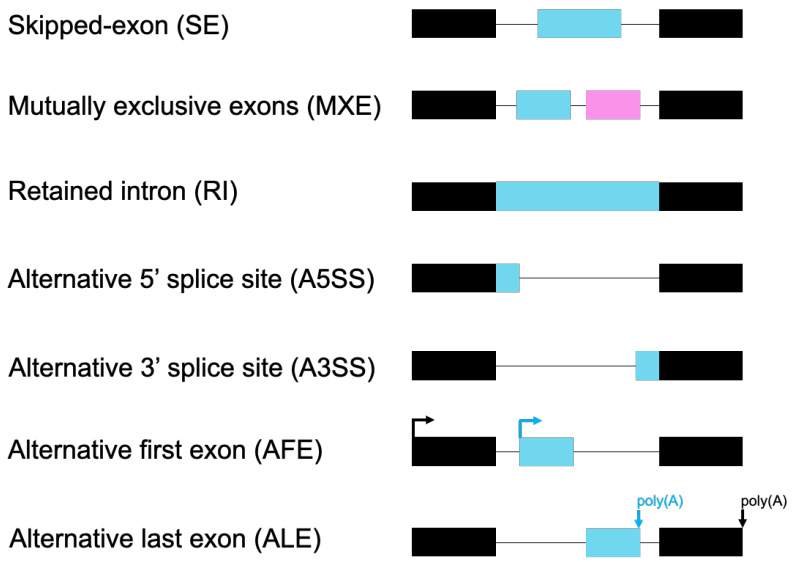


Figure 1.3: Main alternative splicing event types. Different combination of exon and intron inclusion generates different types of alternative splicing events. Constitutive exons (black) are always spliced-in whereas alternative exons (blue and pink) may be spliced-in or spliced-out.

Therefore, alternative splicing increases protein diversity and cellular functionalities by generating multiple isoforms from the same gene. According to the GENCODE v31 database (Harrow et al., 2012), there are 19,975 protein-coding genes, of which 17,228 (86.2%) genes with >1 isoform catalogued (Figure 1.4).

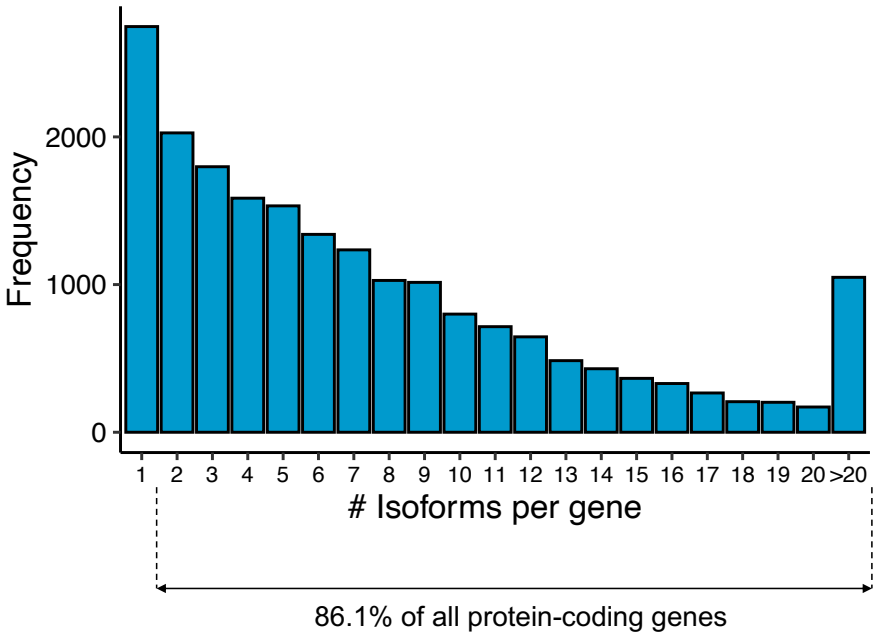


Figure 1.4: Number of isoforms per protein-coding gene. Cumulatively, most genes has >1 isoform catalogued.

1.1.3 Alternative splicing in normal haematopoiesis

In the embryonic stage, haematopoiesis first arise in the yolk sac (A Victor Hoffbrand, 2016). Specifically, both haemopoietic and endothelial cells are simultaneously generated from a common mesodermal precursor cell (haemangioblast). Thereafter, haematopoietic progenitors are generated in both the yolk sac and aorta-gonad-mesonephros (AGM) region. Later, the first haematopoietic stem cells (HSCs) arise from the AGM region and migrate to the foetal liver. Finally, from the foetal liver, HSCs migrate and colonise the bone marrow, where they reside throughout adult stages of life. An overview of the cell populations in which alternative splicing has been characterised is shown in Figure 1.5.

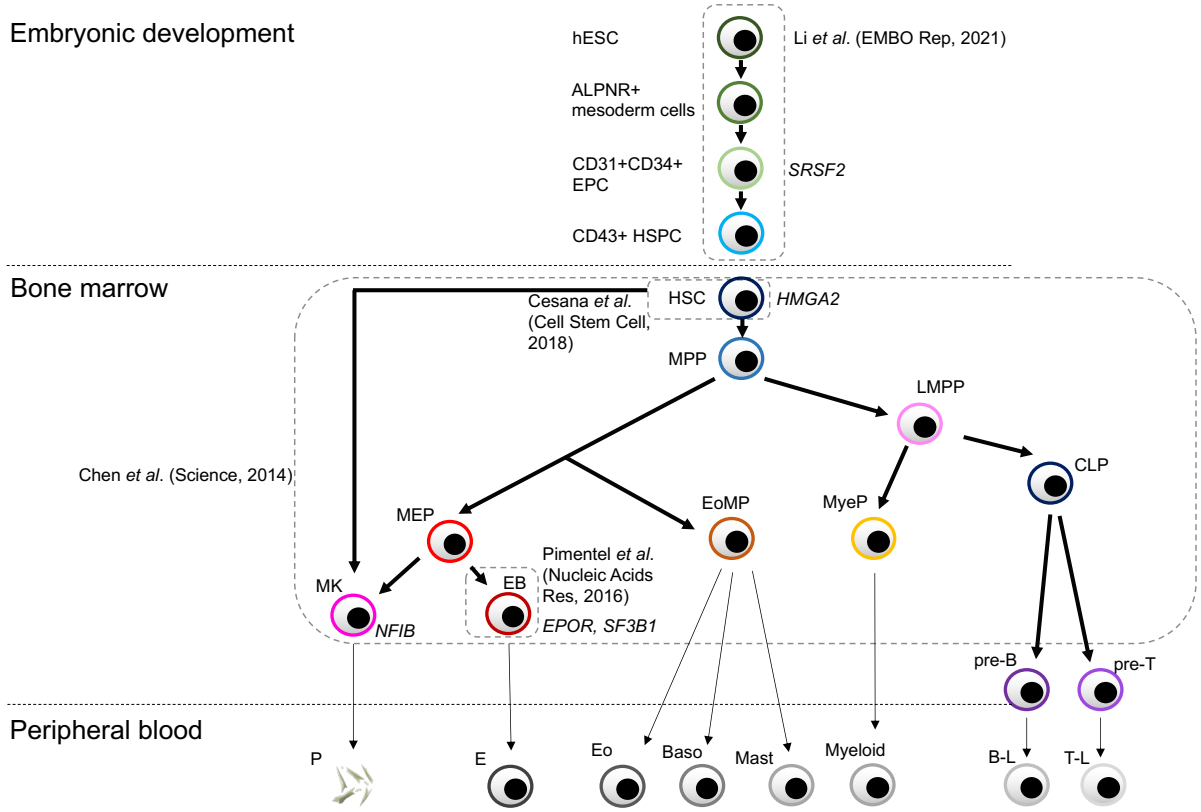


Figure 1.5: Overview of reported studies that investigated alternative splicing and their respective cell population(s) included for analysis. The reported studies are indicated with "*et al*." and the corresponding cell population(s) analysed are indicated within the dashed box. The candidate spliced genes brought forward from high-throughput RNA-seq for validation and functional characterisation indicated in italics and next to the relevant cell population(s). Baso: Basophil; B-L: B lymphocyte; CLP: Common lymphoid progenitor; CMP: Common myeloid progenitor; E: Erythrocyte; EB: Erythroblast; Eo: Eosinophil; EoMP: Eosinophil-basophil-mast cell progenitor; hESC: Human embryonic stem cell; HSC: Haematopoietic stem cell; HSPC: Haematopoietic stem and progenitor cell; LMPP: Lymphoid-primed multipotent progenitor; MK: Megakaryocyte; MPP: Multi-potent progenitor; MyeP: Myeloid progenitor; P: Platelet; T-L: T lymphocyte. Haematopoiesis hierarchy model of bone marrow and peripheral blood adapted from (Psaila & Mead, 2019).

Alternative splicing dynamics during embryonic development were characterised by differentiating human embryonic stem cells (hESCs) into haematopoietic cells, and RNA-seq was performed on cell populations from each differentiation stage (Y. Li, Wang, et al., 2021). The cell populations analysed, in

ascending differentiation stage, were hESCs, ALPNR$^+$ mesoderm cells, CD31$^+$CD34$^+$ endothelial progenitor cells (EPCs), and CD43+ haematopoietic stem and progenitor cells (HSPCs). Expression levels of splicing factors were observed to be relatively constant from hESCs to ALPNR$^+$ mesoderm cells and from EPCs to HSPCs. However, expression levels of splicing factors were significantly reduced from ALPNR$^+$ mesoderm cells to EPCs. One such splicing factor was *SRSF2*. This reduction of *SRSF2* expression level was found to be associated with the exclusion of *NUMB* exon 9. Indeed, two SRSF2 binding motifs were found on this alternative exon. The exclusion of *NUMB* exon 9 regulated NOTCH signalling, which in turn, regulated EPC generation. Therefore, this study demonstrated changes in expression levels of splicing factors as a mechanism for regulating alternative splicing of key genes involved in haematopoietic cell differentiation.

Alternative splicing dynamics in the later stages of the ontogeny of haematopoietic system, namely foetal liver and bone marrow, have been characterised for several early HSPC populations, namely, multipotent progenitor cells (MPP), common lymphoid progenitors (CLPs), common myeloid progenitors (CMPs), and megakaryocyte-erythrocyte progenitors (MEP). These HSPC populations were retrieved using antibodies complementary to cell lineage-specific surface markers coupled with fluorescence-activated cell sorting (FACS) (L. Chen et al., 2014). More mature HSPC populations, such as erythroblasts (EBs) and megakaryocytes (MKs), and terminally differentiated cell populations, such as granulocytes, were obtained by treating progenitor cells with specific growth factors to obtain more terminally differentiated cell population for alternative splicing analysis.

In general, both protein-coding and non-protein coding genes were identified to be differentially spliced across the HSPC compartment, even in the absence of detectable gene expression changes (L. Chen et al., 2014). Notably, around half of differentially spliced isoforms involved at least one protein domain. One such gene that was shortlisted for functional characterisation was *NFIB*. The short isoform of *NFIB* (*NFIB*-S) lacked the DNA binding/dimerization domain and was observed to be enriched in MKs. Knockdown of *NFIB*-S using shRNA demonstrated *NFIB*-S was essential for MK differentiation and maturation. Therefore, this study demonstrated the inclusion or exclusion of protein domains through alternative splicing as a mechanism in key genes involved in haematopoietic cell maturation.

Comparative RNA-seq analysis focusing on only HSCs derived from foetal liver, cord blood, and blood marrow identified several differentially expressed isoforms without concomitant change in gene expression levels across these distinct developmental stages (Cesana et al., 2018). One such gene was *HMGA2*, which demonstrated alternative splicing of the 3'-untranslated region (3'-UTR). The shorter *HMGA2* isoform (*HMGA2*-S) was more highly expressed in CB HSCs compared to FL HSCs. *HMGA2*-S, without the 3'-UTR otherwise present in *HMGA2*-L, escaped miRNA-mediated inhibition, and reinforced HSC-specific program in CB HSCs. Therefore, this study demonstrated miRNA targeting of alternatively spliced 3'-UTR as a mechanism in regulating key genes involved in maintaining developmental stage-specific HSC identity.

In more terminally differentiated cell populations, such as erythroblasts and granulocytes, intron retention was identified as a means of regulating gene expression (Pimentel et al., 2016; Wong et al., 2013). For example, the proportion of genes with intron retention is increased towards more terminally differentiated erythroblast, i.e., from proerythroblasts (proE) to early basophilic erythroblasts (e-basoE), late basophilic erythroblasts (l-basoE), polychromatophilic erythroblasts (polyE), and finally orthochromatophilic erythroblasts (orthoE). Intron retention leads to the inclusion of premature stop codon, which in turn, subjects the isoform to nonsense-mediated decay (NMD) within the nuclear compartment. Example of genes subjected to intron retention in late-stage erythroblast included the erythropoietin receptor *EPOR* and splicing factor *SF3B1*.

Taken together, alternative splicing studies in normal (physiological) haematopoiesis demonstrated alternative splicing-mediated gene regulation via differential expression of splicing factors (e.g., *SRSF2*), splicing out of exons encoding for domains essential for protein-protein interactions (e.g., *NFIB*), splicing out of 3'-UTR otherwise targeted by miRNA (e.g., *HMGA2*), and intron retention leading to NMD (e.g., *EPOR* and *SF3B1*). Therefore, annotation of alternative splicing events with protein domains, *in silico* identification of miRNAs complementary to 3'-UTRs, and *in silico* prediction of intron retention-mediated NMD would be helpful to understand the functional consequence of differentially spliced genes, and hence identify candidate genes for downstream functional characterisation.

1.1.4 Alternative splicing in haematopoietic malignancies

Classical gene expression analysis may underestimate the burden of functional gene disruption in haematological cancers. Indeed, dysregulation of the splicing machinery in haematological cancers have been intensively studied in the last decade (Grosso et al., 2008). This coincided with the advent of high-throughput RNA-seq technologies that enabled transcriptome-wide analysis and the discovery of novel and disease-specific isoforms (Bonnal et al., 2020). There are several mechanisms by which the splicing machinery may be disrupted and consequently contribute to the pathogenesis of haematological cancers: (1) *trans*-acting genetic variants in genes encoding for splicing factors, (2) *cis*-acting genetic variants at splice site and branchpoint sequences, and (3) dysregulation of RNA-binding protein (RBP) expression (Figure 1.6).

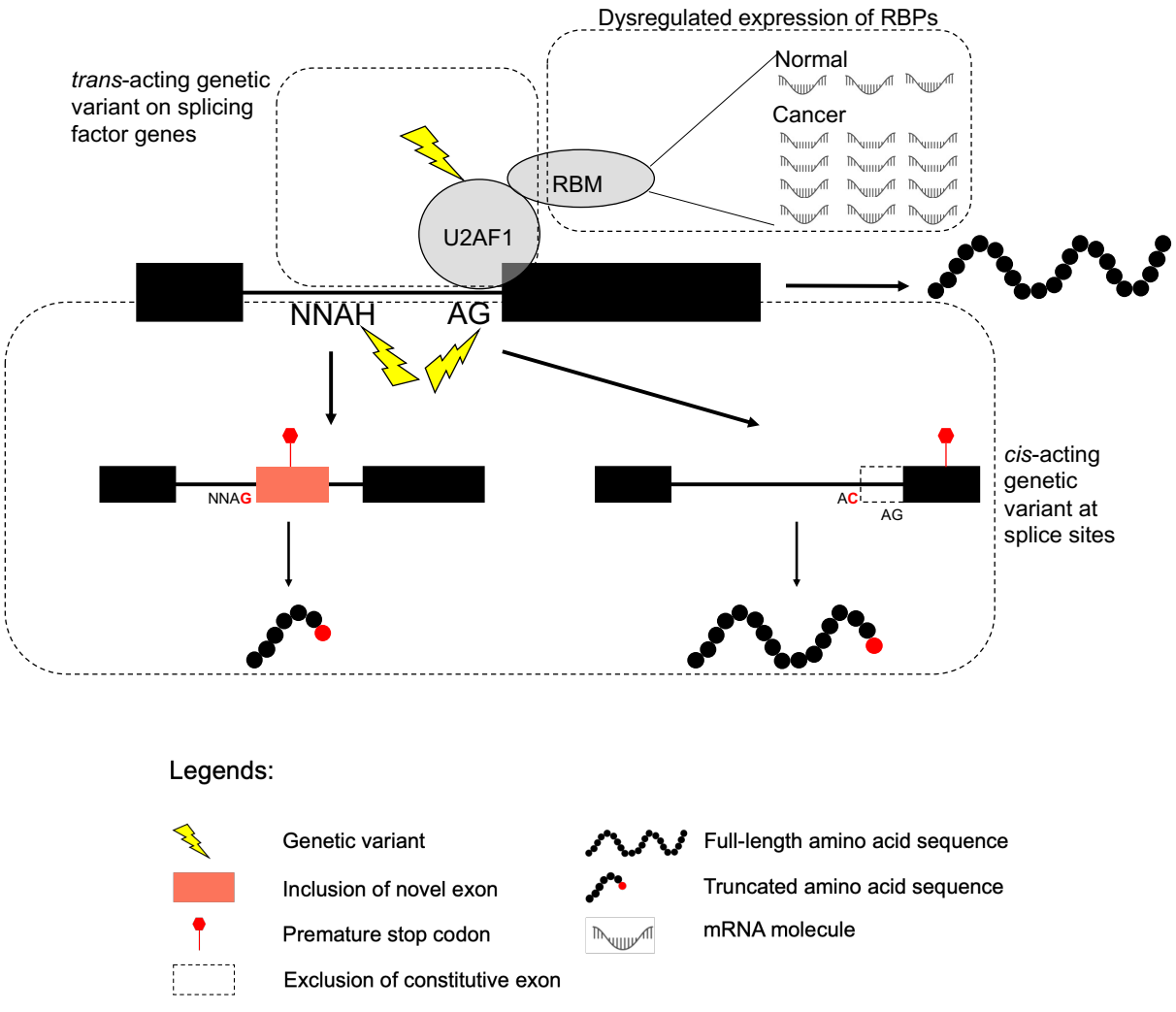


Figure 1.6: Mechanisms by which disruptions in the splicing machinery can lead to aberrant splicing. (Top left) *trans*-acting genetic variants on splicing factor genes leading to aberrant splicing of distant target genes. **(Bottom)** *cis*-acting genetic variants that create *in situ* novel splice sites **(bottom left)** or abrogate canonical splice site **(bottom right)**. **(Top right)** Dysregulated expression of RBPs. AG: Canonical acceptor splice site sequence. H: A/C/T nucleotide.

Whole genome- and exome- sequencing revealed genetic variants in genes encoding for components of the splicing machinery (Graubert et al., 2011; Quesada et al., 2011; Yoshida et al., 2011). In some haematological cancers such as myelodysplastic syndrome (MDS) and myeloproliferative neoplasm (MPN), genetic variants in splicing factors were present in up to half or more of the entire patient cohort (Pellagatti et al., 2018; Schischlik et al., 2019; Shiozawa et al., 2018). The most well investigated splicing factors to date are *SF3B1*, *SFSR2*, *U2AF1*, and *ZRSR2*. This is because genetic variants were most commonly found on these splicing factors.

Genetic variants in these splicing factors lead to transcriptome-wide aberrant alternative splicing pattern characteristic of each splicing factor. For example, hotspot variants in *SF3B1* lead to increased alternative 3' splice site (A3SS) usage and decreased intron retention, hotspot variants in *SRSF*2 lead to skipping or inclusion of alternative exons, hotspot variants in *U2AF1* lead to both A3SS usage and skipping or inclusion of alternative exons, whereas truncating variants in *ZRSR2* lead to increased retention of U12-type introns (Inoue et al., 2021; Shiozawa et al., 2018).

Aside from alternative splicing patterns, the genomic sequence of alternatively spliced exons and their flanking sequences may also be indicative of the specific type of splicing factor. For example, *SF3B1*^{MUT}-associated A3SS tended to be located ~20-24 base-pairs (bp) away from the canonical 3' splice site, *SRSF2*^{MUT}-associated inclusion and skipping of alternative exons tended to occur on exons enriched with CCNG and GGNG motif, respectively, whereas *U2AF1*^{MUT}-associated inclusion and skipping of alternative exons tended to occur on exons with acceptor splice site sequence [C/A]AG and [C/T]AG, respectively (Ilagan et al., 2015; Shiozawa et al., 2018).

Genetic variants in splicing factors may contribute to disease phenotype broadly through two mechanisms. Firstly, the dysregulated alternative splicing may lead to disruption of the translated amino acid sequences. For example, insertion of

A3SS into *ABCB7* gene in *SF3B1*^{MUT} patients creates a premature stop codon (PTC) in the original amino acid sequence, that in turn, leads to nonsense-mediated decay (NMD) and finally down-regulation of the corresponding gene (Shiozawa et al., 2018). Secondly, dysregulated alternative splicing may lead to a change in the translated amino acid sequence but without adversely affecting the protein function. For example, insertion of a novel exon into *IRAK4* in *U2AF1*^{MUT} patients leads to the introduction of a novel protein domain that increases IRAK4 interaction with MyD88 and consequently increases activation of the NF-kB signalling pathway (Smith et al., 2019). Taken together, the former mechanism may lead to inactivation of tumour suppressors or essential genes whereas the latter mechanism may lead to activation of oncogenes, which ultimately contribute to disease development and phenotype.

Table 1.1 summarises the aberrantly spliced genes associated with genetic variants in splicing factors *SF3B1*, *SRSF2*, *U2AF1*, and *ZRSF2*. This table may serve as a useful resource for validation of novel splicing analysis frameworks.

Table 1.1: Aberrantly spliced genes identified in *SF3B1*^{MUT}, *SRSF2*^{MUT}, *U2AF1*^{MUT}, and *ZRSR2*^{MUT} patients. Genes listed here were identified by high-throughput RNA-seq and validated by PCR either in the original sample in which the splicing event was detected or in cell lines transformed or transfected with the corresponding mutated splicing factor.

Genetic variant	RNA-seq			Spliced gene validated (PCR)					Reference
	Cancer type	Sample type	Sample size mutant/ total (%)	Gene	Strand	Direction	Event type	Event coordinate	
SF3B1									
K700E	MDS	BM-MNC	63/125 (50%)	*ERFE*	+	Inclusion	A3SS	chr2:238162731|238162724	(Bondu et al., 2019)
K700E (majority), H662Q	MDS	BM-MNC	3/6 (50%)	*SLC25A37*	+	Inclusion	RI	NOS	(Visconte et al., 2015)
K700E (majority), E662D, K666N, R625L	MDS	BM-MNC (CD34+)	28/84 (33%)	*SEPTIN2*	+	Inclusion	A3SS	chr2: 241335975|241335959	(Pellagatti et al., 2018)
K700E	MDS	BM-MNC	9/11 (82%)	*ENOSF1*	-	Inclusion	A3SS	chr18:683380|683395	(Bergot et al., 2020)
				SF3B1	-	Inclusion	A3SS	chr2:197403035|197403059	
				TMEM14C	+	Inclusion	A3SS	chr6:10724570|10724556	
K700E	MDS	BM-MNC (CD34+)	8/12 (67%)	*ABCB7*	-	Inclusion	A3SS	chrX:75071683|75071704	(Dolatshad et al., 2015; Dolatshad et al., 2016)
				DYNLL1	+	Inclusion	A3SS	chr12:120496410|120496402	
				ENOSF1	-	Inclusion	A3SS	chr18:683380|683395	
				HINT2	-	Inclusion	A3SS	chr9:35813145|35813156	
				SEPTIN6	-	Inclusion	A3SS	chrX:119625379|119625396	
				TMEM14C	+	Inclusion	A3SS	chr6:10724570|10724556	

K700E	MDS	BM-MNC	3/12 (25%)	*GCC2*	+	Inclusion	A3SS	chr2:108486511|108486499	(J. Zhang et al., 2019)
				KANSL3	-	Inclusion	A3SS	chr2:96619763|96619776	
				MAP3K7	-	Inclusion	A3SS	chr6:90560214|90560234	
				ORAI2	+	Inclusion	A3SS	chr7:102436221|102436202	
				PPP2R5A	+	Inclusion	A3SS	chr1:212345803|212345790	
				TTI1	-	Inclusion	A3SS	chr20:38002776|38002793	
				ZNF91	-	Inclusion	A3SS	chr7:23362730|23362739	
K700E (majority), E622D, E622D, R625C/L, N626D, H662Q, K666N/Q/R/T, p.698–700del, p.700–701del, G740E, D781G	MDS	BM-MNC, CD34+	68/214 (32%)	*RFNG*	-	Exclusion	RI	NOS	(Shiozawa et al., 2018)
				RECQL4	-	Exclusion	RI	NOS	
				NDOR1	+	Exclusion	RI	NOS	
				AP1G2	-	Exclusion	RI	NOS	
				PIEZO1	-	Exclusion	RI	NOS	
				ABCB7	-	Inclusion	A3SS	chrX:75071683|75071704	
				PPOX	+	Inclusion	A3SS	NOS	
K700E	CLL	BM-MNC	6/12 (50%)	*CDK8*	+	Inclusion	A3SS	chr13:26397153|26397139	(Z. Liu et al., 2020)
				CEP135	+	Inclusion	A3SS	chr4:56009735|56009713	
				CHTF18	+	Inclusion	A3SS	chr16:794054|794034	
				ELP2	+	Inclusion	A3SS	chr18:36145948|36145934	
				KANSL3	-	Inclusion	A3SS	chr2:96619762|96619776	
				MAP3K7	-	Inclusion	A3SS	chr6:90560214|90560234	
				MICAL1	-	Inclusion	A3SS	chr6:109445862|109445875	
				PPP2R5A	+	Inclusion	A3SS	chr1:212345803|212345790	
				RNF2	+	Inclusion	A3SS	chr1:185091579|185091565	
				SMURF2	-	Inclusion	A3SS	chr17:64578576|64578594	

				TTI1	-	Inclusion	A3SS	chr20:38002776|38002793	
K700E	MDS	BM-MNC/PB	8/12 (67%)	*MAP3K7*	-	Inclusion	A3SS	chr6:90560214|90560234	(Lee et al., 2018)
	CLL		7/13 (54%)						
	AML		5/9 (56%)						
	CMML		3/10 (23%)						
SRSF2									
P95H (majority), P95L, P96_R102del	MDS	BM-MNC (CD34+)	8/84 (9.5%)	*AKAP8*	-	Inclusion	SE	chr19:15369067:15369224	(Pellagatti et al., 2018)
P95H (majority), P95T, P95L	MDS	BM-MNC, CD34	39/214 (18%)	*EZH2*	-	Inclusion	SE	chr7:148818978:148819059	(Shiozawa et al., 2018)
				EZH2	-	Exclusion	SE	chr7:148817222:148817391	
P95H	MDS	BM-MNC/PB	NOS	*CASP8*	+	Exclusion	SE	chr2:201272898:201272942	(Lee et al., 2018)
	CLL	BM-MNC/PB	NOS	*CASP8*	+	Exclusion	SE	chr2:201724889:201274953	
P95H, P95R, P95L	AML	BM-MNC	6/115 (5.2%)	*EZH2*	-	Inclusion	SE	chr7:148818978:148819059	(Cancer Genome Atlas Research et al., 2013; Rahman, Lin, Bradley, Abdel-Wahab, & Krainer, 2020)
P96H	AML	BM-MNC	69/1119 (6.1%)	*IDH3G*	-	Inclusion	SE	chrX:153786801:153786947	(Bamopoulos et al., 2020)
				IDH3G	-	Exclusion	SE	chrX:153790564:153790575	
P95H, P95R, P95L	NA	cell line (HEL)	NA	*HNRNPA2B1*	-	Exclusion	SE	chr7:26193575:26193694	(Liang et al., 2018)
P95H, R86_G93dup	NA	Cell line (K572)	NA	*PRMT2*	+	Inclusion	SE	chr21:46636439:46636547	(Pangallo et al., 2020)
				FAXDC2	-	Exclusion	SE	chr5:154834625:154834728	

P95L	NA	CRISPR/ Cas9-edited normal iPSC	NA	*GNAS*	+	Inclusion	SE	chr20:58898941:58898985	(Wheeler et al., 2022)
U2AF1									
S34, Q157	MDS	BM-MNC, CD34+	14/214 (6.5%)	*EZH2*	-	Inclusion	SE	chr7:148818978:148819059	(Shiozawa et al., 2018)
S34F (majority), S34Y	MDS	BM-MNC, CD34+	13/150 (8.7%)	*DEK*	-	Exclusion	SE	chr6:18258304:18258405	(Graubert et al., 2011; Okeyo-Owuor et al., 2015)
				IFI44	+	Exclusion	A3SS	NOS	
				WASHC4	+	Exclusion	A3SS	NOS	
				SERPINB8	+	Exclusion	A3SS	chr18:63986963|63986874	
				SMN1	+	Exclusion	SE	NOS	
S34F (majority), S34Y, Q157P	AML	BM-MNC	7/169 (4.1%)	*ATR*	-	Inclusion	SE	chr3:142450466:142450603	(Cancer Genome Atlas Research et al., 2013; Ilagan et al., 2015)
S34F	AML	BM-MNC	4/200 (2%)	*CHCHD7*	-	Exclusion	A3SS	chr8:56216389|56216433	(Brooks et al., 2014; Cancer Genome Atlas Research et al., 2013)
				CTNNB1	+	Inclusion	A3SS	chr3:41239819|41239660	
S34F (majority), S34Y, Q157P	AML	BM-MNC	7/110 (6.3%)	*BCOR*	-	Exclusion	A3SS	chrX:40063850|40063952	(Cancer Genome Atlas Research et al., 2013; Shirai et al., 2015)
				GNAS	+	Exclusion	SE	NOS	
				H2AFY	-	Inclusion	MXE	chr5:135352946:135353045:-@chr5:135350823:135350913	
				KDM6A	+	Exclusion	SE	chrX:45060022:45060156	
				KMT2D	-	Inclusion	SE	NOS	
				MED24	-	Inclusion	SE	chr17:40034916:40034972	
				PICALM	-	Inclusion	A3SS	chr11:85981988|85982003	
S34F	MDS	BM-MNC	11/25 (44%)	*ATG7*	+	Inclusion	ALE	NOS	(Cancer Genome Atlas Research et al., 2013; Park et al., 2016)
	AML	BM-MNC	4/169 (2.3%)						

S34F (majority), S34Y, Q157P	MDS	BM-MNC (CD34+)	6/23 (26%)	*IRAK4*	+	Inclusion	SE	chr12:43771220:43771365	(Cancer Genome Atlas Research et al., 2013; Smith et al., 2019)
S34F, R156H, Q157P, Q157R	AML	BM-MNC	6/160 (3.4%)						
S34F, I24T	NA	Cell line (K572)	NA	*RHBDD2*	+	Inclusion	SE	chr7:75881365:75881486	(Pangallo et al., 2020)
				H2AFY	-	Inclusion	MXE	chr5:135352946:135353045:-@chr5:135350823-135350913	
S34F	NA	Ery, gran derived from CD34+ cells	NA	*H2AFY*	-	Inclusion	MXE	chr5:135352946:135353045:-@chr5:135350823-135350913	(Yip et al., 2017b)
				STRAP	+	Exclusion	SE	chr12:15883541:15883676	
				SMARCA5	+	Exclusion	SE	chr14:143540363:143540495	
				ITGB3BP	-	Inclusion	SE	chr1:63510065:63510181	
				ATR	-	Inclusion	SE	chr3:142450466:142450603	
S34F	NA	CRISPR/Cas9-edited normal iPSC	NA	*GNAS*	+	Inclusion	SE	chr20:58898941:58898985	(Wheeler et al., 2022)
ZRSR2									
Various truncating and missense variants	MDS	BM-MNC	8/12 (67%)	*FRA10AC1*	-	Inclusion	RI	chr10:93694938:93698135	(Madan et al., 2015)
				WDR41	-	Inclusion	RI	chr5:77459125:77463094	
Various truncating and missense variants	MDS	BM-MNC	8/18 (44%)	*LZTR1*	+	Inclusion	RI	chr22:20996113:20996696	(Inoue et al., 2021)
	AML	BM-MNC/PB	9/427 (2.1%)						

A3SS: Alternative 3’ splite site; ALE: Alternative last exon; AML: Acute myeloid leukaemia; BM-MNC: Bone marrow mononuclear cells; CLL: Chronic lymphocytic leukaemia; Ery: Erythrocytes; Gran: Granulocytes; MDS: Myelodysplastic syndrome; NA: Non-applicable; NOS: Not specified

It is noteworthy that published studies have focused on characterising aberrant alternative splicing mediated by common splicing factor hotpot variants such as *SF3B1*^{K700}, *SRSF2*^{P95} and *U2AF1*^{S34}. Aberrant alternative splicing associated with less common, but clinically relevant hotspot variants, remains to be comprehensively studied and validated. For example, *U2AF1*^{Q157} is associated with worse prognosis compared to *U2AF1*^{S34} in primary myelofibrosis patients (Tefferi et al., 2018), but due to the rarity of this hotspot variant, candidate genes aberrantly spliced by this hotspot variant have yet to be identified in patients or experimentally validated. Moreover, different hotspot variants may yield distinct alternative splicing profile (Pangallo et al., 2020; Shiozawa et al., 2018) and hence this may have implications for variant-specific targeted therapy.

Aside from *trans*-acting variants in splicing factors, *cis*-acting variants at donor and acceptor splice sites have been shown to also contribute to aberrant splicing. Specifically, changes in, and by extension, the abrogation of canonical donor splice site sequence (GT) and acceptor splice site sequence (AG) lead to increased skipping of the corresponding exon (Group et al., 2020; Jayasinghe et al., 2018; Kahles et al., 2018). On the other hand, variants that introduce novel splice sites within the introns lead to the inclusion of novel exons (Group et al., 2020). As a consequence, novel amino acids sequences are generated from these novel isoforms and subsequently presented on the cell surface via nonsense-mediated decay (NMD) pathway that may be amenable to immunotherapy (Jayasinghe et al., 2018; Kahles et al., 2018). It is noteworthy that studies which investigated the impact of *cis*-acting genetic variants on alternative splicing have mainly focused on solid tumours. It would be of particular interest to investigate the extent of contribution of *cis*-acting genetic variants on alternative splicing in haematological tumours which on average have lesser mutational burden compared to solid tumours (Alexandrov et al., 2013).

In addition to genetic variants, aberrant expression of components of the splicing machinery can also lead to aberrant splicing. For example, up-regulation of RNA-binding protein *RBM39* is associated with transcriptome-wide increased in exon-skipping and intron retention, and CRISPR-mediated knock-out (KO) of *RBM39* revealed its up-regulation to be essential for acute myeloid leukaemia (AML) survival (E. Wang et al., 2019).

Lastly, aberrant splicing may occur even in the absence of detectable *trans*- or *cis*-acting variants or aberrant RBP expression. For example, *EZH2* which is known to

harbour truncating variants in myeloid neoplasm has shown to be aberrantly spliced in a substantial number of AML patients without any detectable *EZH2* truncating variants (Rivera et al., 2021). Specifically, aberrantly spliced *EZH2* constituted three-quarters of all patients with reduced functional *EZH2* compared to one-quarter of that contributed by *EZH2*^{MUT}. Furthermore, transcriptome-wide aberrant splicing has been shown to have prognostic value in AML patients (Anande et al., 2020), thus suggesting that aberrant splicing contributes to patient heterogeneity and disease outcome in addition to genetic variants and gene expression programme.

Taken together, the different mechanisms of aberrant splicing, attributable to genetic variants or aberrant gene expression or otherwise, collectively constitute a non-negligible proportion of haematological cancer patients. In haematological cancers with substantial proportion of patients with splicing factor genetic variants, such as MDS and MPN, there remains limited success in targeted therapy centred on patients with splicing factor genetic variants (E. Wang & Aifantis, 2020). On the other hand, haematological cancers with few somatic variants such as AML, aberrant splicing, while being able to serve as an additional avenue for targeted therapy, remains an untap area for therapy development. Therefore, there is an urgent need to identify splicing-based biomarkers that may lead to novel and improved therapies for haematological cancer patients.

1.1.5 Motivation for single-cell alternative splicing analysis

Whole tissue samples used for bulk RNA-seq represent heterogeneous cell populations. For example, solid tumours of epithelial origin, such as breast cancer, consist primarily of cancer stem cells, epithelial cells, infiltrating immune cells, supporting blood vessels, and connective tissues (L. Ren et al., 2021). Another example is the bone marrow that consists primarily of mature haematopoietic cells, haematopoietic stem and progenitor cells (HSPCs), mesenchymal stem cells, endothelial cells, fibroblasts and other non-haematopoietic cell types (Dolgalev & Tikhonova, 2021).

One approach to enrich for cell populations of interest is to perform microdissection prior to high-throughput sequencing. For example, microdissection may be performed on breast cancer tissues to segregate the breast epithelium from the stromal cells, and then subsequently sequence and analyse both compartments

separately (Lessi et al., 2019). Another approach is fluorescence-activated cell sorting (FACS) using antibodies complementary to specific cell surface markers to enrich for specific cell types prior to sequencing. The latter approach is used extensively to enrich for haematopoietic and non-haematopoietic cell types to characterise the bone marrow niche in both health and disease states (Y. Chen et al., 2018). This approach requires prior knowledge of the surface markers of the cell populations of interest, and therefore precludes the characterisation of rare cell populations and discovery of novel cell populations.

Single-cell genomic and transcriptomic analysis is able to deconvolute presumed phenotypically homogeneous cell populations, based on cell surface markers, into biologically relevant novel and rare sub-populations. For example, in chronic myeloid leukaemia (CML), single-cell analysis of $CD34^{+}CD38^{-}$ cell population revealed two *BCR-ABL*$^{+}$ cell populations that displayed proliferative or quiescence gene expression signature (Giustacchini et al., 2017). The latter cell population was found to be the dominant cell population in patients one year after receiving tyrosine kinase inhibitor. In myelofibrosis (MF) patients, single-cell analysis of CD34+ cell population revealed an expanded megakaryocyte cell population in patients relative to healthy donors. Further sub-clustering analysis of the megakaryocyte cell population revealed distinct sub-populations, only one of which expressed *AURKA*, a target of alisertib (Psaila et al., 2020). Furthermore, detailed single-cell analysis of haematopoietic stem cell/multipotent progenitor (HSC/MPP) compartment across several stages of human development revealed sub-populations of cells distinguishable by cell cycle gene expression programme (Roy et al., 2021). Specifically, quiescence gene expression programme was increased in the direction of early foetal life to adulthood. Conversely, proliferative gene expression programme was increased in the direction of adulthood to early foetal life. More recently, single-cell analysis of total mononuclear cells from COVID-19 patients revealed cell populations with aberrant gene expression programme such as dysfunctional mature neutrophils and monocytes with down-regulated antigen presentation pathway in severe COVID-19 patients (Schulte-Schrepping et al., 2020; Yao et al., 2021).

To date, most single-cell studies focused on gene expression analysis. Alternative splicing represents an underappreciate layer of complexity underlying gene expression programme (Song et al., 2017). In haematological cancers with substantial proportion of patients with splicing factor genetic variants, such as MDS and MPN, it

would be of particular interest to characterise aberrant splicing patterns specific to certain cell populations. For example, identification of aberrantly spliced genes in the erythroid compartment can reveal possible mechanisms by which aberrant splicing may lead to dysregulated erythroid production, and by extension, disease phenotype, such as anaemia. Furthermore, identification of aberrantly spliced genes in the disease-propagating HSPC compartment may reveal biomarkers amenable to targeted therapy.

It is also conceivable that single-cell alternative splicing analysis would be insightful in haematological cancers without substantial proportion of patients with splicing factor genetic variants, such as AML. Specifically, single-cell alternative splicing analysis may complement existing single-cell gene expression analysis by identifying the specific isoforms that underlie candidate genes detected from differential gene expression analysis. Afterall, following candidate gene selection from high-throughput sequencing and analysis, specific isoforms would need to be identified for downstream functional studies. For example, specific exons would need to be identified for CRISPR-mediated knocked-out of the corresponding gene or specific isoforms would need to be identified for overexpression analysis in cell lines to study the cellular phenotype engendered by the corresponding gene of interest. It is noteworthy that although isoforms of a given gene of interest may be retrieved from publicly available databases such as GENCODE (Harrow et al., 2012), Ensembl (Yates et al., 2020), and genome browsers (Karolchik et al., 2003; Robinson et al., 2011), alternative splicing analysis may reveal novel isoforms not yet reported in these publicly available databases and, more importantly, identify isoforms that are specifically expressed in the researcher's dataset, and therefore pinpoint the most biologically relevant isoforms for downstream functional studies.

1.2 Technological advances in single-cell alternative splicing analysis

Early technologies for single-cell alternative splicing analysis are reverse transcription polymerase chain reaction (RT-PCR) and single-molecule fluorescence *in situ* hybridisation (smFISH). Later and more advanced technologies are plate- and droplet-based single-cell library preparation methods for high-throughput sequencing. While RT-PCR and smFISH are pioneer methods, they remain important tools for validation of candidate alternative splicing events detected from plate- and droplet-

based methods, and are therefore not obsolete for single-cell alternative splicing analysis. The advantages and disadvantages of early and advanced technologies and the ways they complement each other will be explored here.

In the next section, we will review the computational tools available to accommodate single-cell alternative splicing analysis from scRNA-seq data generated from high-throughput next generation sequencing. Briefly, the two most popular tools to date are BRIE (Huang & Sanguinetti, 2021) and Expedition (Song et al., 2017). BRIE infers alternative splicing based on genomic sequence features and sequencing reads whereas Expedition quantifies alternative splicing based on sequencing reads alone. Other tools such as SCATS (Y. Hu, Wang, & Li, 2020) and DESJ-detection (S. Liu et al., 2021) quantify alternative splicing at the pseudo-bulk level by combining all cells for a given pre-defined cell type.

1.2.1 RT-PCR

The earliest method used to investigate alternative splicing in single cells is the RT-PCR. This method utilises sequence-specific primers and reverse transcriptase enzyme to isolate and amplify the isoforms of interest, follow by subjecting the PCR products to gel electrophoresis to check for the absence or presence of the isoforms.

The isoform sequences would need to be known *a priori* because sequence-specific primers are needed to isolate the isoforms of interest. Before the advent of next-generation sequencing, several approaches for identifying alternative splicing for characterisation in single cells have been reported. The first approach is to perform RT-PCR on whole tissue and subjecting the PCR products to gel electrophoresis. Any PCR products whose length do not conform to previously reported isoforms will be subjected to Sanger sequencing. The genomic or amino acid sequences are then aligned to known isoforms to identify novel insertions (exon inclusion) and deletions (exon exclusion) (Castro et al., 2007; Graf et al., 2005; Steinboeck & Kristufek, 2005). The second approach is to compare the alignments of all previously reported tissue-specific cDNA clones against one another (Kanumilli et al., 2006). The third approach is to collate all previously characterised isoforms from the literature (Kumazaki, Mitsui, Hamada, Sumida, & Nishiyama, 1999; Springer, McGregor, Fink, & Fischer, 2003). The expression of these known and novel isoforms detected from whole tissue will then be subsequently characterised at the single-cell level.

Characterisation of alternative splicing in single cells from a specific cell population found that most single cells express only one isoform, but rarely two or more isoforms in the same cell (Castro et al., 2007; Graf et al., 2005; Kanumilli et al., 2006; Springer et al., 2003). For example, analysis of four different isoforms of the *PPT-A* gene in 144 single cells derived from neurones of the nodose ganglion found that each cell expresses only one of the four isoforms (Springer et al., 2003). Another example is the analysis of two different isoforms of the *Cd6* gene in mature T cells and thymocytes whereby >75% of single cells express only one of the two isoforms (Castro et al., 2007).

The observation of the expression of one dominant isoform in single cells is in contrast to that observed at the whole tissue level whereby multiple isoforms are found to be expressed. This may be due to the heterogeneous cell populations with varying isoform expressions that constitute the whole tissue of interest. Indeed, human atrial cells are found to express two different isoforms of the *CACNA1C* gene. However, separation of human atrial cells into cardiomyocytes and non-cardiomyocytes found that cardiomyocytes express only the short isoform of *CACNA1C* whereas non-cardiomyocytes express both isoforms (Graf et al., 2005).

Taken together, early single-cell analysis of alternative splicing events has already demonstrated the power of single cells to reveal cell type-specific isoform expressions that may be missed or are difficult to deconvolute at the bulk level. However, the low-throughput approach by RT-PCR precludes multiplexing large number of single cells and isoforms for analysis. Moreover, the requirement for knowing the isoform nucleotide sequences beforehand may preclude the discovery of novel isoforms. Nevertheless, RT-PCR remains an important tool for validating both known and novel isoforms identified from high-throughput next-generation sequencing platforms (Falcao et al., 2019; Song et al., 2017).

1.2.2 smFISH

smFISH utilises sequence-specific fluorescent probes to bind to mRNA molecules of interest. These mRNA molecules are then visualised and enumerated using microscopy. There are two approaches for delineating the different isoforms of a given gene using smFISH, namely using probes complementary to isoform-specific exons or splice junctions.

In the former approach, fluorescent probes are designed to target exons that can distinguish the different isoforms. For example, in the case of alternative last exon usage, whereby isoform-A and -B utilise different exons as their last exon, these exons can be distinguished by using fluorescent probes with two different colours. Multiple probe copies are used for the same isoform to increase the signal-to-background noise ratio for isoform detection. Analysis of two different isoforms of *CAPRIN1* gene in single cells derived from HeLa and Rpe1 cell lines revealed the dominant isoform to be 20 times more abundant than the minor isoform (Waks, Klein, & Silver, 2011). On the other hand, the two different isoforms of *MKNK2* are found to be in roughly equal proportions in these cell lines (Waks et al., 2011).

In the latter approach, fluorescent probes are designed to target isoform-specific exon-exon junctions (splice junctions). In contrast to the former approach, the splice junction of interest in each mRNA molecule may only be targeted using a single fluorescent probe (Figure 1.7). This may severely influence the sensitivity of the assay. Indeed, analysis of *ErbB4* isoforms using splice junction-specific probes identified on average 1-2 mRNA molecules per cell (Erben, He, Laeremans, Park, & Buonanno, 2018) whereas analysis of *CAPRIN1* isoforms using exon-specific probes identified 100-200 mRNA molecules per cell (Waks et al., 2011).

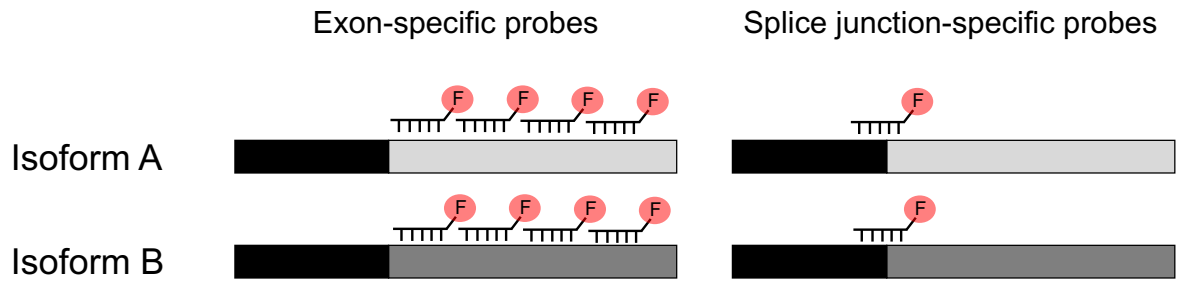


Figure 1.7: Comparison of the relative number of probes targetable to isoform-specific exons and splice junctions.

One method to overcome the low sensitivity of splice junction-specific probes for smFISH is to first apply a modified non-fluorescent probe that can be amplified *in situ* after binding to the splice junction of interest. The amplified probe, now containing multiple tagging sites for the fluorophores, can be tagged by the fluorophores and the signals subsequently detected using microscopy (X. Ren et al., 2018). Using this

method, analysis of three different isoforms of *CD45* gene in single Jurkat T cells revealed *CD45RB* isoform to constitute ~50% of *CD45* molecules detected whereas the remaining molecules consisted of the *CD45RO* and *CD45RA* isoforms in roughly equal proportions (X. Ren et al., 2018).

Similar to RT-PCR, smFISH revealed the dominant expression of one type of isoform for a given gene in single cells. smFISH also restricts the analysis to a small number of different isoforms in single cells and requires prior knowledge of the isoform sequences. Nevertheless, both RT-PCR and smFISH remain important tools for validating both known and novel isoforms identified from high-throughput next-generation sequencing platforms (Falcao et al., 2019; Song et al., 2017).

1.2.3 Plate-based library preparation

Next-generation sequencing, also known as massively parallel sequencing, has enabled high-throughput whole-transcriptome analysis from a wide variety of tissue types (Kahles et al., 2018; Wen & Leong, 2019). Early library preparation techniques were specifically developed for whole (bulk) tissue. Nevertheless, these techniques required large amount of starting material, typically in micrograms (µg), which made library preparation from single cells infeasible.

mRNA-Seq is the first reported assay for preparing single cells for next-generation sequencing (F. Tang et al., 2009). In this method, single cells are individually isolated and lysed. RNA molecules containing poly(A) tails are captured using poly(T) primers anchored to a universal primer sequence. After the first round of cDNA synthesis, synthetic poly(A) tails are attached to the 5'-ends of the cDNA molecules using terminal deoxynucleotide transferase. Poly(T) primers anchored to a different universal primer sequence are subsequently used to amplify the cDNA molecules. After several rounds of PCR amplification, the cDNA molecules are mechanically sheared with sonication and the small-sized fragments are ligated with adaptor sequences for short-read RNA-seq. As a proof of principle, mRNA-Seq was first applied on single cells derived from mouse oocytes, blastomeres, inner cell mass (ICM) of blastocytes, and embryonic stem cells (ESCs) (F. Tang et al., 2010; F. Tang et al., 2009). Notably, *Dppa4* was demonstrated to be alternative spliced whereby the splice junction specific to NM_001018002 isoform was found to be more highly

expressed in ICM whereas the splice junction specific to NM_028610 isoform was found to be more highly expressed in ESC.

Smart-seq employed similar library preparation strategy as mRNA-Seq (Ramskold et al., 2012). One main innovation of Smart-seq is that instead of appending synthetic poly(A) tails, Smart-seq appends several non-templated C nucleotides (CCC) to the 5'-ends of the cDNA molecules after the first round of cDNA synthesis. Template switching oligos (TSOs) are then base-paired with these C nucleotides prior to the next round of amplification. After several rounds of PCR amplification, the cDNA molecules may be mechanically sheared as per mRNA-Seq protocol (F. Tang et al., 2009). Alternatively, the cDNA molecules may be fragmented using Tn5 transposase (tagmentation) prior to adaptor sequence ligation for short-read RNA-seq. As a proof of principle, Smart-seq was first applied on single cells derived from PC3, LnCaP, and T24 cell lines (Ramskold et al., 2012). Notably, *NEDD4L* was found to be alternative spliced whereby exon 4 was observed to be expressed across LNCaP, but not across T24 single cells.

Following up from Smart-seq, Smart-seq2 further increased both yield and length of the cDNA libraries from single cells (Picelli et al., 2013; Picelli et al., 2014). To this end, Smart-seq2 exchanges a single guanylate for a locked nucleic acid (LNA) guanylate at the TSO 3'-end (rGrG+G), increases Mg+ concentration in the SMARTer buffer, adds dNTPs prior to RNA denaturation (as opposed to adding dNTPs in the reverse transcription master mix), adds betaine into the reaction mix, and eliminates the bead extraction step. Similar to Smart-seq, tagmentation may be used to achieve smaller fragments amenable for short-read RNA-seq. Alternatively, the tagmentation step may be skipped to preserve the full-length cDNA molecules for long-read RNA-seq (Byrne et al., 2017). To date, Smart-seq2 remains the most widely used plate-based library preparation protocol for both singe-cell gene expression and alternative splicing analysis compared to other methods (Hayashi et al., 2018; Islam et al., 2011; F. Tang et al., 2009; L. Wu et al., 2015). Indeed, Smart-seq2 combined with short- or long-read RNA-seq has been used to characterise the alternative splicing landscape in a variety of cell types including induced pluripotent stem cells, neural progenitor cells, motor neurons, endoderm cells, immune cells, and cancer cells (Linker et al., 2019; Manipur, Granata, & Guarracino, 2019; Singh et al., 2019; Song et al., 2017).

More recently, Smart-seq3 was developed. The main innovation of Smart-seq3 is the attachment of a unique molecular identifier (UMI) to each TSO, effectively

tagging each cDNA molecule at the 5'-end (Hagemann-Jensen et al., 2020). This enables *in silico* isoform reconstruction using sequencing reads with the same UMI tag and subsequent quantification of isoforms at the single-molecule level. Smart-seq3 analysis of single cells derived from HEK293FT cell line demonstrated that half of the molecules per cell could be assigned to a known isoform, and isoforms of up to 4kb have been reconstructed successfully.

Although plate-based library preparation methods are labour-intensive and low-throughput, indeed most plate-based studies to date analysed on average only a few hundred single cells, they nevertheless are the first methods to enable transcriptome-wide analysis of alternative splicing events in single cells when coupled with next-generation sequencing platforms.

1.2.4 Droplet-based library preparation

Heterogeneous cell populations such as those of the bone marrow mononuclear cells and central nervous system consist of many cell populations with differing population sizes. Low-to-moderate throughput library preparation methods such as plate-based methods may not be sufficiently sensitive for gene and alternative splicing analysis in rare cell populations. Droplet-based library preparation methods were developed to automate single-cell library preparation process and to increase the number of cells for next-generation sequencing.

inDrop (indexing droplets) and Drop-Seq are pioneers in developing droplet-based library preparation protocols for single cells (Klein et al., 2015; Macosko et al., 2015). Both techniques utilise a custom microfluidic device containing separate inlets for cells, barcoded primers, lysis buffer, and oil to generate discrete droplets for each cell. Each barcoded primer consists of the cell barcode, UMI sequence, and poly(T) tail. In each droplet, the cell is lysed, and the barcoded primers are used to generate cDNA molecules. Next, the droplets are broken, and the cDNA molecules are collectively amplified prior to downstream next-generation sequencing.

The key difference between inDrop and Drop-Seq is the medium of delivery of the barcoded primers. In the former, the barcoded primers are linked to polyacrylamide mesh via photo-cleavable linkers and are encapsulated in hydrogels. The barcoded primers are then released in the droplets using ultraviolet (UV) light for cDNA generation in free suspension within the droplet (Klein et al., 2015). In the latter, the

barcoded primers are attached to the surface of microparticles (beads) and cDNA generation occurs on the beads themselves (Macosko et al., 2015).

10x Genomics employs similar droplet-based library preparation protocols for single cells, and at the same time, addresses several limitations of inDrop and Drop-Seq (Zheng et al., 2017). Firstly, not all droplets will contain single cells after the droplet-generation step in inDrop and Drop-Seq. For example, only ~10% of droplets are occupied by cells in inDrop (Klein et al., 2015) whereas 10x Genomics demonstrated capture rate of ~50% while maintaining negligible number of droplets containing >1 cell (doublets). This will mitigate sample wastage, especially for low-volume samples. Secondly, 10x Genomics developed an 8-channel microfluidic chip that enables multiplexing of up to 8 samples in a single run. Thirdly, 10x Genomics developed a computational workflow called CellRanger to conveniently pre-process single-cell next-generation sequencing data, including sequence alignment, cell barcode and UMI sequence correction, and gene expression quantification. To date, 10x Genomics remains the most popular high-throughput droplet-based library preparation method.

The development of droplet-based library preparation methods was motivated by gene expression profiling of large cell populations. Therefore, in contrast to plate-based method, the original droplet-based studies did not demonstrate alternative splicing as a proof of principle for their methods. Nevertheless, studies on alternative splicing analysis using droplet-based data are gradually emerging (Dehghannasiri, Olivieri, Damljanovic, & Salzman, 2021; Kaminow, Yunusov, & Dobin, 2021).

1.3 Computational approaches for single-cell alternative splicing analysis

There already exists several robust and popular analytical frameworks to enable comprehensive characterisation of gene expression profiles in single cells, such as Seurat (Satija, Farrell, Gennert, Schier, & Regev, 2015), SingCellaR (G. Wang et al., 2022), Monocle (Trapnell et al., 2014), and Scanpy (Wolf, Angerer, & Theis, 2018). These analytical frameworks provide functionalities for gene expression normalisation, batch/donor correction, dimension reduction, clustering analysis and cell type identification, differential expression analysis, and pathway enrichment analysis.

Current analytical frameworks for single-cell alternative splicing analysis have not match more established gene expression analytical frameworks in terms of providing wide-ranging functionalities to enable comprehensive characterisation of the splicing landscape in single cells. Current analytical frameworks for single-cell alternative splicing analysis may be broadly categorised into whether alternative splicing is characterised at the level of gene, isoform, or splice junction level.

1.3.1 Gene-level analysis

Early single-cell alternative splicing analysis inferred differential isoform usage for a given gene without explicitly quantifying the isoform expression levels. Computational frameworks that utilise this approach include SingleSplice (Welch, Hu, & Prins, 2016), logistic regression (Ntranos, Yi, Melsted, & Pachter, 2019), and ISOP (Vu et al., 2018). The common features shared by these frameworks and their unique features will be elaborated here.

SingleSplice first performs *de novo* assembly using short reads to construct the longest piece of transcript for a given gene (Welch et al., 2016). This conceptual transcript is termed a directed, acyclic splice graph. All possible isoforms are subsequently detected using this conceptual transcript as a reference. These isoforms are referred to as paths through the graph. Next, alternative splicing modules (ASMs) representing alternative splicing events across the different isoforms are detected and their corresponding expressions are quantified. Using this information, SingleSplice assesses whether the observed ASM-A to ASM-B ratio is above the variation in ratio due to technical noise. Genes with observed ASM-A to ASM-B ratio that exceeds the simulated variation in ratio due to technical noise are considered as differentially spliced. This process is applicable to genes with two or more ASMs detected.

Similar to SingleSplice, ISOform-Patterns (ISOP) focuses on the ratio of two isoforms for a given gene to detect differentially splice genes (Vu et al., 2018). Specifically, ISOP categorises the isoform expression patterns into three broad classes. Pattern I/II represent equal expression of both isoform A and B in a given cell population. Pattern V/VI represent bimodal expression of isoform A (some cells express the isoform while other cells do not) and unimodal expression of isoform B in a given cell population. Pattern X/XI represent bimodal expression of both isoform A

and B in a mutually exclusive manner, i.e., when isoform A is expressed, isoform B is not, and vice versa.

It is noteworthy that both SingleSplice and ISOP only enable detection of differentially spliced genes within a cell population, but not across two or more different cell populations (Vu et al., 2018; Welch et al., 2016). Logistic regression was subsequently showed to be able to detect differentially splice genes between two cell populations (Ntranos et al., 2019). Logistic regression is a supervised machine learning classification algorithm used to predict the probability of an observation (single cell) belonging to a group (cell population A or B) based on several given features (isoform A and B expression). Ntranos *et al*. used logistic regression to assess whether the two most highly expressed isoforms for a given gene were able to distinguish two different cell populations. Genes whose isoforms have the predictive power to distinguish the two different cell populations are considered to be differentially spliced.

The aforementioned frameworks were shown to be able to detect genes that were differentially spliced, but not differentially expressed within or across cell populations (Ntranos et al., 2019; Vu et al., 2018; Welch et al., 2016). For example, SingleSplice detected differentially spliced genes within a presumed homogeneous population of single cells derived from mouse embryonic stem cells that reflected subpopulations of single cells at different cell cycle stages, i.e., G1, S, and G2M (Welch et al., 2016). Furthermore, logistic regression identified *CD45* to be differentially spliced between naïve T cells (CD4+CD45RA+CD25-) vs memory T cells (CD4+CD45RO+) in the absence of any detectable gene expression changes (Ntranos et al., 2019). Hence, these early applications of single-cell alternative splicing frameworks have demonstrated that alternative splicing represents an additional layer of complexity underlying and invisible at gene expression level

1.3.2 Exon-level analysis

While gene-level alternative splicing analysis frameworks enable the detection of differentially spliced genes, they neither explicitly quantify the isoform expression levels nor do they quantify the degree or extent of isoform changes within or across different cell populations.

Exon-level alternative splicing events are measured in terms of percent spliced-in (PSI). For a given alternative splicing event in a given cell, a PSI of 100 means that only one dominant isoform is expressed in the cell, specifically the isoform with the alternative exon spliced in. A PSI of 0 also means that only one dominant isoform is expressed in the cell, specifically the isoform with the alternative exon spliced out. 0 < PSI < 100 means that both isoforms are co-expressed in the cell (Figure 1.8).

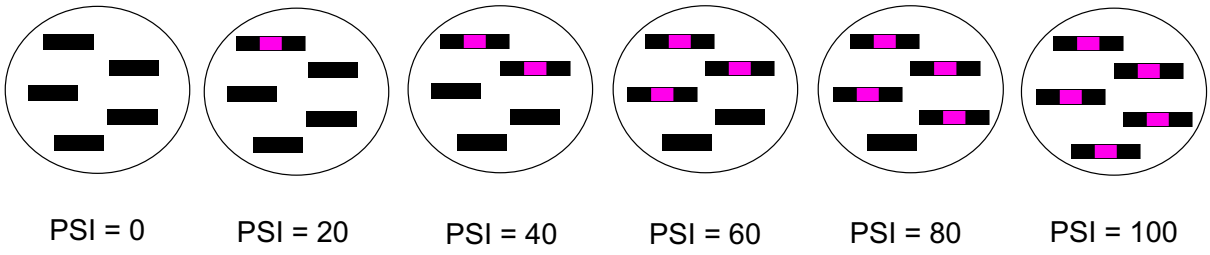


Figure 1.8. Example of PSI values arranged in ascending order. PSI values are defined as the percentage of isoforms with the alternative exon spliced in (pink).

Current analytical frameworks employ either one of the two approaches for PSI estimation, namely Bayesian- or sequencing read-based approach.

1.3.2.1 Bayesian approach for PSI estimation

BRIE (Bayesian regression for isoform estimation) is the first analytical framework developed specifically for single-cell alternative splicing analysis (Huang & Sanguinetti, 2017). This Bayesian model combines an informative prior together with a likelihood to predict PSI values. The informative prior consists of 735 curated genomic sequence features such as alternative exon sequence conservation score (phastCons), splice site motifs, and intron length. The likelihood consists of sequencing reads aligning to the alternative exon body and its flanking constitutive exon bodies. This likelihood is a mixture model borrowed from MISO – a bulk alternative splicing software (Katz, Wang, Airoldi, & Burge, 2010).

For a given alternative exon in a given cell, when the coverage is high, the model places more weight on the likelihood (sequencing reads) to predict the PSI value. When the coverage is moderate, the model places similar weights on both the likelihood (sequencing reads) and the informative prior (genomic sequence features)

to predict the PSI value. Lastly, when coverage is low, the model places more weight on the informative prior (genomic sequence features) to predict the PSI value. Therefore, BRIE should in theory be able to circumvent the issue of low coverage by imputing the PSI values in single cells with low coverage. Nevertheless, the accuracy of BRIE to predict PSI values in low coverage scenarios has not been systematically assessed in the original study (Huang & Sanguinetti, 2017).

Recently, it was revealed that PSI value imputation of alternative splicing events in cells with low coverage leads to inaccurate PSI estimation by BRIE (W. Liu & Zhang, 2020). Specifically, the cell-to-cell correlation for PSI values in single cells derived from homogeneous cell populations is poor when alternative splicing events with low coverage are included for analysis. It was discovered that at low coverage, BRIE tends to predict PSI values of ~50. This is consistent with the likelihood (sequencing reads) mixture model that BRIE borrowed from MISO which defines a PSI value of 50 by default at low coverage (Katz et al., 2010). PSI value of 50 means that both isoforms (isoform A that splice in the alternative exon and isoform B that splice out the alternative exon) are co-expressed in the same cell. The default value of 50 is sensible in bulk alternative splicing analysis because more than one isoform is likely to be expressed in bulk samples (Katz et al., 2010; S. Liu et al., 2021). However, it has been widely reported that single cells are more likely to express only one dominant isoform (W. Liu & Zhang, 2020; Westoby, Herrera, Ferguson-Smith, & Hemberg, 2018). This translates to either PSI of 100 (the alternative exon spliced in) or 0 (the alternative exon spliced out). Hence, the affinity to predict exon inclusion rate ~50 by BRIE in low coverage settings may not be suitable for single cells. Indeed, removing splicing events and cells with low coverage improves the cell-to-cell correlation of the exon inclusion rate estimated by BRIE (W. Liu & Zhang, 2020). It is noteworthy that BRIE presently does not automatically identify or filter away splicing events or single cells with insufficient reads.

Notwithstanding the accuracy of PSI estimation at low coverage, the Bayesian model remains an elegant approach for estimating PSI values because it is possible to incorporate additional informative priors, aside from genomic sequence features, to improve PSI estimation. For example, incorporation of methylation profile, such as methylation of alternative exon, is able to significantly, albeit very marginally, improve the accuracy of PSI estimation (Linker et al., 2019).

Presently, the utility of genomic sequence features as an informative prior has only been demonstrated for skipped-exon (SE) alternative splicing event. It would be of particular interest to assess the usefulness of genomic sequence features as an informative prior for other alternative splicing event types, namely mutually exclusive exons (MXE), retained intron (RI), alternative 5' and 3' splice sites (A5SS and A3SS, respectively) and alternative first and last exons (AFE and ALE, respectively). Afterall, alternative splicing events other than SE have been shown to play important roles in both health and disease states (Shiozawa et al., 2018; Smart et al., 2018). Moreover, genomic sequence features as informative prior has only been curated for human and mouse genomes. It would therefore also be of particular interest to assess genomic sequence features as an informative prior for other mammalian species.

1.3.2.2 Read-based approach for PSI estimation

Expedition is the first read-based alternative splicing analytical framework developed for single-cell analysis (Song et al., 2017). In contrast to Bayesian approach which incorporates both genomic sequence features and sequencing reads to estimate PSI values, the read-based approach incorporates only sequencing reads for estimating PSI values. Specifically, splice junction reads are used to compute the PSI values. The read-based approach offers several advantages over the Bayesian approach for estimating PSI values.

Firstly, the read-based approach does not depend species-specific genomic sequence features to estimate the PSI values. Therefore, this approach is applicable to any species with well-annotated genome. For example, human, mouse, cow, chicken, xenopus, zebrafish, amphioxus, sea urchin, fruit fly, centipede, Caenorhabditis elegans, planarian, sea anemone, and Arabidopsis thaliana (Tapial et al., 2017).

Secondly, the PSI values estimated by the read-based approach are based on observed sequencing reads and therefore are more likely to represent true biological phenomena (Song et al., 2017). For example, a PSI value of 75 reflects three-quarters of the sequencing reads supporting the inclusion (spliced in) of the alternative exon whereas one-quarter of the sequencing reads supporting the exclusion (spliced out) of the alternative exon. Moreover, single cells in which alternative exons with low-to-zero coverage (dropouts) are excluded from downstream analysis due to the

uncertainty in estimating the PSI values. On the other hand, the Bayesian approach uses a PSI value of 50 as a prior, hence the PSI value assigned to alternative exons with low-to-zero coverage would be skewed towards 50 (Huang & Sanguinetti, 2017; W. Liu & Zhang, 2020). A PSI value of 50 as a prior may not always represent the true biology phenomena, especially for lowly expressed genes with high dropout rates. Indeed, majority of alternative exons with moderate-to-high coverage have PSIs of ~100 or ~0 in single cells (W. Liu & Zhang, 2020; Marinov et al., 2014; Song et al., 2017).

Despite the advantages of read-based approach compared to Bayesian approach for estimating PSI values, Expedition has only been applied by one study (excluding the original publication) to date (W. Liu & Zhang, 2020) whereas BRIE has been applied by at least five studies (excluding the original publication) (Falcao et al., 2018; Y. Li, Chen, et al., 2021; W. Liu & Zhang, 2020; Manipur et al., 2019; Munoz et al., 2019). Both Expedition and BRIE are available since 2017. One possible reason for the low uptake of Expedition is the large computational power required; Expedition recommends at least 16 cores. This high computational requirement may preclude users without access to advanced computational infrastructure. For example, yours truly was not able to run Expedition before October 2020 because the computer cluster which yours truly had access to at that time had only 4 cores.

SCATS is a more recent exon-level read-based alternative splicing software (Y. Hu et al., 2020). SCATS aggregates cells based on user-defined cell types (pseudobulk) prior to splicing analysis. Notably, SCATS is able to analyse UMI-containing sequencing reads generated from plate- or microfluidic-based (e.g., Fluidigm C1 instrument), but not droplet-based platforms.

Several read-based alternative splicing analytical framework developed for bulk RNA-seq analysis have been applied to single-cell datasets (Y. Chen et al., 2018; Marinov et al., 2014). This is not surprising given that the utilisation of splice junction reads to estimate PSI values is not unique to single-cell analysis. Nevertheless, these bulk-level alternative splicing software do not include features to address the technical noise inherent to scRNA-seq data, including validation of alternative splicing events and modelling of high dropout rates.

It is noteworthy that similar to BRIE, Expedition, and SCATS currently only computes the PSI values for skipped-exon (SE) and mutually exclusive exons (MXE) (Song et al., 2017). Other alternative splicing event types including retained intron (RI),

alternative 5' and 3' splice sites (A5SS and A3SS, respectively) and alternative first and last exons (AFE and ALE, respectively) also play important roles in health and diseased states (Shiozawa et al., 2018; Smart et al., 2018).

1.3.2.3 Modelling PSI distributions

The percent spliced-in (PSI) values for a given alternative splicing event across a cell population can take any values between 0-100. Therefore, the PSI values may be modelled using the beta distribution. The beta distribution is defined by two parameters, namely α and β. By estimating the values of α and β, the PSI distribution may be grouped into discrete categories ("modalities"). The first set of modalities was proposed by Expedition and consists of five modalities, namely included (PSI ~ 100), excluded (PSI ~ 0), bimodal (PSI ~ 100 and PSI ~ 0), middle (PSI ~ 50), and multimodal (uniform distribution) (Song et al., 2017) (Figure 1.9).

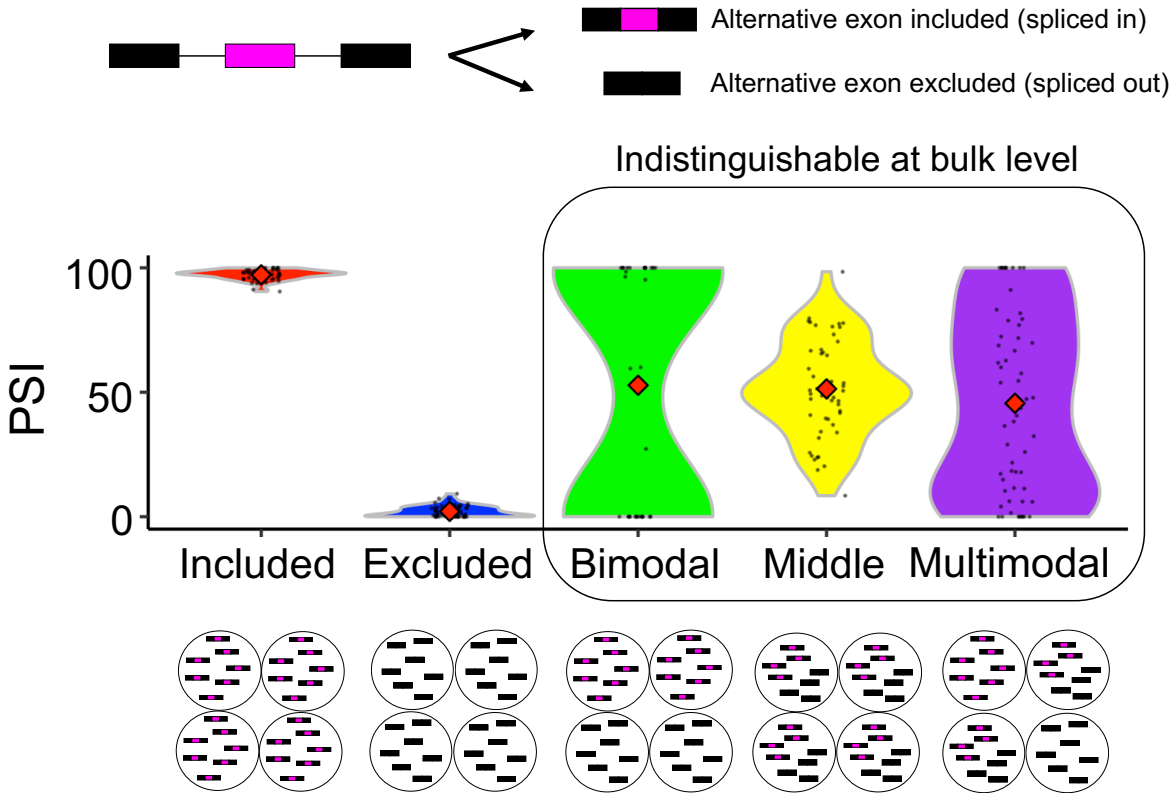


Figure 1.9. Modalities conceived by Expedition. (Top) Two possible isoforms when analysed at the exon level, i.e., one isoform in which the alternative exon is spliced in and another in which the alternative exon is spliced out. **(Middle)** The five modalities by Expedition. Red diamonds represent the average PSI values across the single-cell

population and also the expected PSI values at the bulk level. **(Bottom)** The proportion of isoforms with the alternative exon spliced in (pink) or spliced out in each cell for the respectively modalities. Four single cells shown as representatives of the entire cell population.

Included and excluded modalities may be represented accurately at the bulk level because the average PSI values across a cell population for these two modalities are similar to that at the bulk level. However, it may not be possible to distinguish the bimodal, middle, and multimodal splicing patterns at the bulk level because the average PSI values across a cell population for these modalities are similar, i.e., ~50. Therefore, modality analysis may reveal single-cell alternative splicing patterns that may be indistinguishable at the bulk level.

Modality classification of skipped-exon (SE) and mutually exclusive exons (MXE) in induced pluripotent stem cells (iPSCs), neural progenitor cells, and motor neurons revealed included, excluded, and bimodal modalities to be the most common modality types. They account for ~50%, ~30%, and ~20% of alternative splicing events, respectively (Song et al., 2017). On the other hand, middle and multimodal modalities collectively account for only <1% of alternative splicing events.

While bimodal modality accounts for a significantly proportion (~one-fifth) of alternative splicing events, a recent study using simulated and empirical data have shown that majority of bimodal splicing patterns are false positives, i.e., misclassified (Buen Abad Najar, Yosef, & Lareau, 2020). False bimodal splicing patterns may be attributed to PCR amplification bias of the minor isoform arising during single-cell library preparation, especially when the starting material is low (Figure 1.10). The presence of misclassified bimodal splicing patterns identified from RNA-seq has also been validated using quantitative polymerase chain reaction (qPCR) whereby some bimodal splicing patterns identified in RNA-seq are found to be either included or excluded splicing pattern in qPCR (Song et al., 2017).

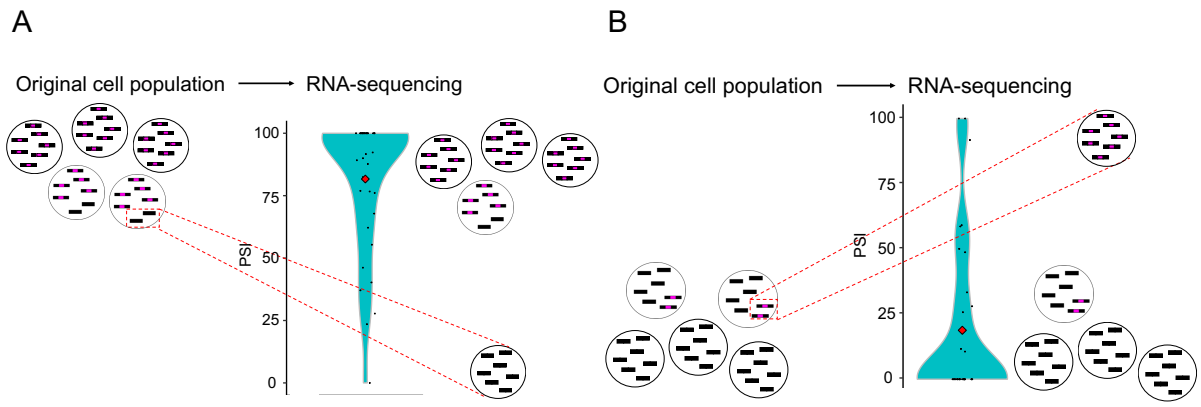


Figure 1.10: PCR amplification bias of minor isoform leads to false bimodal splicing pattern reflected in scRNA-seq data. PCR amplification bias for minor isoform that **(A)** exclude (splice out) or **(B)** include (splice in) the alternative exon (pink) leading to false bimodal splicing patterns.

One approach to mitigate false bimodal classification is to only include highly expressed alternative splicing events, such as alternative splicing events whose genes have inferred mRNA molecular counts of at least 10 (Buen Abad Najar et al., 2020; Qiu et al., 2017). However, this approach would preclude majority of genes from alternative splicing analysis, especially genes with moderate-to-low expression. Therefore, an alternative approach is needed to distinguish true from false bimodal classification to enable more accurate modality assignment but without sacrificing genes with moderate-to-low expression.

It is noteworthy that current single-cell alternative splicing frameworks do not incorporate any approach to identify and make corrections to false bimodal classifications.

1.3.2.4 Differential splicing analysis

Differential gene expression analysis between different cell populations is the cornerstone in gene expression studies as it can identify candidate genes for downstream experimental validation and characterisation. Similarly, after estimating and modelling percent spliced-in (PSI) values, differential splicing analysis is an important step for identifying differentially spliced events for downstream analysis (Shiozawa et al., 2018).

BRIE compares the percent spliced-in (PSI) values between two cells to identify differentially spliced events between this pair of cells (Huang & Sanguinetti, 2017). In order to identify differentially spliced events within a given cell population, all possible pair-wise comparisons would need to be performed. This will lead to an exponential increase in computational power and time with increasingly larger cell populations. One approach to circumvent this limitation is to only perform the differential splicing analysis in a subsample of the cell population. For example, only 20 of 1,205 (1.7%) single cells derived from mouse embryonic cells were included for differential splicing analysis in this study (Huang & Sanguinetti, 2017). Using this down-sampling approach, ~1-2% of alternative splicing events are found to be differentially spliced in single cells derived from human HCT116 cells and mouse embryonic cells (Huang & Sanguinetti, 2021). Although pair-wise comparison of cells may identify differentially spliced events within a cell population, this approach does not enable comparison between groups of cell populations.

Expedition allows for differential splicing analysis between two groups of cell populations. Differentially spliced events are defined as alternative splicing events that exhibit modality change from one cell population to another. For example, *PKM* exon 9 undergoes modality change when induced pluripotent stem cells (iPSCs) differentiate into motor neurons whereby *PKM* exon 9 demonstrates an included modality in the former but a bimodal modality in the latter cell population (Song et al., 2017). Therefore, differential splicing analysis based on modality change allows for qualitative assessment of splicing changes between two cell populations.

BRIE (mode 2) enables quantitative assessment of splicing changes between two cell populations (Huang & Sanguinetti, 2021). To this end, for each alternative splicing event, BRIE (mode 2) fits two models: One model with the cell group information (cell population A and B) provided and another model with the cell group information left out. The strength of the association between an alternative splicing event and the cell groups is represented as evidence lower bounds (ELBO), which approximates the Bayes factor (BF). ELBO and BF are the frequentist equivalent for *P* values. It is noteworthy that BRIE (mode 2) only utilises sequencing reads, without incorporating genomic features, to infer PSI values for the purpose of differential splicing analysis. For cells with missing PSI values (due low-to-no coverage) for a given alternative splicing event, BRIE (mode 2) imputes these missing values using the mean PSI values across cells with sufficient coverage. This approach assumes

the mean PSI is representative of the overall cell population and therefore this approach of differential splicing analysis is recommended only for comparing homogeneous, but not heterogeneous, cell populations (Huang & Sanguinetti, 2021).

1.3.3 Splice junction-level analysis

Plate-based library preparation protocols, such as Smart-seq2, preceded droplet-based library preparation protocols, such as 10x Genomics. Plate-based library preparation protocols yield more-or-less uniform coverage across the isoform molecules that enables exon-level alternative splicing analysis (Picelli et al., 2013; Picelli et al., 2014). Indeed, correcting for 3'-bias coverage during percent spliced-in (PSI) estimation does not substantially improve the precision of PSI values (J. Zhang, Kuo, & Chen, 2015). Therefore, pioneering single-cell alternative splicing frameworks, such as BRIE and Expedition, are developed for analysing splicing events at the exon-level in RNA-seq data generated from plate-based library preparation methods (Huang & Sanguinetti, 2017; Song et al., 2017).

Nevertheless, droplet-based library preparation protocols are increasingly more widely applied compared to plate-based because of the former's ability to process more cells within a shorter period of time in an almost automated fashion (Zheng et al., 2017). The additional challenges during single-cell alternative splicing analysis presented by droplet-based preparation protocols compared to that of plate-based include higher drop-out rates, perverse 3'- or 5'-end coverage bias, and much larger number of cells. Specifically, due to the 3'- or 5'-end coverage bias, exon-level alternative splicing analysis is not feasible for RNA-seq data generated from droplet-based library preparation protocols (Figure 1.11). Therefore, there is an urgent demand for novel single-cell alternative splicing frameworks to address these challenges presented by RNA-seq data generated from droplet-based library preparation protocols.

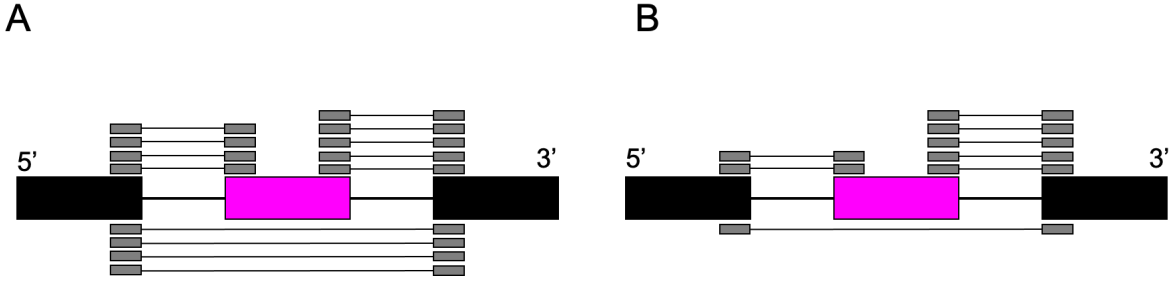


Figure 1.11. Comparison of coverage uniformity across isoform molecules in RNA-seq data generated from plate- and droplet-based library preparation protocols. (A) More-or-less uniform coverage across isoform molecules in RNA-seq data generated from plate-based library preparation protocols enables precise estimation of the PSI value of the alternative exon (pink). **(B)** Coverage decreases from 3' to 5'-end in RNA-seq data generated 3'-bias droplet-based library preparation protocols. This precludes exon-level analysis of alternative splicing events.

SICILIAN was introduced to call high confident splice junctions from RNA-seq data generated from 10x Genomics (Dehghannasiri et al., 2021). To this end, SICILIAN incorporates several variables into a generalised linear model to yield a confidence score for each splice junction. Examples of these variables are the number of supporting reads and the alignment scores. Splice junctions above a user-defined confidence score will be retained. Next, only annotated (previously reported) splice junctions are further retained. Collectively, these steps will mitigate false positive splice junctions (sequencing artifacts) from being included for downstream analysis. Nevertheless, SICILIAN does not perform downstream analysis such as differential splicing analysis. Furthermore, SICILIAN only evaluates splice junctions at the bulk-level, i.e., library/sample-level. On the other hand, STARsolo is able to quantify splice junction expression at the single-cell level, but similar to SICILIAN, it does not provide functionalities for downstream analysis such as differential splicing analysis (Kaminow et al., 2021).

DESJ-detection was introduced to perform differential splicing analysis at the splice junction level, but it only supports RNA-seq data generated from plate-, but not, droplet-based platforms (S. Liu et al., 2021). Moreover, DESJ-detection aggregates cells based on user-defined cell types (pseudobulk) prior to splicing analysis, and hence may underestimate splicing complexing/heterogeneity within a presumed homogenous cell population.

Therefore, there is an urgent need for a more comprehensive analytical framework to provide end-to-end support for users for single-cell alternative splicing analysis of RNA-seq data generated from droplet-based library preparation protocols. Ideally, this framework should provide functionalities for pre-processing such as selection of confident (annotated) splice junctions and assigning each splice junction

to its respective gene, and also for downstream analysis such as differential splicing analysis, pathway enrichment analysis, functional annotation, and visualisation of candidate splice junction expression on low-dimensional space such as uniform manifold approximation and projection (UMAP).

1.3.4 Visual-based validation

Differential splicing analysis between sample groups may yield hundreds, or even thousands, of differentially spliced events (Shiozawa et al., 2018; Song et al., 2017). Therefore, it may not be feasible to bring forward all differentially spliced events for downstream validation such as qPCR or smFISH. One approach to validate differentially spliced events is to visually inspect the coverage and read alignments in a genome browser (Shiozawa et al., 2018; Smart et al., 2018). This visual-based inspection approach of validating initial discoveries from high-throughput next-generation sequencing platforms have been used for genetic variant analysis (Nik-Zainal et al., 2014), gene expression analysis (X. Wang et al., 2015), and chromatin accessibility analysis (Corces et al., 2018).

While current genome browsers are suitable for visual inspection of alternative splicing events at the bulk level, it may not be immediately applicable for visual inspection of alternative splicing events at the single-cell level. Current approach for visual inspection of alternative splicing events in single cells include selecting a subsample of single cells from the entire cell population (Huang & Sanguinetti, 2017; Manipur et al., 2019; Munoz et al., 2019), aggregating read alignments of single cells according to their sample group (e.g. tissue type) (Falcao et al., 2018) or displaying the read alignment profile for all single cells (Munoz et al., 2019; Song et al., 2017).

The read alignment profile of only a subsample of single cells may not capture or may not be representative of the splicing profile of the entire cell population, especially for heterogeneous cell populations. On the other hand, aggregating read alignments of single cells may obscure any variability in splicing patterns underlying the entire cell population. Lastly, displaying the read alignment profile for all single cells included in the study, whereby each panel represents one cell, may not be feasible when the sample size is very large. Moreover, current genome browser displays coverage or read counts instead of percent spliced-in (PSI) values.

Millefy is first software developed specifically for visualising read alignment profile at single-cell resolution (Ozaki, Hayashi, Umeda, & Nikaido, 2020). Nevertheless, Millefy only accommodates visualisation of single-cell gene expression profile.

Taken together, there is an urgent need to develop a visualisation framework for visual inspection of alternative splicing events at single-cell resolution. This will enable initial screen of alternative splicing events identified from high-throughput RNA-seq prior to selection of candidate (true positive) alternative splicing events for downstream experimental validation and functional studies.

1.4 Aims

Single-cell alternative splicing analysis is able to provide additional biological insights, on top of single-cell gene expression analysis, in both healthy and diseased states. Yet, there remains a paucity of single-cell alternative splicing studies. For single-cell studies that did investigate alternative splicing, the alternative splicing analysis was secondary to gene expression analysis and lacked independent validation of candidate alternative splicing events. It is conceivable that this is due to the lack of a comprehensive analytical framework for single-cell alternative splicing analysis and candidate alternative splicing selection. Therefore, the overarching objective of this DPhil project is to address these shortcomings with the ultimate goal of applying our developed tools in blood cancer patients, specifically myeloid neoplasm patients with spliceosome variants, at the single-cell level, e.g., comparing spliceosome-mutant vs wildtype HSCs from with the same patient (intra-patient comparison) and across patients (inter-patient comparison). To this end, we aim to:

a. Develop an analytical framework for single-cell alternative splicing detection, quantification, differential analysis, and functional annotation to enable comprehensive transcriptome-wide characterisation of the alternative splicing landscape at single-cell resolution.

b. Develop an analytical framework for single-cell visual validation of alternative splicing events from (a) based on sequencing reads to identify true positive alternative splicing events.

c. Develop an analytical framework for identifying clinical-relevant and druggable alternative splicing events from (c) for downstream functional studies.

d. Demonstrate the application of analytical frameworks from (a-c) on single-cell genomic and transcriptomic datasets generated from a range of myeloid neoplasms with genetic variants in splicing factor genes.

2 Materials and methods

2.1 MARVEL analysis for scRNA-seq datasets generated from the plate-based method

2.1.1 Pre-processing of publicly available datasets

MARVEL stands for **M**ulti-modal **A**pproach to **R**e**VE**a**L** alternative splicing dynamics at single-cell resolution. The main goal of MARVEL is to perform global characterisation of the splicing landscape between different cell populations to reveal differentially spliced genes in health and diseased states. To this end, the functions of MARVEL are to perform isoform detection and quantification, splicing pattern (modality) assignment, differential gene and splicing analysis, and functional annotation of differentially spliced genes. MARVEL is implemented as an R package and is available on GitHub: https://github.com/wenweixiong/MARVEL.

To benchmark MARVEL against existing software and demonstrate the application of MARVEL on scRNA-seq data generated from plate-based library preparation methods, we selected three datasets consisting of human cell lines from different cell types (Linker et al., 2019; Song et al., 2017; Trapnell et al., 2014) and one dataset consisting of heterogeneous mouse cell populations (Falcao et al., 2018). We primarily chose human cell lines because they consist of homogeneous cell populations and therefore are suitable for optimising MARVEL prior to its application on human samples consisting of heterogeneous cell populations, such as haematopoietic stem and progenitor cells (HSPCs)

Raw FASTQ files were downloaded from Sequence Read Archive (SRA). The adaptor and nucleotide sequences from the 3'-end with Phred quality scores < 20 [*-q 20*] were trimmed using TrimGalore (version 0.5.0) (Martin, 2011). Trimmed reads were mapped to the human or mouse reference genome (GRCh38 or GRCm38) for human and mouse datasets, respectively, using STAR aligner (version 2.6.1d) with default settings. In the first round of alignment, the splice junctions expressed in each sample (cells and matched-bulk samples if available) were detected. In the second round of alignment, the binary alignment map (BAM) files were generated. This second round of alignment additionally generates the splice junction counts files for each sample based on the list of splice junctions detected across all samples from the first round of alignment.

For each sample, the total number of reads were retrieved from the log files generated from TrimGalore, and the total number of reads successfully mapped to the reference genome were tabulated using Samtools (version 1.9) [*stats*]. The alignment rate was then computed as the total number of mapped reads divided by the total number of reads and then multiplied by 100. Samtools [*view*] was used to tabulate the total number of mitochondrial reads. The percentage of mitochondrial reads was subsequently computed as the total number of mitochondrial reads divided by the total number of reads and then multiplied by 100. The alignment rate, the total number of mapped reads, and the percentage of mitochondrial reads were used to stratify samples into those that passed or failed sequencing quality control (QC).

The first dataset consists of induced pluripotent stem cells (iPSCs), neural progenitor cells (NPCs), and motor neurons (MNs) (Song et al., 2017). Both NPCs and MNs were differentiated *in vitro* from iPSCs. Libraries were prepared using Smart-seq protocol (Ramskold et al., 2012). In total, 206 single cells and 8 matched bulk samples were sequenced. One-hundred and seventy-four single cells and all bulk samples were sequenced in 100bp paired-end (PE) mode, while 32 single cells were sequenced in 100bp single-end (SE) mode. Among the single cells, 16 single cells were removed because either they were annotated as outliers by the original study or did not pass sequencing QC. For single cells sequenced in PE mode, single cells that passed sequencing QC were defined as having > 70% alignment rate and the percentage of mitochondrial reads < 15% (Figure 2.1A-C). For single cells sequenced in SE mode, single cells that passed sequencing QC were defined as having > 70% alignment rate (Figure 2.1D-F). In total, 190 single cells consisting of 62 iPSCs, 68 NPCs, and 60 MNs were included for downstream analyses.

The second dataset consists of myoblast cells cultured and then sequenced at 0-, 24-, 48-, and 72-hours (hrs) (Trapnell et al., 2014). In total, 372 single cells and 12 matched bulk samples were sequenced. Libraries were prepared using Smart-seq protocol (Ramskold et al., 2012) and sequenced in 100bp PE mode. Among the single cells, 45 single cells were removed because either they were annotated as control wells by the original study or did not pass sequencing QC. Single cells that passed sequencing QC were defined as > 75% alignment rate, > 100,000 mapped reads, and < 20% of mitochondrial reads (Figure 2.1G-I). In total, 327 single cells consisting of 82, 85, 88, and 72 single cells at 0-, 24-, 48-, and 72-hrs, respectively, were included for downstream analyses.

The third dataset consists of iPSCs and endoderm cells differentiated *in vitro* from iPSCs (Linker et al., 2019). In total, 192 single cells were sequenced. Libraries were prepared using G&T-seq (Macaulay et al., 2016) and sequenced in 125bp PE mode. Fifty-six single cells were removed because they either were annotated as unknown cell types by the original study or did not pass sequencing QC. Single cells that passed sequencing QC were defined as having > 75% alignment rate, > 100,000 mapped reads, and < 20% of mitochondrial reads (Figure 2.1J-L). In total, 136 single cells consisting of 83 iPSCs and 53 endoderm cells were included for downstream analyses.

The fourth dataset consists of single cells derived from the spinal cord of mice induced with experimental autoimmune encephalomyelitis (EAE) and control mice (Falcao et al., 2018). In total, 2,208 single cells were sequenced. Libraries were prepared using Smart-seq2 protocol (Picelli et al., 2014) and sequenced in 50bp SE mode. One-hundred and forty-four single cells were removed because they did not pass sequencing QC. Single cells that passed sequencing QC were defined as having > 50% alignment rate, > 40,000 mapped reads, and < 55% of mitochondrial reads (Figure 2.1M-O). An additional 8 cells annotated as doublets by the original publication were removed. In total, 2,056 single cells consisting of 1,078 EAE and 978 control mice single cells were included for downstream analyses.

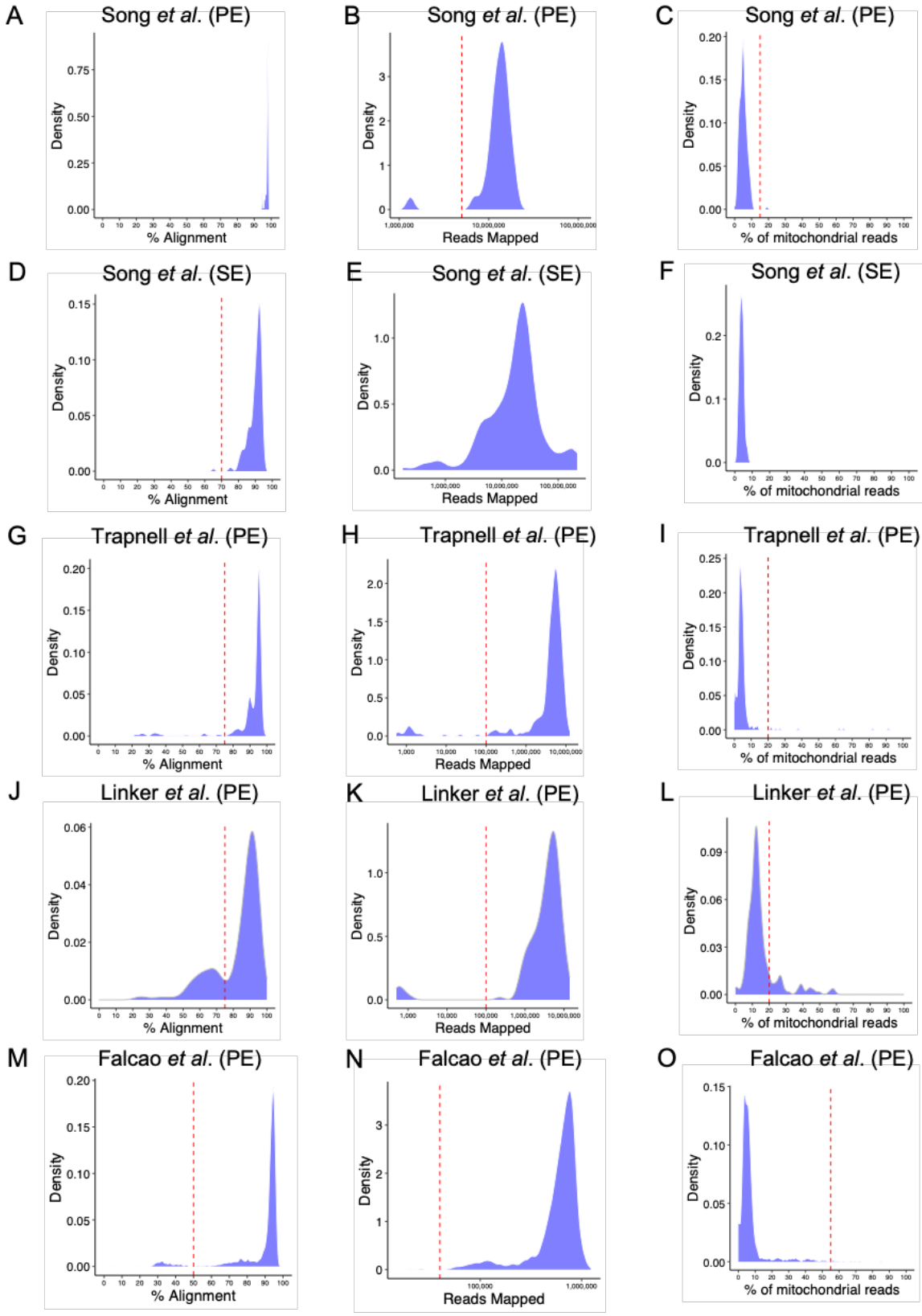

A
Song et al. (PE)
Density
% Alignment
B
Song et al. (PE)
Density
Reads Mapped
C
Song et al. (PE)
Density
% of mitochondrial reads
D
Song et al. (SE)
Density
% Alignment
E
Song et al. (SE)
Density
Reads Mapped
F
Song et al. (SE)
Density
% of mitochondrial reads
G
Trapnell et al. (PE)
Density
% Alignment
H
Trapnell et al. (PE)
Density
Reads Mapped
I
Trapnell et al. (PE)
Density
% of mitochondrial reads
J
Linker et al. (PE)
Density
% Alignment
K
Linker et al. (PE)
Density
Reads Mapped
L
Linker et al. (PE)
Density
% of mitochondrial reads
M
Falcao et al. (PE)
Density
% Alignment
N
Falcao et al. (PE)
Density
Reads Mapped
O
Falcao et al. (PE)
Density
% of mitochondrial reads

Figure 2.1. Sequencing QC metrices by alignment rate, number of mapped reads, and percentage of mitochondrial reads. Sequencing metrices for single cells from **(A-F)** Song *et al.* **(G-I)** Trapnell *et al.*, **(J-L)** Linker et al., and **(M-O)** Falcao *et al.*. Red dotted lines represent the thresholds used to stratify cells into those that passed or failed sequencing QC. For alignment rate and the number of reads mapped, cells to the right of the dotted lines were considered as cells passing sequencing QC. For the percentage of mitochondrial reads, cells to the left of the dotted lines were considered as cells passing sequencing QC. Only cells passing all three sequencing QC metrices were included for downstream analyses. PE: Paired-end; SE: Single-end.

2.1.2 Pre-processing of in-house datasets

ScRNA-seq generated in-house using plate-based methods was included in "Section 6 Application of developed computational pipelines" to demonstrate the application of our computational framework in myeloid neoplasm patients. Specifically, scRNA-seq of a myelodysplastic syndrome (MDS) patient with *SF3B1*^{K666} and *SF3B1*^{K626} variants was generated by Affaf Aliouat under the supervision of Sten Eirik Jacobsen, Eva Hellström-Lingberg, and Seshi Ogawa. ScRNA-seq of myeloproliferative neoplasm (MPN)_patients with *U2AF1*^{S34} or *U2AF1*^{Q157} variants was generated by Alba Rodriguez-Meira under the supervision of Adam Mead. ScRNA-seq was pre-processed as described in "Section 2.1.1 Pre-processing of publicly available datasets".

2.1.3 Gene expression quantification

Gene expression in transcript per million (TPM) for each sample was quantified using RSEM with default settings (B. Li & Dewey, 2011).

2.1.4 Isoform detection and quantification

For cell types without matched bulk samples, single-cell BAM files belonging to the same cell type were merged using Samtools (version 1.9) to create pseudo-bulk BAM files. For each dataset, known and novel isoforms were subsequently detected in the corresponding bulk BAM files using StringTie (version 2.1.4). The detected isoforms from each bulk BAM file were merged to create a dataset-specific Gene

Transfer File (GTF) (Pertea et al., 2015). Therefore, this dataset-specific GTF is a catalogue of all known and novel isoforms detected in a given dataset.

Next, custom R scripts were used to format the StringTie-generated GTF to match the standard GENCODE GTF. This ensures that the GTF can be recognised by downstream third-party software.

It is noteworthy that the *de novo* detection of isoforms and subsequent generation of dataset-specific GTF was only feasible for relatively longer sequencing reads such as >=100bp in PE mode (Linker et al., 2019; Song et al., 2017; Trapnell et al., 2014). Relatively shorter sequencing reads such as 50bp in SE mode (Falcao et al., 2018) do not often straddle across exon-exon junctions and also have shorter overhangs, and hence less powered for *de novo* isoform detection. Therefore, for RNA-seq datasets with relatively shorter sequencing reads, we downloaded publicly available GENCODE GTF for downstream alternative splicing detection.

Next, alternative splicing detection and quantification were performed using BRIE (Huang & Sanguinetti, 2021), Expedition (Song et al., 2017), and MARVEL.

For BRIE, the *briekit-event* module was used to detect alternative splicing event from the GTF, and the *briekit-event-filter* function was subsequently used to filter for high-quality skipped-exon (SE) splicing events in lenient mode [*--add_chrom chrX --as_exon_min 10 --as_exon_max 100000000 --as_exon_tss 10 --as_exon_tts 10 --no_splice_site*]. Next, the number of reads supporting the constitutive and alternative exons were tabulated using the *brie-count* function. The percent spliced-in (PSI) values were computed using three modes (0, 1, and 2), using the *brie-quant* function. At low coverage where sequencing reads do not provide sufficient information to compute PSI values confidently, the different modes offer different strategies for imputing PSI values in this setting.

Mode 0 imputes PSI values at low coverage using a default setting of 50. Therefore mode 0 assumes that a given single cell expresses both included (alternative exon spliced in) and excluded (alternative exon spliced out) isoforms. Mode 1 imputes PSI values at low coverage using a Bayesian prediction approach from genomic sequence features. The correlation between genomic sequence features and PSI values was first determined using the *briekit-factor* function. When genomic sequence features are not informative for a given alternative splicing event, mode 1 imputes PSI values at low coverage using a default setting of 50. Lastly, mode 2 imputes PSI values at low coverage using the mean PSI across the cell population.

For Expedition, the *outrigger index* function was used to detect both SE and mutually exclusive exons (MXE) splicing events directly from the splice junction files generated from the STAR aligner. Next, the PSI values for each alternative splicing event were computed using the *outrigger psi* function. For a given alternative splicing event in a given cell, the PSI value was re-coded as a missing value (NA) when the event was supported by less than 10 splice junction reads.

For MARVEL, a third-party software, rMATS (version 4.1.0) was used to detect SE, MXE, retained intron (RI), alternative 5' and 3' splice sites (A5SS and A3SS, respectively) (Shen et al., 2014). MARVEL was used to further detect alternative first and last exons (AFE and ALE) using the *DetectEvents* function. Prior to computing PSI values, MARVEL first ensures (validates) that each alternative splicing event is supported by at least 10 splice junction reads (Figure 2.2). This is to mitigate false positive alternative splicing events detected by StringTie and rMATS. Next, the PSI value for each validated alternative splicing event was computed using the splice junction read counts (Figure 2.3). For a given alternative splicing event in a given cell, the PSI value was re-coded as a missing value (NA) when the event was supported by less than 10 splice junction reads. Both validation of alternative splicing event and subsequent PSI computation were performed using the *ComputePSI* function.

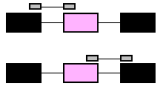
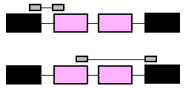
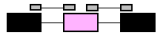
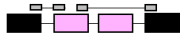
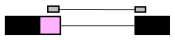
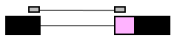
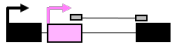
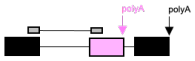
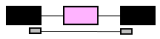
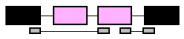
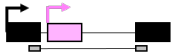
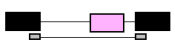
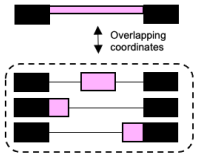
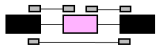
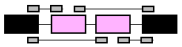
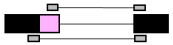
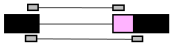
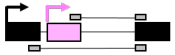
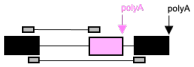
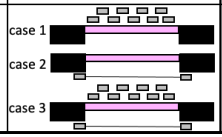
Description	Alternative splicing event type							Validated?
	SE	MXE	RI	A5SS	A3SS	AFE	ALE	
Only one of both $SJ_{included}$ found			N/A	N/A	N/A	N/A	N/A	No
Only $SJ_{included}$ found			N/A					No
Only $SJ_{excluded}$ found			N/A					No
Intron overlapping with exonic coordinates	N/A	N/A		N/A	N/A	N/A	N/A	No
Both $SJ_{included}$ and $SJ_{excluded}$ found			NA					Yes
$Reads_{included}$ or/and $SJ_{excluded}$ found	N/A	N/A		N/A	N/A	N/A	N/A	Yes

Figure 2.2: Validation of alternative splicing events using splice junction reads. A3SS: Alternative 3' splice site; A5SS: Alternative 5' splice site; AFE: Alternative first exon; ALE: Alternative last exon; MXE: Mutually exclusive exon; N/A: Non-applicable; RI: Retained intron; SE: Skipped-exon. SJ: Splice junction.

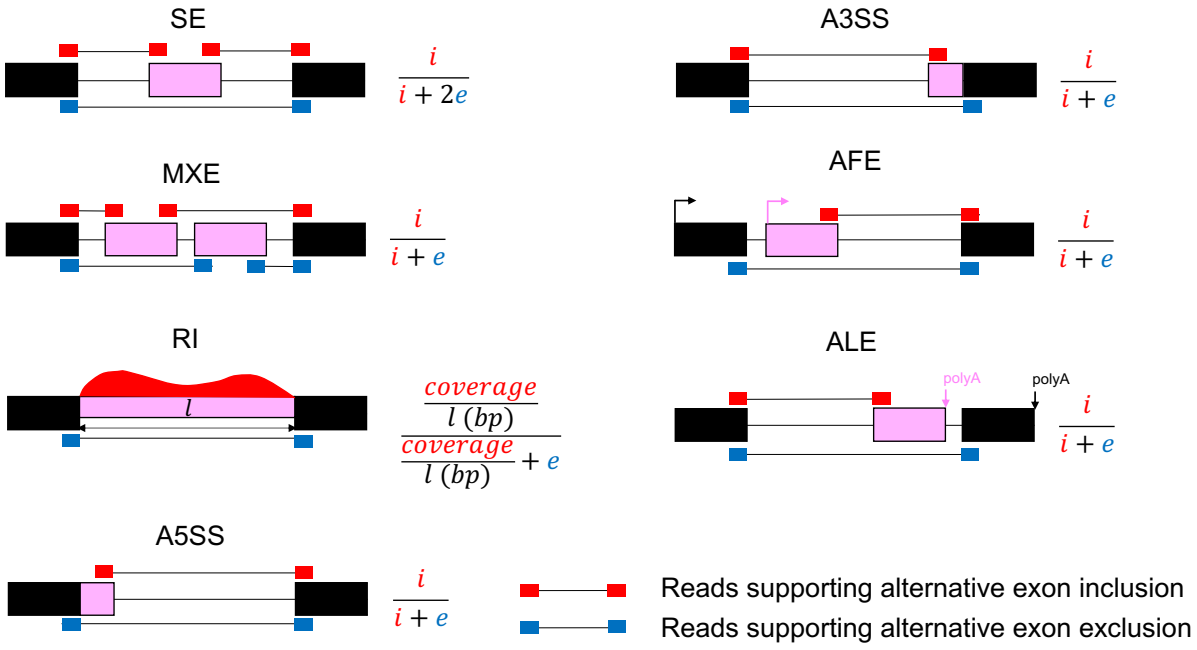


Figure 2.3: Formula for computing the PSI values for each alternative splicing event type. PSI values were computed as the total splice junction counts supporting the inclusion (spliced in) of the alternative exon divided by the total splice junction counts supporting the inclusion (spliced in) or exclusion (spliced out) of the alternative exon. The red splice junctions (*i*) represent the splice junctions supporting the inclusion (splicing in) of the alternative exon. The black splice junctions (*e*) represent the splice junctions supporting the exclusion (splicing out) of the alternative exon. For RI, coverage/read counts, in lieu of splice junction, supports the inclusion (splicing in) of the intron.

2.1.5 DNA sequence conservation score analysis

The DNA sequence conservation score of each alternative exon was computed using the *gscores* function from the GenomicScores package (Puigdevall & Castelo, 2018). Specifically, the conservation score of each alternative exon represented the average conservation score of each nucleotide that constituted the alternative exon. The DNA sequence conservation scores were based on the UCSC phastCons conservation scores for the human genome (GRCh38) calculated from multiple alignments with other 99 vertebrate species (Siepel et al., 2005). For each type of the seven alternative splicing events, the correlation between the conservation score of each alternative exon and percent spliced-in (PSI) values was assessed.

2.1.6 Modality assignment

Song *et al.* first proposed the concept of modality, which is the assignment of PSI distributions into discrete categories (Song et al., 2017). The five modalities introduced were included, excluded, bimodal, middle, and multimodal (Figures 2.4 and 1.8). An alternative splicing event may be assigned to one of these modalities, depending on its PSI distribution across a given cell population.

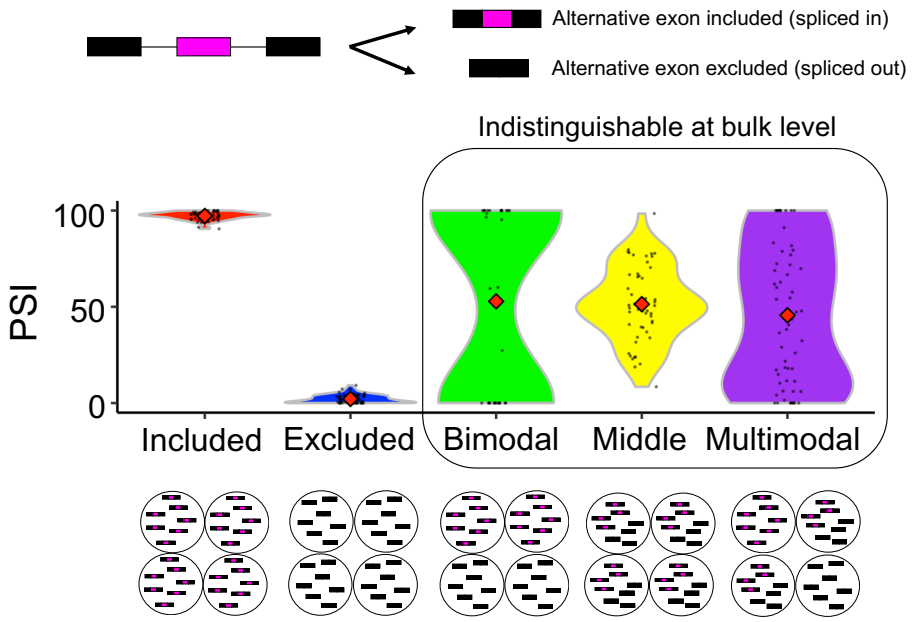


Figure 2.4: The five modalities originally proposed by Song *et al.* (Song et al., 2017). (See also Figure 1.8)

The PSI values for a given alternative splicing event in a given cell population can take any values between 0-100, therefore the PSI distribution may be modelled using a beta distribution. The shape of the beta distribution requires two parameters to be estimated, α and β. MARVEL used a maximum likelihood estimation to determine the α and β values of the PSI distribution. This is implemented using the *fitdist* function from the fitdistrplus R package (Delignette-Muller & Dutang, 2015). Based on the α and β values, MARVEL sequentially assigned the distribution of the alternative splicing event into one of the modalities as follows. "Sequentially" here means that PSI distribution that did not meet criteria (a) will be assessed for criteria (b), and so on.

a. Bimodal : $\alpha < 0.5$ & $\beta < 0.5$

b. Included : $\alpha > 2.0$ & $\beta < 1$ OR $\alpha:\beta > 2$

c. Excluded : $\beta > 2.0$ & $\alpha < 1$ OR $\beta:\alpha > 2$

d. Middle : $\alpha > 1.5$ & $\beta > 1.5$

Any PSI distributions that do not meet any of the above criteria will be assigned as multimodal. The following figures illustrate the variations of PSI distributions for a given modality based on a range of α and β values (Figure 2.5).

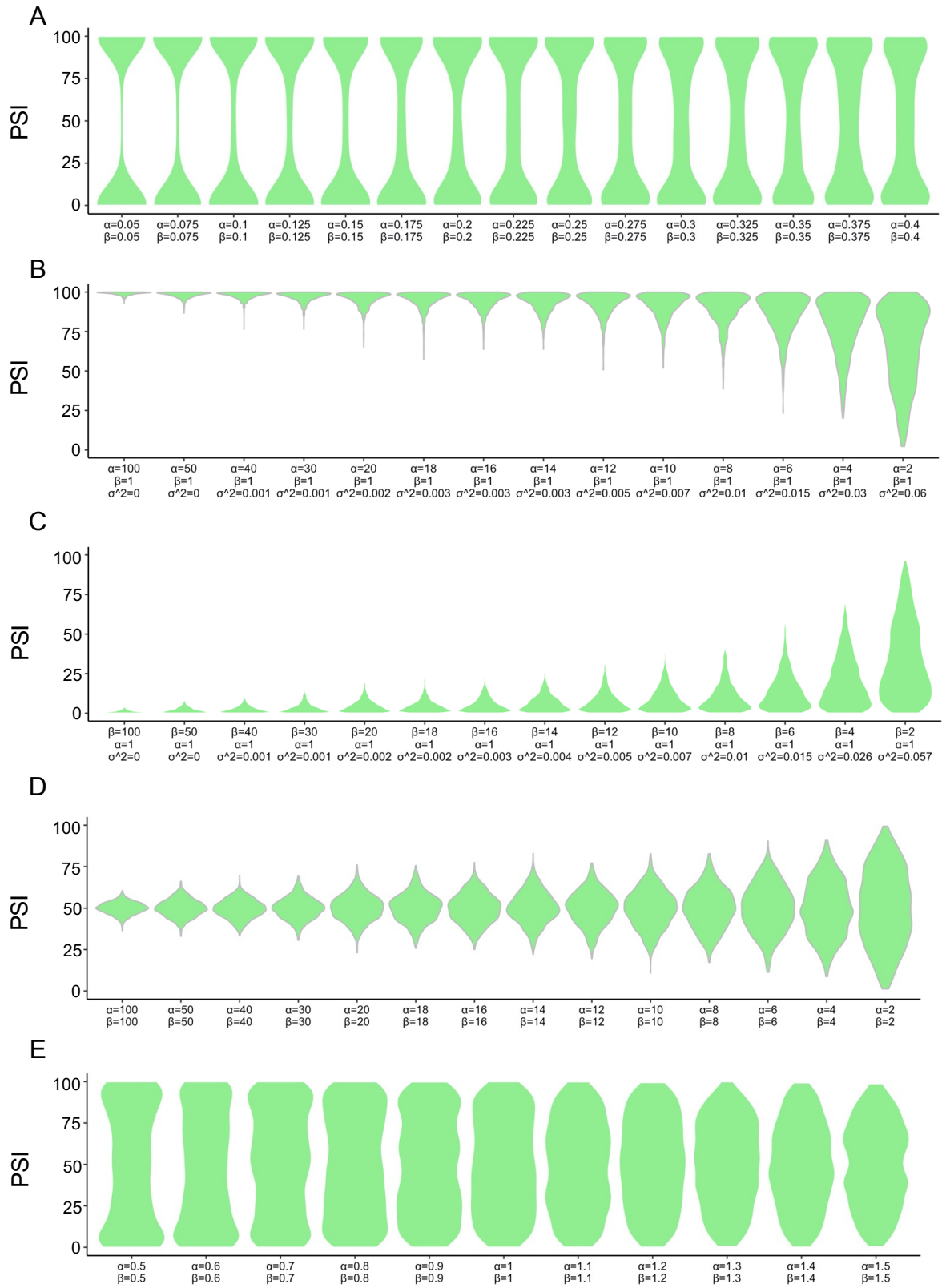


Figure 2.5: The PSI distributions for each modality based on a range of α and β values modelled using the beta distribution. PSI distributions for **(A)** bimodal, **(B)** included, **(C)** excluded, **(D)** middle, and **(E)** multimodal.

For included and excluded modalities, we noted that the PSI distributions with longer tails were associated with larger variance (Figure 2.5B-C). Therefore, we applied a heuristic threshold of 0.001 for variance (σ^2) to further stratify the included and excluded modalities into primary and dispersed (Figure 2.6). The included and excluded primary modalities have smaller σ^2 (< 0.001), whereas the included and excluded dispersed modalities have larger σ^2 (> 0.001). Therefore, the PSI values of the primary modalities clustered more tightly together, whereas the PSI values of dispersed modalities were more spread out (dispersed).

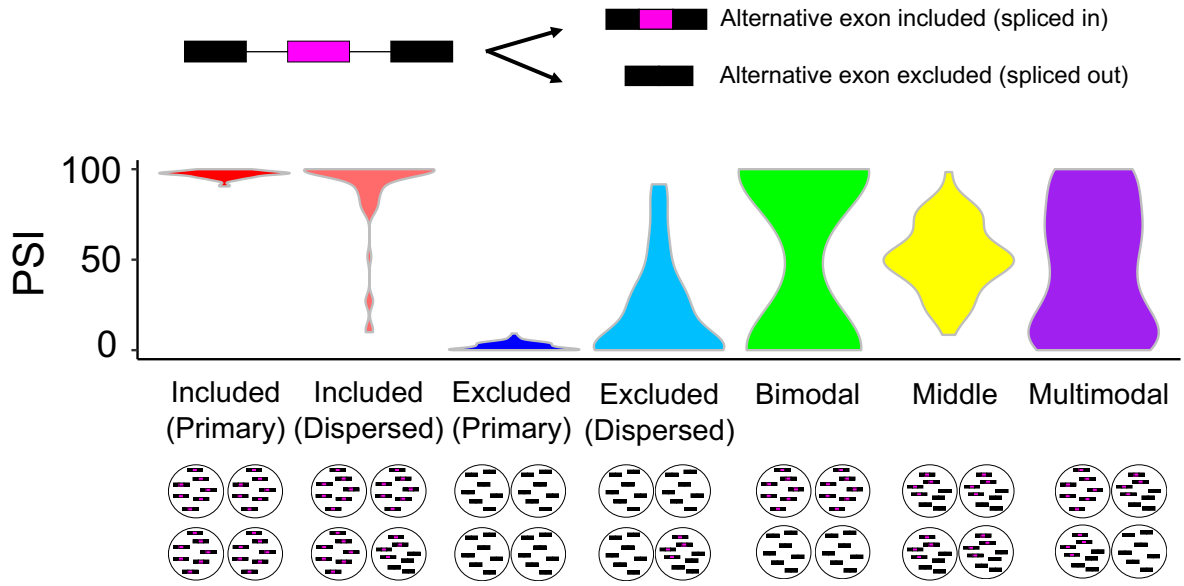


Figure 2.6: The further stratification of included and excluded modalities into primary and dispersed, based on the variance (spreading out) of the PSI distributions.

MARVEL assigned the PSI distribution of each alternative splicing event into one of the seven modalities using the *AssignModality* function.

2.1.7 Bimodal modality adjustment

A recent study showed that a significant proportion of bimodal PSI distributions were due to single-cell library preparation artifacts from amplification bias of minor isoforms. This amplification bias may result in false bimodal distributions. (Figure 1.9)

(Buen Abad Najar et al., 2020). Notably, the current modality assignment algorithm from Expedition did not consider these artifacts when assigning modalities (Song et al., 2017). Indeed, a disproportionate number of alternative splicing events were assigned as bimodal by Expedition (~10-20%) compared to just ~1% when alternative splicing events when only highly expressed genes were included for modality assignment. Including only highly expressed genes were shown to mitigate the false bimodal classification (Buen Abad Najar et al., 2020). However, the latter approach would preclude low-to-moderately expressed genes with relevant biological functions from alternative splicing analysis.

We first sought to identify any distinguishing features between true and false bimodal PSI distributions. To this end, we tabulated a catalogue of alternative splicing events with true or false bimodal PSI distributions. Alternative splicing events with true bimodal PSI distributions were retrieved from RNA-seq data previously validated using qPCR or smFISH (Song et al., 2017). Additional true bimodal PSI distributions were retrieved from modality assignment of highly expressed alternative splicing events as previously described (Buen Abad Najar et al., 2020). These highly expressed alternative splicing events were defined as events whose corresponding genes had inferred mRNA counts of at least 10 in at least 25 cells. mRNA counts were inferred using Monocle3 (Trapnell et al., 2014). Furthermore, alternative splicing events with false bimodal PSI distributions were retrieved from RNA-seq data previously validated using qPCR (Song et al., 2017). In total, 45 and 7 alternative splicing events with true and false bimodal PSI distributions, respectively, were included for identifying distinguishing features between true and false bimodal PSI distributions.

Furthermore, 17,252 highly expressed alternative splicing events with true non-bimodal (included, excluded, middle or multimodal) PSI distributions were retrieved as previously described (Buen Abad Najar et al., 2020). Altogether, 45 bimodal and 17,259 non-bimodal (17,252 from highly expressed alternative splicing events and 7 from qPCR validation) alternative splicing events were included to assess the sensitivity, specificity, negative predictive value, and precision of bimodal/non-bimodal assignment by Expedition and MARVEL.

2.1.8 Differential splicing analysis

Differential expression analysis is the cornerstone of both single-cell and bulk RNA-seq experiments. scRNA-seq analysis software, such as Seurat, use Wilcoxon rank-sum test as the default statistical test for differential gene expression between two groups of single cells (Satija et al., 2015). This approach detects differences in the average gene expression values between two groups of single cells but may not be suitable for detecting differences in percent spliced-in (PSI) values between two groups of single cells.

For example, alternative splicing events with bimodal, middle, and multimodal PSI distributions have average PSI values of ~50, and therefore it would not be straightforward to distinguish these three types of PSI distributions from the average PSI values alone (Figure 2.6). Furthermore, PSI distributions with large variance, such as included dispersed, excluded dispersed, bimodal, and multimodal, may decrease statistical power to detect any differences in average PSI values between two groups of single cells. Therefore, current statistical approaches for differential gene expression analysis may not directly apply to differential splicing analysis.

Instead of assessing the differences in average PSI values between two groups of single cells, MARVEL implemented statistical tests to assess the differences in overall PSI distribution between two groups of single cells. For comparing mean PSI values, MARVEL implemented Wilcoxon rank-sum test and t-test. For comparing overall PSI distributions, MARVEL implemented Kolmogorov-Smirnov test, Anderson-Darling test, and D Test Statistics (DTS) (Dowd, 2020).

To perform differential gene and splicing analysis, users may use the *CompareValues* function by MARVEL. Users may subsequently plot the differential analysis results, such as in the form of a volcano plot, by using the *PlotDEValues* function. Furthermore, to characterise the modality changes of alternative splicing events between two groups of single cells, users may use the *ModalityChange* function. Lastly, to simultaneous explore the changes in PSI values of alternative splicing events relative to their corresponding changes in gene expression values between two groups of single cells, user may use the *IsoSwitch* function.

2.1.9 Modality dynamics

After detection of differentially spliced events, MARVEL is able to classify the modality change of a given differentially spliced event between two cell populations. To this end, MARVEL leverages on its bimodal-adjusted modality assignment algorithm above. For a given differentially spliced event, MARVEL stratifies the modality change from one population to the next into one of three categories, namely explicit, implicit, or restricted. Explicit modality change involves the main modalities (included, excluded, bimodal, middle, and multimodal), e.g., from included to bimodal. Implicit modality change involves primary or dispersed modalities, e.g., from included primary to included dispersed. Restricted modality change refers to two cell populations having the same modality notwithstanding having significantly different PSI distributions. Users may categorise the modality change of differentially spliced events using the *ModalityChange* function.

2.1.10 Gene-splicing relationships

After detection of differentially spliced events, MARVEL is also able to classify the relationship between a given differentially spliced gene and its corresponding splicing event(s) into one of four categories, namely coordinated, opposing, isoform switching, and complex. Coordinate relationships indicate the change in mean gene expression value is in the same direction as the change in PSI value between two cell populations. On the other hand, opposing relationships indicate the change in mean gene expression value is in the opposite direction to the change in PSI value between two cell populations. Isoform switching indicates that the gene is not differentially expressed, but the splicing event is significantly spliced, between two cell populations. Lastly, a complex relationship constitutes a combination of coordinated, opposing, and/or isoform switching. For example, a given gene may be more highly expressed in cell population A relative to B, and its corresponding splicing event no. 1 is also more highly spliced in among cell population A relative to B, but splicing event no. 2 is more highly spliced in among cell population B relative to A. Users may perform gene-splicing relationship analysis using the *IsoSwitch* function.

2.1.11 Pathway enrichment analysis

MARVEL may perform pathway enrichment analysis to assess whether the differential spliced genes are functionally related and hence belong to the similar

biological pathways. It is conceivable that genes that are functionally related may be co-ordinatedly regulated, and as a consequence differentially spliced, between two groups of single cells. Therefore, pathway enrichment analysis is useful in assessing whether MARVEL detects biological relevant differentially spliced genes (Huang & Sanguinetti, 2017).

The pathway enrichment analysis and subsequent collapsing of redundant pathways were implemented using the *enrichGO* and *simplify* functions, respectively, from the clusterProfiler package (T. Wu et al., 2021). Users may use the *BioPathways* function for pathway enrichment analysis and subsequently use the *BioPathways.Plot* function to plot the top or user-specified pathways. MARVEL currently supports pathway enrichment analysis for mouse and human genomes.

2.1.12 Nonsense-mediated decay prediction

To determine the functional consequence of differentially spliced genes, MARVEL predicts whether the inclusion (spliced in) of alternative exons or introns leads to the introduction of premature stop codons (PTCs) and consequently downstream nonsense-mediate decay (NMD) of the isoforms.

For a given differentially spliced gene, MARVEL first retrieves all protein-coding isoforms from the GTF file, and further subsets isoforms where the alternative exon or intron identified from differential splicing analysis is located within the open reading frame (ORF). Next, the alternative exon or intron is inserted into the isoform and the resulting transcribed mRNA sequence is retrieved using the *getSeq* function implemented by the Biostrings package (Pagès, 2021) (Figure 2.7). Then, the mRNA sequence is translated into amino acid sequence using *translate* function implemented by the same package. The distance of the PTC, if any, relative to the final splice junction is then noted. Isoforms with PTC > 50bp away from the final splice junction are considered to be subjected to NMD. A differentially spliced gene is considered to be subjected to NMD if at least one isoform was found to be subjected to NMD.

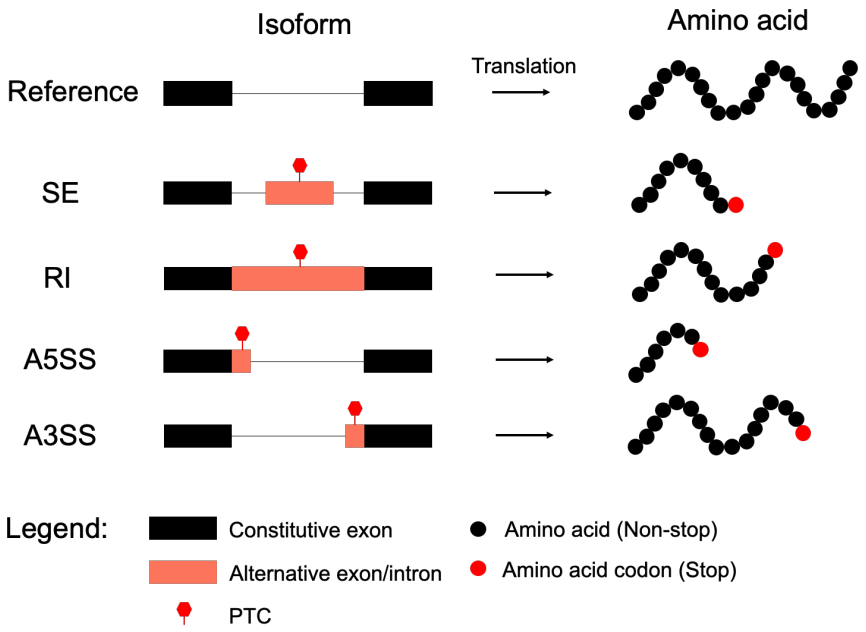


Figure 2.7: The *in silico* prediction of PTCs. The mRNA sequence without the alternative exon or intron (reference) is first retrieved, and the resulting amino acid sequence is retrieved after the inclusion (spliced in) of the alternative exon (SE/A5SS/A3SS) or intron (RI) into the mRNA sequence. PTC: Premature stop codon.

The introduction of PTCs and subsequent NMD can be predicted using the *FindPTC* and *CompareExpr* functions by MARVEL. The association between gene expression levels and the NMD status of genes may also be visualised using the *CompareExpr* function. Lastly, individual genes subjected to NMD may be visualised on the -log10(FDR) vs log2(FC) volcano plot generated from differential gene expression analysis using the *AnnoVolcanoPlot* by MARVEL to identify genes most affected (down-regulated) due to NMD. These genes may serve as candidate spliced genes for downstream experimental validation.

2.2 MARVEL analysis for droplet-based scRNA-seq datasets

2.2.1 Pre-processing of publicly available datasets

To demonstrate the utility of MARVEL on scRNA-seq data generated from the droplet-based library preparation method, we retrieved two RNA-seq datasets.

The first dataset was generated from induced pluripotent stem cells (iPSCs) and cardiomyocytes differentiated from iPSCs (Ou et al., 2021). This single-cell data

was prepared using Chromium Single Cell 3′ Reagent Kit (version 2) and sequenced in 150bp paired-end (PE) mode. In total, four libraries were prepared, consisting of iPSCs, and cardiomyocytes at days 2, 4, and 10 post-differentiation. The second dataset consists of brain tissues derived from autism spectrum disorder (ASD) patients and healthy control (Velmeshev et al., 2019). This single-cell data was prepared using Chromium Single Cell 3′ Reagent Kit (version 2) and sequenced in 100bp paired-end (PE) mode.

For the iPSC-cardio dataset, the FASTQ files were downloaded from the Sequence Read Archive (SRA) and aligned to the human reference genome (GRCh38) using Cell Ranger (version 2.1.1). For the brain tissue dataset, the FASTQ files were aligned to the GRCh38 reference genome using Cell Ranger v7.0.0 with the *include-introns true* option because nuclei mRNAs (the starting material in this study) contain higher proportion of unspliced intronic reads compared to cytoplasmic mRNAs. The resulting BAM files were shuffled using Samtools (version 1.9) prior to re-alignment using STARsolo (version 2.7.8a) (Kaminow et al., 2021). The resulting unique molecular identifier (UMI)-collapsed gene and splice junction count files were used for downstream quality control and analyses.

For the iPSC-cardio dataset, SingCellaR was used for identifying high-quality cells and subsequent integration of single cells from the different libraries (G. Wang et al., 2022). For each library, high-quality cells were identified based on the number of UMIs and genes detected per cell using the *plot_UMIs_vs_Detected_genes* function (Figure 2.8). High-quality cells were defined and retained using the *filter_cells_and_genes* function. Next, the UMI values of the remaining cells were normalised against their respective total counts (library size) and multiplied (scaled) by 10,000 using the *normalize_UMIs* function. Potential confounding factors that may skew the gene expression values, e.g., the total UMI counts and percentage of mitochondrial counts, were adjusted for using the *remove_unwanted_confounders* function. Next, highly variable genes were identified using the *get_variable_genes_by_fitting_GLM_model* function. Each pre-processed library was saved as a separated R object. The individual pre-processed libraries for iPSCs and day-10 cardiomyocytes were integrated into a single R object using the *preprocess_integration* function. Linear dimension reduction using principal component analysis (PCA) was performed using the *runPCA* function. Next, the first 10 principal components (PCs) were used for further non-linear dimension reduction

using the *runTSNE* function. The normalised gene expression matrix and tSNE embeddings by SingCellaR, raw gene and splice junction counts by STARsolo, and GTF were provided as the inputs for MARVEL. In total, 11,244 iPSCs and 6,240, 8,635, and 5,937 cardiomyocytes at day-2, -4, and -10, respectively, were included for analysis.

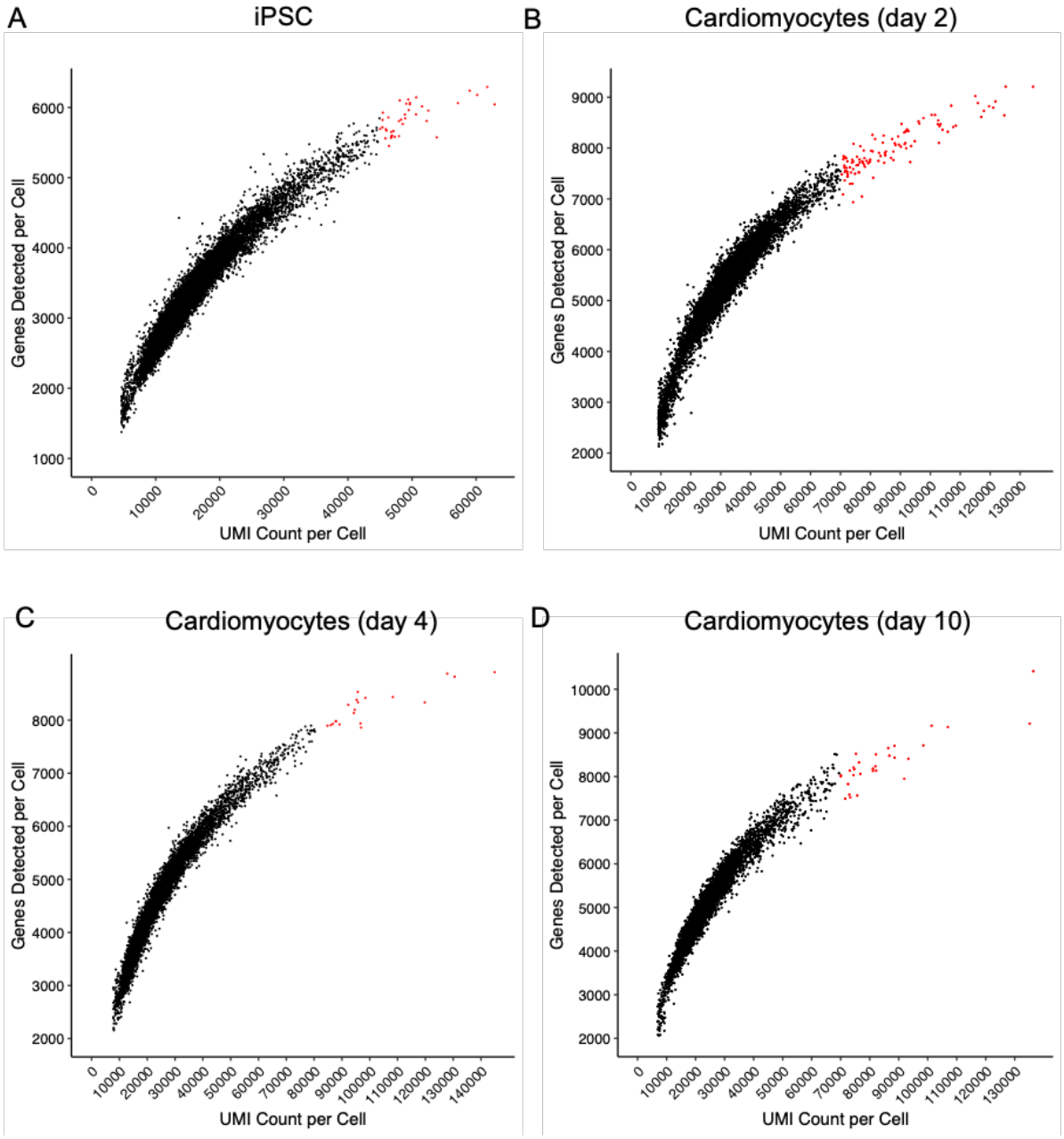


Figure 2.8: Quality control (QC) of single cells using total UMI and genes detected. Red points denote cells that failed QC and were excluded. Black points denote cells that passed QC and were included in downstream analyses.

For the brain tissue dataset, all 104,559 cells reported by Supplementary table 2 of the original study were retrieved and included for analysis. The tSNE coordinates were similarly retrieved from Supplementary table 2 of the original study. The published tSNE coordinates, raw gene and splice junction counts by STARsolo, and GTF were provided as the inputs for MARVEL.

Further pre-processing was performed using MARVEL. First, each gene was annotated with its corresponding gene type, e.g., protein-coding, pseudogene etc., using the *AnnotateGenes.10x* function. Next, the exons that constituted each splice junction were annotated using the *AnnotateSJ.10x* function as either (1) both exons consisting of known exons as per reported in the GTF file or (2) only one or none of the exons consisting of known exons. Each exon that constituted the splice junctions was also annotated as either (1) uniquely mapped to a single exon or (2) mapping to multiple exons. Only splice junctions whose both exons were identified as uniquely mapped and reported in the GTF ("known exons") were considered high-quality splice junctions and were retained for downstream analyses using the *ValidateSJ.10x* function. Users also have the option to include splice junctions whose one end mapped to a known exon and another end mapped to a novel exon for downstream analysis by specifying *keep.novel.sj=FALSE* option in the *ValidateSJ.10x* function. Then, only protein-coding genes and splice junctions of protein-coding genes were retained using the *FilterGenes.10x* function. Finally, only cells (barcodes) overlapping in both gene and splice junction data were retained using the *CheckAlignment.10x* function.

2.2.2 Pre-processing of in-house datasets

ScRNA-seq generated in-house using droplet-based methods was included in "Section 6 Application of developed computational pipelines" to demonstrate the application of our computational framework in myeloid neoplasm patients. Specifically, scRNA-seq of myelodysplastic syndrome (MDS) patients with *SRSF2*^{P95} variant was generated by Juseong Lee under the supervision of Andrea Pellagatti and Jacqueline Boultwood. ScRNA-seq was pre-processed as described in "Section 2.2.1 Pre-processing of publicly available datasets".

2.2.3 Splice junction usage quantification

Due to the sparsity of read counts, compounded by perverse 3'/5' coverage bias (Figure 1.10), of RNA-seq data generated from droplet-based library preparation methods, isoform analysis was performed at the splice junction-level in lieu of exon-level. For the same reasons, the percent spliced-in (PSI) of a given splice junction will be computed at the cell type (population) level, in lieu of at the single-cell level. Therefore, the PSI value of a given splice junction is the total splice junction counts divided by the total gene counts across all single cells of a given cell type (Kaminow et al., 2021). The mathematical formula is described as:

$$PSI_{sj,cell\ type} = \frac{\sum_{cell \in cell\ type} Count_{sj,cell}}{\sum_{cell \in cell\ type} Count_{gene,cell}}$$

The numerator of the equation indicates the total splice junction counts of a given splice junction across all cells in a given cell population. The denominator indicates the corresponding total gene counts across all cells in the same cell population.

The cell type will have to be defined *a priori* by the user. The cell type may be identified by its cellular phenotype (e.g., induced pluripotent stem cells (iPSCs) vs cardiomyocytes) (Ou et al., 2021), or using cell surface markers (e.g., CD34+ for haematopoietic stem and progenitor cells), or gene expression signatures (Roy et al., 2021).

2.2.4 Differential splicing analysis

MARVEL adopts a permutation-based approach for assessing differentially spliced junctions between two cell populations (Efremova, Vento-Tormo, Teichmann, & Vento-Tormo, 2020).

For a given splice junction, the cell type labels of the single cells constituting the two cell types are shuffled (permutated) (Figure 2.9). Next, PSI values for each of the two simulated cell populations are computed. The difference in the PSI values between the two simulated cell populations is then noted ($\Delta PSI_{permutated}$). This is iterated 1,000 times, and these values will form the null distribution.

Then, the observed PSI values for each of the two cell populations are then computed and the difference in the PSI values between the cell populations is then

noted ($\Delta PSI_{observed}$). The $\Delta PSI_{observed}$ is compared against the null distribution to obtain the *P* value. Specifically, the proportion of $|\Delta PSI_{permutated}| > |\Delta PSI_{observed}|$ will constitute the *P* value. (Figure 2.10). For example, if 50 out of 1,000 of $|\Delta PSI_{permutated}|$ constituting the null distribution are larger than $|\Delta PSI_{observed}|$, then the *P* value is 0.05.

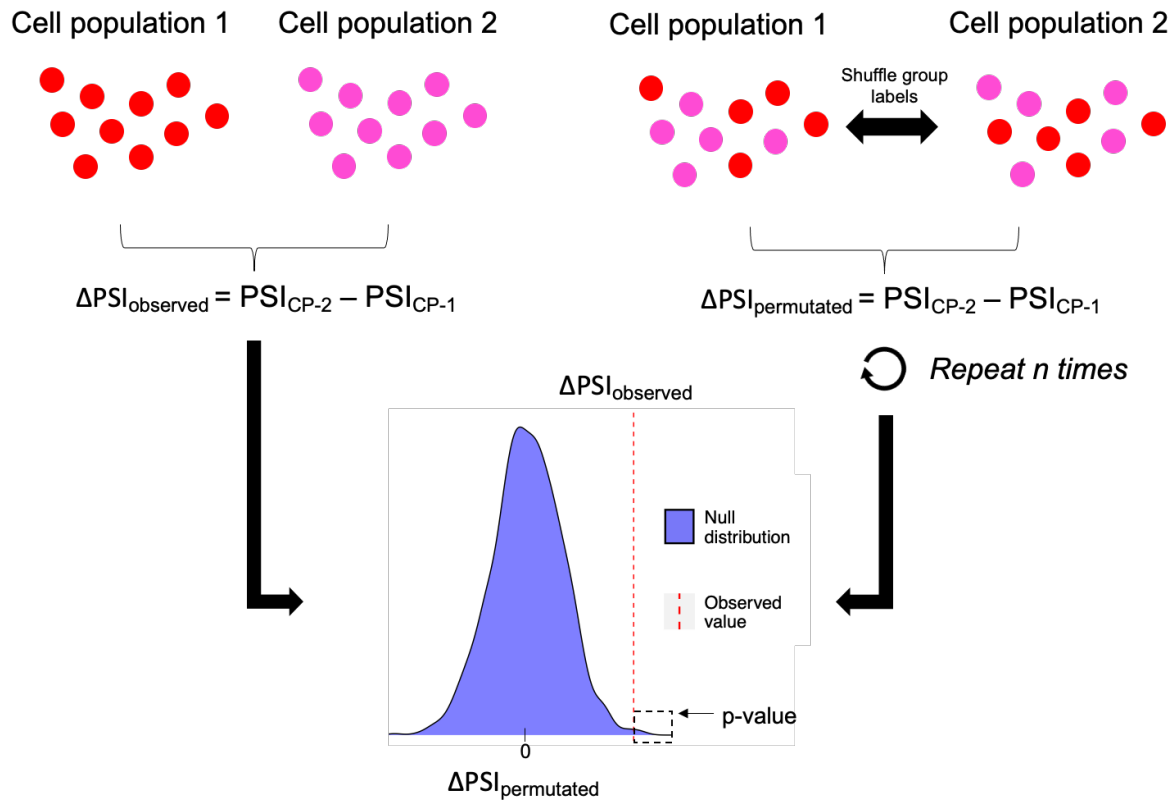


Figure 2.9: Permutation-based approach for assessing differences in PSI values between two cell populations, e.g., health vs disease states. The permutated difference in PSI values is computed several times to construct the null distribution and then the observed difference in PSI values is mapped to the null distribution to obtain the *P* value. CP: Cell population.

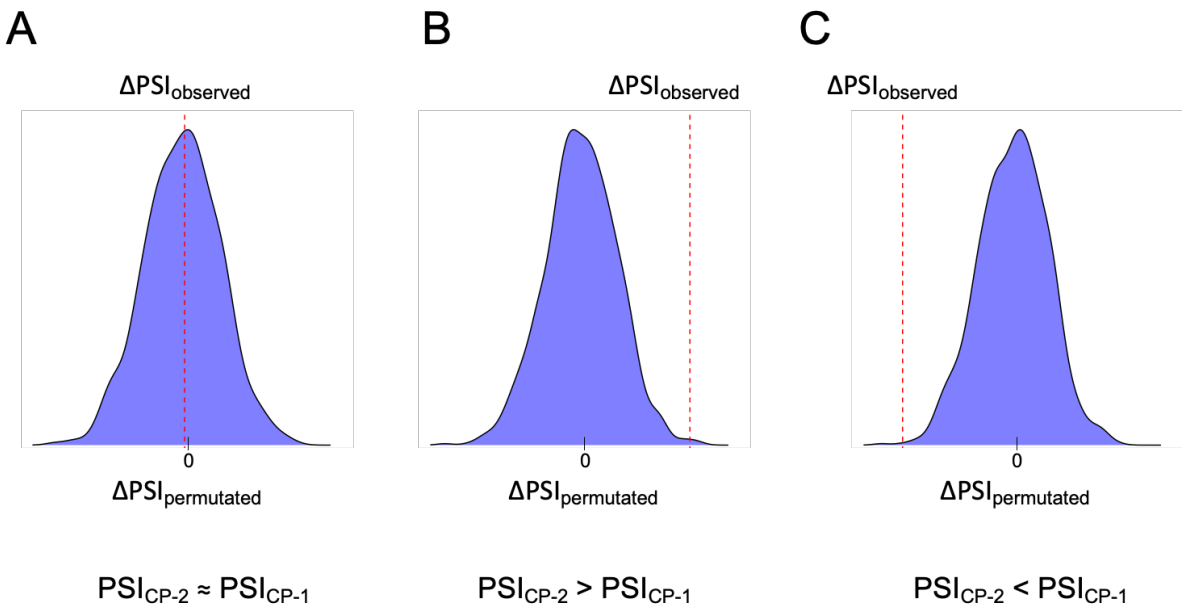


Figure 2.10: Three possible scenarios of the permutation-based approach. (A) No statistical difference between the PSI value in cell population 1 vs. cell population 2. **(B)** PSI value in cell population 2 is significantly higher ($P < 0.05$) than in cell population 1. **(C)** PSI value in cell population 2 is significantly smaller ($P < 0.05$) than in cell population 1. CP: Cell population.

2.2.5 Differential gene expression analysis

MARVEL employs Wilcoxon rank-sum test as the default statistical test for differential gene expression analysis between two cell populations, as recommended by the Seurat tutorial (Satija et al., 2015). MARVEL also provides the MAST statistical framework to allow for users to adjust for selected co-variates, such as sequencing batch and donor ID, during for differential gene expression analysis (Finak et al., 2015). When MAST is selected, MARVEL automatically computes the number of genes detected per cell (gene detection rate) and include this co-variate into the zero-inflated regression model. This is because gene detected rate is recommended as an important co-variate by the MAST tutorial. To identify differentially expressed genes, a likelihood ratio test (LRT) is performed to compare the model with and without the cell group information. Users may call the *CompareValues.Exp.Spliced* and *CompareValues.Genes.10x* functions to perform differential gene expression analysis for plate- and droplet-based sequencing data, respectively.

2.3 Visualisation of alternative splicing events using VALERIE

2.3.1 Estimating single-cell PSI values

VALERIE stands for **V**isualising **AL**ternative splicing **E**vents from single-cell **RI**bonucleic acid-sequencing **E**xperiments. The main goal of MARVEL is to perform visual validation of differentially spliced exons in order to classify a given splicing event as true or false positive. To this end, the main function of VALERIE is to visualise alternative splicing events across different cell populations using sequencing reads. This will allow users to validate differentially spliced events identified by MARVEL described in the previous section. VALERIE is implemented as an R package. It is currently hosted on GitHub: https://github.com/wenweixiong/VALERIE.

Using a single function (*PlotPSI*), users may plot the sequencing reads profile of the alternative splicing event of interest across different cell populations. The *PlotPSI* function is a wrapper for plotting the different alternative splicing event types, including SE, MXE, RI, A5SS, and A3SS, by executing the *PlotPSI.SE.Pos*, *PlotPSI.SE.Neg*, *PlotPSI.MXE.Pos*, *PlotPSI.MXE.Neg*, *PlotPSI.RI.Pos*, *PlotPSI.RI.Neg*, *PlotPSI.A5SS.Pos*, *PlotPSI.A5SS.Neg*, *PlotPSI.A3SS.Pos*, and *PlotPSI.A3SS.Neg* functions. The **Pos* and **Neg* functions are for plotting alternative splicing event types located on the positive or negative strand of the genome. The framework of plotting functions for each alternative splicing event type are similar and are elaborated below.

The genomic coordinates of the constitutive exons are first retrieved. Because some exons are very long relative to the alternative exon, this may lead to an under-emphasis of the alternative exon in the final figure, i.e., the alternative exon appears very small compared to the constitutive exons. Therefore, we provided users the option to specify the maximum length of the constitutive exons to be shown on the final figure. Any base pairs exceeding the maximum length are trimmed off (censored) in the final figure. This upper limit may be specified using the *cons.exon.cutoff* option.

The BAM files of single cells will be read into R using the *readGAlignments* function implemented by the GenomicAlignments package (Lawrence et al., 2013). To enable quick reading of the BAM files, only sequencing reads corresponding to the genomic coordinates of the alternative splicing event will be read into R. This is achieved using the *ScanBamParam* option of the *readGAlignments* function.

Then, the per-base percent spliced-in (PSI) values are computed using the sequencing reads retrieved from the BAM files. First, the number of overlapping reads (spliced in) across a given base is computed using the *coverage* function. Second, the number of overlapping reads (spliced in) across a given base and the number of skipped reads (spliced out) across the base are summed up as the total coverage using the *granges* function implemented by the GenomicRanges package (Lawrence et al., 2013), followed by the *coverage* function (Figure 2.11). This distinction is not often explicitly made clear by others when computing coverage, but nevertheless, it is this distinction that differentiates coverage contributed by only overlapping reads (spliced in) from coverage contributed by both overlapping (spliced in) and skipped reads (spliced out).

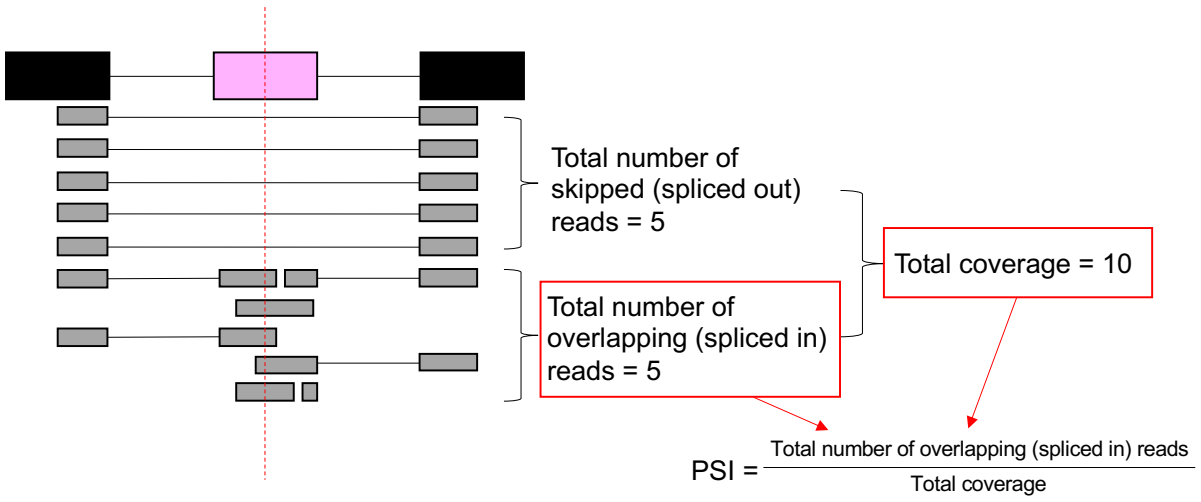


Figure 2.11: The distinction between overlapping (spliced in) and skipped (spliced out) reads when computing coverage for PSI estimation. The figure shows an example of computing PSI value for a given base position (red dashed line). The black exons represent the constitutive exons while the pink exon represents the alternative exon. The grey bars represent the sequencing reads. The black solid lines between the exons represent the intronic regions while the black solid lines between the sequencing reads represent the split region within the sequencing reads, i.e., the region not expressed (spliced out) by the cell. The latter is represented as N CIGAR string in the BAM files and are recognised as “skipped” regions by software such as the GenomicRanges package.

The per-base PSI value is then computed as the total number of overlapping reads divided by the total coverage. If the total coverage is below a user-defined threshold, then the PSI value will be re-coded as missing, i.e., NA. This can be specified by the user using the *min.coverage* option.

2.3.2 Plotting single-cell PSI values

The final figure is ready to be generated using the per-base PSI values computed. The final figure consists of three panels (Figure 2.12), showing the sequencing reads profile for *PKM* exon 8-9 mutually exclusive exons (MXE) in induced pluripotent stem cells (iPSCs), neural progenitor cells (NPCs), and motor neurons (MNs) (Song et al., 2017). This splicing event has been validated using single-molecule fluorescence *in situ* hybridization (smFISH) by the original publication.

The top panel demonstrates the per-base PSI values inferred from the sequencing reads profile. The columns represent the bases whereby the black intervals represent the constitutive exons, whereas the brown intervals represent the alternative exons. Each row represents a cell, and the cells are grouped according to the user-defined cell populations. The values within the heatmap represent the scaled PSI values across the column. The yellow-white-blue gradient intensity represents the scaled PSI values across the columns. Grey colour represents base position with low coverage, i.e., coverage below minimum coverage defined by the user. The heatmap was generated using the *pheatmap* function implemented by the heatmap package.

The middle panel shows the aggregated (mean) PSI values for each cell population. We can infer that there is a decrease in exon 9 (the 3' alternative exon) usage when iPSCs differentiated into either NPCs or MNs. Conversely, there is an increased in exon 8 (the 5' alternative exon) usage when iPSCs differentiated into either NPCs or MNs. This panel is generated by the ggplot2 package.

The bottom panel shows the -log10 of adjusted p-values derived from assessing the differences in PSI values for each base across the cell populations. We can infer that PSI values are significantly different at the intervals corresponding to the alternative exons. This suggests differential usage of alternative exons during iPSCs differentiation into NPCs and MNs. On the other hand, the PSI values are not significantly different at the intervals corresponding to the constitutive exons. This suggests no differential usage of constitutive exons during iPSCs differentiation into

NPCs and MNs. This re-affirms the status of the constitutive exons. This panel was also generated by the ggplot2 package.

The statistical tests provided by VALERIE for assessing the per-base difference in PSI values across two cell populations include Wilcoxon rank-sum test, t-test, Anderson-Darling test, and D Test Statistics (Dowd, 2020). Additionally, the statistical tests provided by VALERIE for assessing the per-base difference in PSI values across more than two cell populations include Kruskal-Wallis test and Analysis of Variance (ANOVA). The type of statistical test to use may be specified using the *method* option. Lastly, the method for adjusting the p-values for multiple testing may be specified using *method.adj* function.

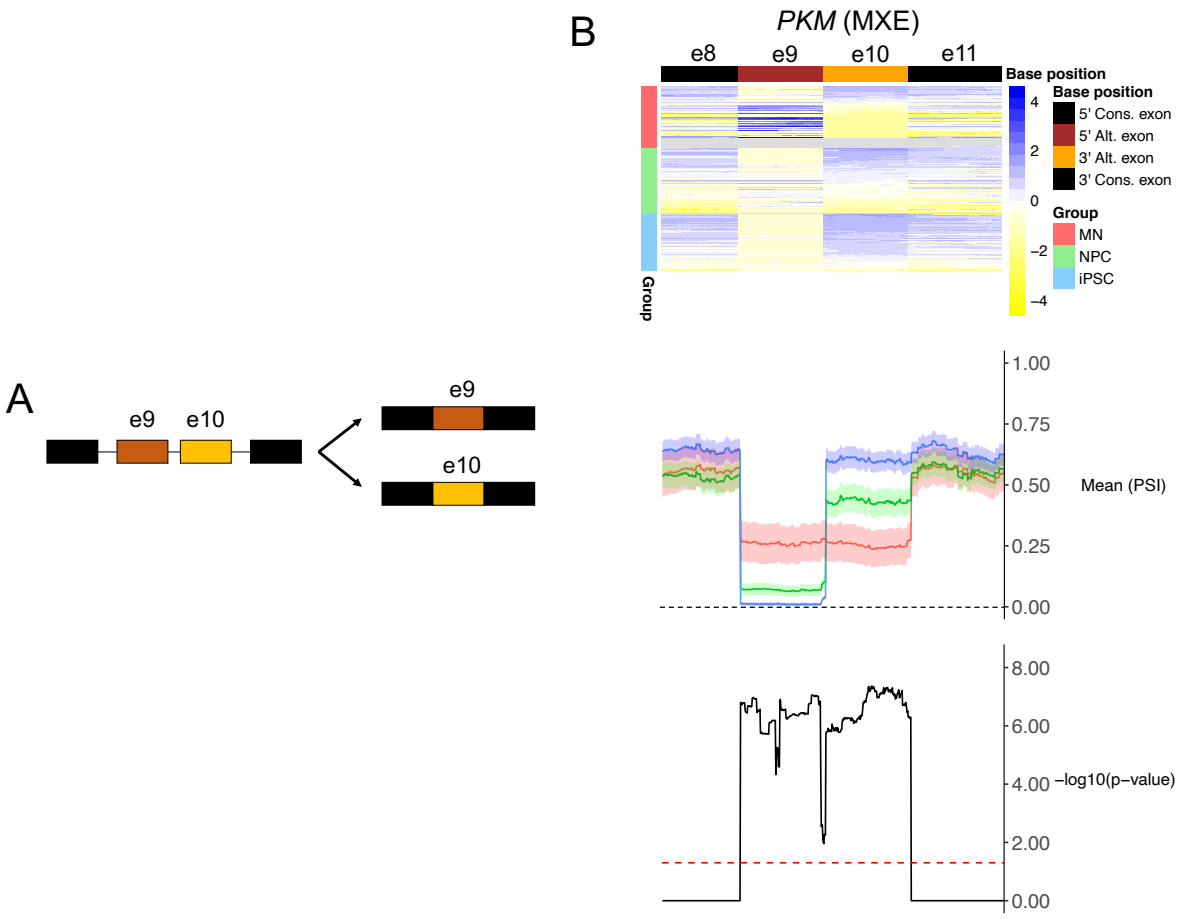


Figure 2.12: Single-cell PSI profile generated from sequencing reads by VALERIE. (A) Two possible isoforms from *PKM* based on MXE 9 and 10. **(B, top)** Heatmap of PSI values scaled across the columns. The columns and rows represent the genomic bases and single cells, respectively. **(B, middle)** Aggregated per-base

PSI values by the three cell populations. **(B, bottom)** -log10(p-values) derived from assessing the differences in the per-base PSI values across the three cell populations. *P* values were defined from Kruskal-Wallis and adjusted for multiple testing using Bonferroni correction.

2.4 IMPACT

2.4.1 Pre-processing of myeloid neoplasm datasets

IMPACT stands for **I**ntegrated **M**yeloid neoplasm **P**latform for prioritising **AC**tionable alternative splicing events for **T**argeted Therapy. The main goal of IMPACT is to bring together publicly available myeloid cancer, clinical, and cell line resources to prioritise clinically relevant and druggable (actionable) alternative splicing events identified from our scRNA-seq datasets.

For the clinical component of IMPACT, we have tabulated acute myeloid leukaemia (AML) data from the BeatAML and The Cancer Genome Atlas (TCGA) AML cohorts (Cancer Genome Atlas Research et al., 2013; Tyner et al., 2018).

For BeatAML, patient information such as survival data and diagnosis were retrieved from Vizome repository (http://www.vizome.org/aml/). It is noteworthy that the patient information on Vizome is more up-to-date than that provided in the supplementary table of the original publication. The gene expression information in fragment per kilobase of transcript per million (FPKM) and splice junction counts was retrieved from the National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Portal (Z. Zhang et al., 2021). The genotypes for selected genes, *SF3B1*, *SRSF2*, and *U2AF1* were retrieved from the cBioPortal for Cancer Genomics (Gao et al., 2013).

In total, 702 records were retrieved, of which 636 records diagnosed with either *de novo* AMLs (n=540) or transformed AMLs (n=96) were retained. Only samples of patients that were collected within one month of the patients' enrolment into the study (n=544) were further retained. Then, only samples with both DNA-sequencing (genotype) and RNA-seq data available at the same timepoint (n=367) were further retained. Of these samples, 345 patients had one sample, whereas 6 patients had multiple samples. We collapsed the genotype, gene expression values, and splice junction counts of these patients with multiple samples. Specifically, a patient is considered to be a carrier of a genetic variant if at least one of his/her samples carries

the genetic variant. Gene expression values and splice junction counts across multiple samples were collapsed using the average values across the samples. In total, 360 patients with 360 representative samples were included for downstream analyses. Of these patients, 322 had complete survival data, i.e., vital status and time to death (for diseased individuals) or time to last follow-up (for fortunate individuals).

Additional 33 healthy donor RNA-seq data were identified, of which 31 donors had one sample, whereas a one donor had multiple samples. We collapsed the gene expression values and splice junction counts of this patient with multiple samples. Specifically, the gene expression values and splice junction counts across multiple samples were collapsed using the average values across the samples. In total, 32 healthy donors with 32 representative samples were included for downstream analyses.

For TCGA AML, patient information such as survival data, gene expression information in fragment per kilobase of transcript per million (FPKM) and splice junction counts was retrieved from the National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Portal (Z. Zhang et al., 2021). The genotypes for selected genes, *SF3B1*, *SRSF2*, and *U2AF1* were retrieved from a previous publication (Yoshimi et al., 2019). Patient, gene expression, and genotype information were all open access. Splice junction counts were controlled access and access was granted and approved through our application to the Database of Genotype and Phenotype (dbGaP) (Project ID: 22325).

Of the 200 AML patients, 150 patients had both genotype and splice junction data available and therefore were included for downstream analyses. Of these patients, 132 had complete survival data. Unlike BeatAML, each patient in TCGA AML is represented by only one sample, and therefore we did not need to collapse the data for any multi-sample patients. Furthermore, TCGA AML only consist of *de novo* AML patients but no transformed AML patients.

2.4.2 Pre-processing of drug sensitivity datasets

2.4.2.1 *in vitro* drug screen

We identified 35 human cell lines of haematopoietic and lymphoid origins from the Cancer Cell Line Encyclopaedia (CCLE) project (Ghandi et al., 2019) that may be

used to screen for drug sensitivity and subsequently prioritise druggable alternative splicing events (Table 2.1).

Table 2.1: Disease stages and demographics of the human haematopoietic and lymphoid origins cell lines included for IMPACT.

No.	Cell Line	Disease Stage	Age (Years)	Ethnicity	Gender
1	OCIAML3	Primary	57	Caucasian	Male
2	OCIM1	Primary	62	Caucasian	Unknown
3	NB4	Primary	23	Caucasian	Female
4	KASUMI1	Primary	7	Asian	Male
5	KASUMI6	Primary	64	Asian	Male
6	GDM1	Primary	66	Caucasian	Female
7	AML193	Primary	13	African American	Female
8	NOMO1	Primary	31	Asian	Female
9	SKNO1	Unknown	Unknown	Asian	Unknown
10	OCIAML5	Primary	77	Caucasian	Male
11	OCIAML2	Primary	65	Caucasian	Male
12	ME1	Primary	40	Asian	Male
13	HEL9217	Primary	30	Caucasian	Male
14	HEL	Primary	30	Caucasian	Male
15	SIGM5	Primary	63	Caucasian	Male
16	MV411	Primary	10	Caucasian	Male
17	MONOMAC1	Primary	64	Caucasian	Male
18	MONOMAC6	Primary	64	Caucasian	Male
19	MUTZ3	Primary	29	Caucasian	Male
20	KO52	Primary	46	Asian	Male
21	SET2	Primary	71	Asian	Female
22	M07E	Primary	0.5	Caucasian	Female
23	KG1	Primary	59	African American	Male
24	PLB985	Unknown	Unknown	Caucasian	Unknown
25	PL21	Primary	24	Asian	Male
26	SKM1	Primary	76	Asian	Male
27	MOLM16	Primary	77	Asian	Female
28	MOLM13	Primary	20	Asian	Male
29	THP1	Primary	1	Asian	Male
30	TF1	Primary	35	Asian	Male
31	HL60	Primary	35	Caucasian	Female
32	CMK	Primary	0.8333	Asian	Male
33	P31FUJ	Primary	Unknown	Asian	Male
34	F36P	Metastasis	68	Asian	Male

35	EOL1	Primary	33	Caucasian	Male

RNA-seq files in FASTQ format were downloaded from Sequence Read Archive (SRA). The adaptor and nucleotide sequences from the 3'-end with Phred scores < 20 [*-q 20*] were trimmed using TrimGalore (version 0.6.5) (Martin, 2011). Trimmed reads were mapped to the human reference genome (GRCh38) using STAR aligner (version 2.6.1d) with the default settings. The first round of alignment is used to detect the splice junctions expressed in each sample. The trimmed reads were mapped to the reference genome in a second round to generate the binary alignment map (BAM) files. This second round of alignment also generates the splice junction count files for each sample based on the list of splice junctions detected across all samples in the first round of alignment.

Gene expression in transcript per million (TPM) for each sample was quantified using RSEM with the default settings (B. Li & Dewey, 2011).

The genetic variants of selected genes of interest, such as *TP53* and splicing factor genes, were retrieved from the Dependency Map (DepMap) Portal (Tsherniak et al., 2017).

The drug sensitivity read-out for each cell line was retrieved from the Cancer Therapeutics Response Portal (CTRP) (Basu et al., 2013). First, the drug sensitivity read-out in terms of area under the curve (AUC) for each compound ID and cell line experiment ID was retrieved (v20.data.curves_post_qc.txt). Next, the compound metadata such as the compound name, status (clinically approved), and gene or proteins targets were retrieved (v20.meta.per_compound.txt). Then, the cell line experiment metadata, such as cell line name and cancer type, were retrieved (v20.meta.per_experiment.txt and v20.meta.per_cell_line.txt). All information was integrated into a master drug sensitivity table.

In total, 35 cell lines and 545 compounds were available for downstream analyses.

2.4.2.2 *ex vivo* drug screen

The BeatAML study isolated mononuclear cells from AML patients and interrogated them against a panel of drug compounds (Tyner et al., 2018). The drug sensitivity read-out in terms of area under the curve (AUC) for each compound on each sample was retrieved from Supplementary Table 10 of the original publication.

The family names of which the compounds belong to were retrieved from Supplementary Table 11 of the original publication. Of the 360 patients identified from Section 2.1.4, we identified 251 patients whose samples had drug sensitivity data. Each patient was represented by one sample, and therefore we did not need to collapse the drug sensitivity data for patients with multiple samples as per Section 2.1.4. In total, 251 patients and 122 compounds were available for downstream analyses.

2.4.2.3 *in silico* drug screen

The PSI values of candidate alternative splicing events were computed using the splice junction counts derived from the CCLE cancer cell lines and BeatAML patients. These candidate alternative splicing events were derived from differential alternative splicing events detected from MARVEL and subsequently visually validated using VALERIE. The PSI values were assessed for any correlation with the AUC values across the cancer cell lines and patients for each compound. This approach was used to identify candidate compounds for skipped-exon (SE), mutually exclusive exons (MXE), alternative 5' and 3' splice sites (A5SS, A3SS), and alternative last and first exons (ALE, AFE).

The gene expression values may also be assessed for any correlation with the AUC values across the cancer cell lines and patients for each compound. This gene-centric, in lieu of splicing-centric, approach may be used to identify candidate compounds for retained introns (RIs) associated with down-regulation of gene expression (Inoue et al., 2021).

2.5 Adjunct computational pipelines

Aside from splicing-oriented computational pipelines, MARVEL, VALERIE, and IMPACT, we further developed additional computational pipelines to facilitate single-cell splicing analysis. These adjunct computational pipelines, namely two-tier sample demultiplexing, and variant calling and genotyping assignment for single-cell DNA-seq and RNA-seq, were based on the earlier published framework for TARGET-seq (Rodriguez-Meira et al., 2019).

2.5.1 Two-tier demultiplexing for single-cell DNA-/RNA-seq

Previously, each cell in a given well had a unique barcode and therefore cells may simply be demultiplexed by well ID using bcl2fastq software (Rodriguez-Meira et al., 2019). To multiplex more cells from different plates, each cell in a given well now share the same barcode with cells with the same well IDs from other plates. For example, the cell in Plate 1 well A1 will have the same barcode as the cell in Plate 2 well A1. To differentiate the cell in well A1 in Plate 1 from Plate 2, plate barcodes were added to the read construct. Therefore, we developed a demultiplexing pipeline to demultiplex the sequencing reads by well followed by plate, so that each FASTQ file represents a unique cell (Figure 2.13).

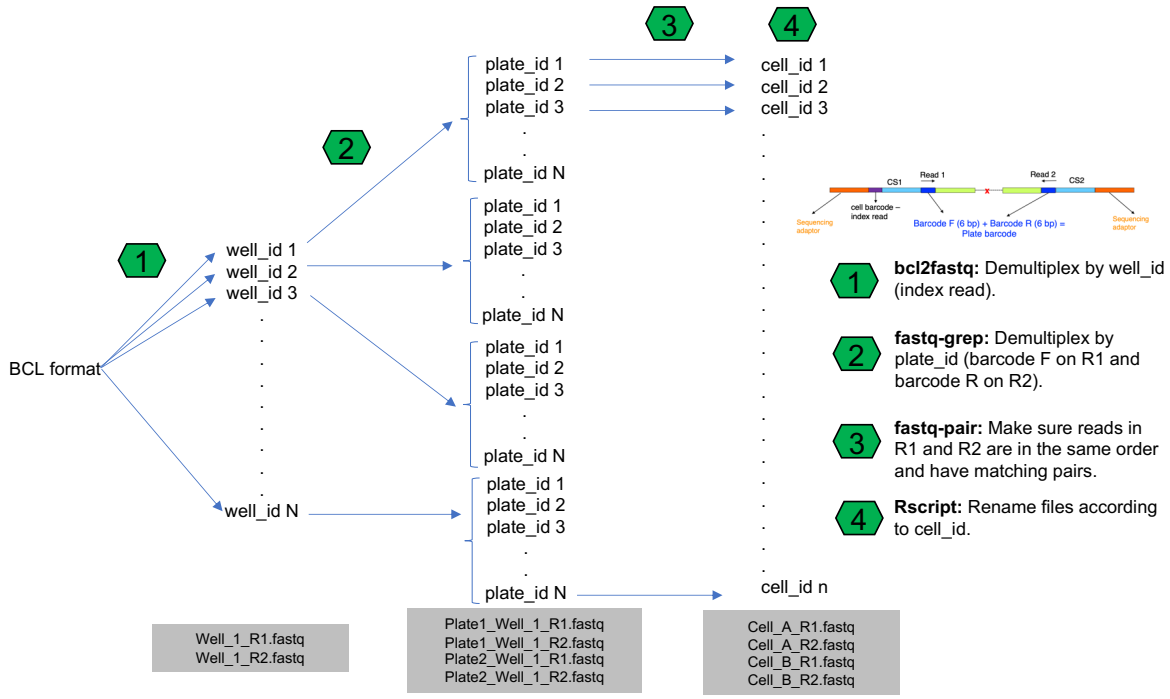


Figure 2.13: Overview of the two-tier demultiplexing pipeline. Sequencing reads are first demultiplexed by well, and then by plate.

First, bcl2fastq (version 2.20.0.422) was used to demultiplex the sequenced reads by wells. However, each FASTQ file at this stage consists of sequencing reads from other plates with the same well ID. Therefore, we used *fastq-grep* module from fastq-tools (version 0.8.3) to demultiplex the sequencing reads by plates based on the plate barcodes. Next, we ensured that the order sequencing reads in Read 1 FASTQ file matches that of Read 2 FASTQ file by using the *fastq_pair* module from fastq-pair.

At this stage, each FASTQ file represents one unique cell and is named according to its plate and well ID. Hence, we used a custom R script to rename the files to reflect the user's preferred cell IDs.

2.5.2 Variant calling for single-cell DNA-seq

Previously, the TARGET-seq genotyping pipeline for single-cell DNA/RNA-seq was optimised for single base substitutions, also known as single nucleotide variants (SNVs). But it was not optimised for detecting insertions or deletions (Rodriguez-Meira et al., 2019). Moreover, the genotyping pipeline for the single-cell DNA-seq used an RNA-seq aligner, namely STAR aligner (Dobin et al., 2013), for aligning the DNA-seq reads to the human reference genome. We updated the pipeline to more robustly identify indels, and developed two separate pipelines for calling variants from DNA and RNA reads.

To this end, we first aligned the sequencing reads to the human reference genome (hg19/GRCh37) using Burrows-Wheeler Aligner (BWA; version 0.7.17) (H. Li & Durbin, 2009). It is noteworthy that while Bowtie2 is another popular aligner (Langmead & Salzberg, 2012), it has the propensity to discard sequencing reads with a large deletion (Hasmad et al., 2016). Therefore, we proceeded with BWA, in lieu of Bowtie2, as our choice of the aligner.

Next, the resulting SAM files were converted to BAM files using the *view* module from Samtools (version 1.9) (H. Li et al., 2009). The DNA reads were then separated from RNA reads using custom Perl scripts (Rodriguez-Meira et al., 2019). Duplicate DNA reads were then flagged but not removed using the *MarkDuplicates* module from Picard (version 2.3.0). This is to ensure the duplicate reads will not be removed by downstream variant detection software. For example, MuTect2 will remove duplicate reads by default prior to the variant detection (Cibulskis et al., 2013). This will severely limit the power to detect variants in amplicon-based libraries such as TARGET-seq (J. Li et al., 2018; J. Li et al., 2019; Wen et al., 2018). This is in contrast to hybridisation capture/probe-based libraries whereby duplicate reads need to be removed prior to variant detection (Ng et al., 2016).

Mutect2 module from The Genome Analysis Toolkit (GATK) (McKenna et al., 2010) was used to detect indels. The *mpileup* module from Samtools was used to detect SNVs (Rodriguez-Meira et al., 2019). Lastly, the coverage at the variant site

was quantified using the *coverage* module from bedtools (version 2.27.1) (Quinlan & Hall, 2010).

2.5.3 Variant calling for scRNA-seq

The sequencing reads were aligned to the human reference genome (hg19/GRCh37) using the STAR aligner (version 2.6.0c). The RNA reads were then separated from the DNA reads using custom Perl scripts (Rodriguez-Meira et al., 2019). The duplicate RNA reads were then flagged, but not removed, using the *MarkDuplicates* module from Picard (version 2.3.0). Next, the sequencing reads corresponding to splicing intervals were hard-clipped using the *SplitCigarReads* module by GATK (Schischlik et al., 2019).

The *Mutect2* module from The Genome Analysis Toolkit (GATK) (McKenna et al., 2010) was used to detect indels. The *mpileup* module from Samtools was used to detect SNVs (Rodriguez-Meira et al., 2019). Lastly, the coverage at the variant site was quantified using the *coverage* module from bedtools (version 2.27.1) (Quinlan & Hall, 2010) (Figure 2.14).

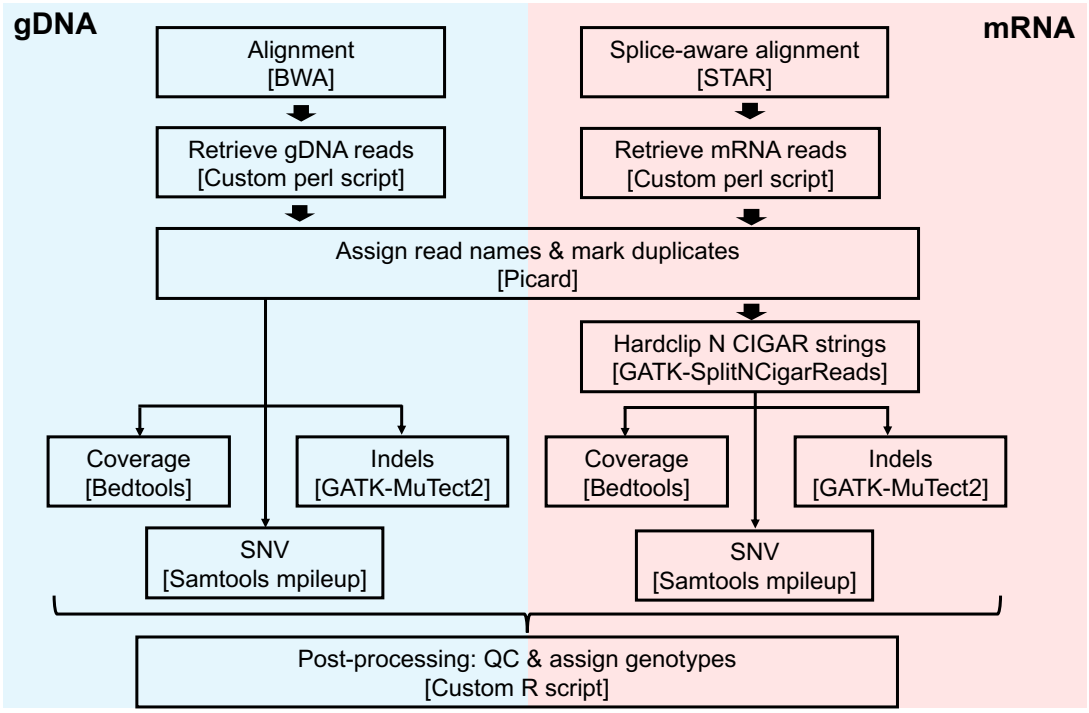


Figure 2.14: Overview of variant calling pipeline for single-cell DNA- and RNA-seq.

2.5.4 Genotype assignment

For each pre-defined variant site, the number of reads corresponding to the reference (wildtype) and alternative (variant) alleles for indels and SNVs were tabulated from the outputs of *Mutect2* and *mpileup* modules, respectively.

Here, we introduced a novel genotype scoring system to assign each variant site to one of the three genotypes: wildtype, heterozygous, or homozygous. We used the chi-square (χ^2) test to compare the observed frequency of reference and alternative alleles against the expected fraction of reference and alternative alleles corresponding to the three genotypes. The expected fraction of the reference alleles was 0.999, 0.5, and 0.001, and the expected fraction of the alternative alleles was 0.001, 0.5, and 0.999 for wildtype, heterozygous, and homozygous genotypes, respectively. The χ^2 statistics were then tabulated for each fitted model and converted to genotype scores using the following formula:

$$Score_{genotype} = \frac{1}{log10(\chi^2 + 1)}$$

The genotype assigned to the variant site will be based on the genotype model with the highest score.

Next, we computed the variant (alternative) allele frequency (VAF) and reassigned variant sites with 2 < VAF < 4 and 96 < VAF < 98 as “ambiguous”. For cells with no variants called at the variant sites by the variant callers (either due to the absence of the variants or the variants were present below the detection limit), the coverage from bedtools was used to assign the genotype. Specifically, cells with variant sites having coverage above the threshold determined from “blank” controls were assigned as “wildtype”, whereas cells with variant sites having coverage below this threshold will be assigned as “low coverage” (Figure 2.15).

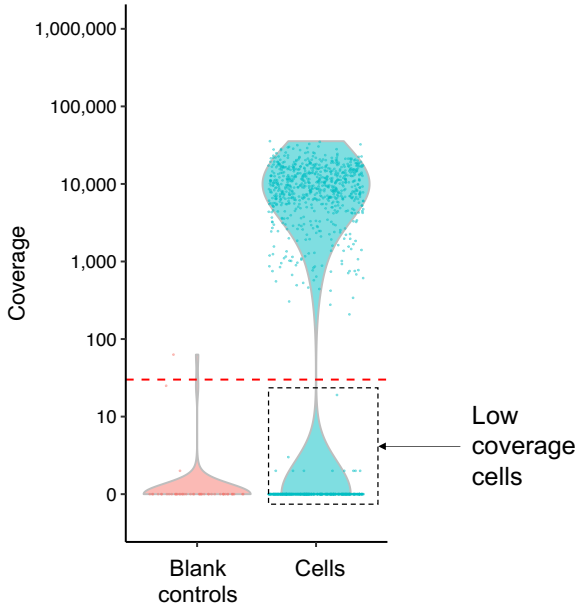


Figure 2.15: The coverage distribution for an example variant used to determine the variant calling threshold. The coverage threshold at the variant site for a given cell, below which, the cell will be annotated as "low coverage" is determined from the blank controls. Specifically, the threshold (red line) is calculated based on the coverage of the blank controls using the formula: (Q3 + IQR*1.5) + 30 reads.

Taken together, each variant site may take one of the five genotype assignments: wildtype, heterozygous, homozygous, ambiguous, or low coverage (Figure 2.16).

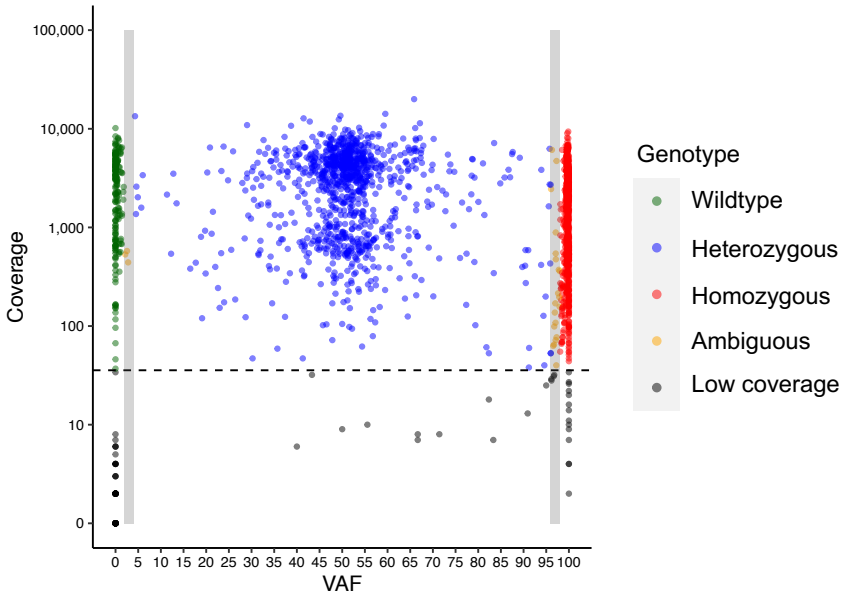


Figure 2.16: The five genotype classes for an example variant relative to coverage and variant allele frequency (VAF). The black dashed line indicates the coverage, below which, the cells were annotated as “low coverage” based on blank controls.

3 MARVEL: A novel computational tool for transcriptome-wide characterisation of alternative splicing landscape at single-cell resolution

3.1 Benchmarking percent spliced-in estimation

Percent spliced-in (PSI) measures the degree of alternative exon inclusion in a given isoform. There are mainly two approaches for computing PSI values at the single-cell level. The first approach is the Bayesian approach which combines an informative prior, such as genomic sequence features, with sequencing reads to predict the PSI values (Huang & Sanguinetti, 2017, 2021). The second approach utilises only sequencing reads, specifically splice junction reads, to quantify the PSI values (Song et al., 2017). Therefore, the PSI value represented by the former is a predicted probability, whereas the PSI value represented by the latter is an observed percentage.

The Bayesian approach has been demonstrated for skipped-exon (SE) splicing events, but not other splicing event types. Therefore, we assessed the predicted PSI values using a genomic sequence feature for each event type. Sequence conservation score (phastCons score) was used as the representative genomic sequence feature for the assessment because it is most strongly correlated with PSI values compared to other genomic features (Linker et al., 2019).

For SE splicing events, we observed the phastCons scores are highly correlated with PSI values in iPSCs (R=0.74) (Figure 3.1A) and in all datasets (median R=0.75) (Figure 3.1B). There was also little variation in the correlation values across different datasets, suggesting that phastCons score was not cell type-dependent for SE splicing events.

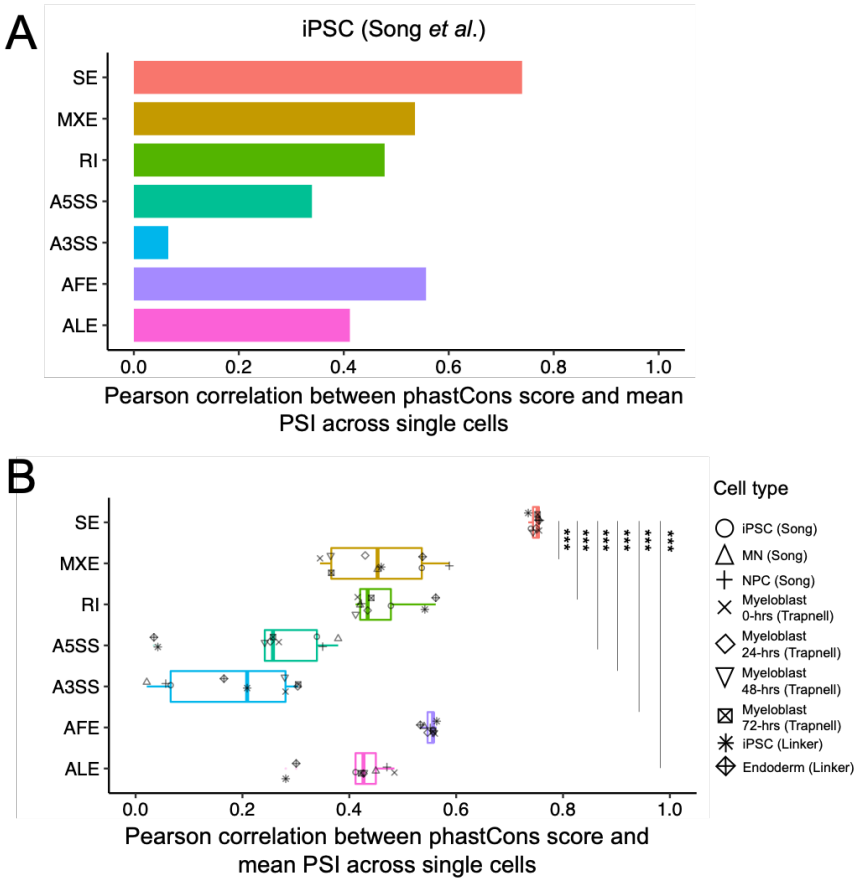


Figure 3.1: Assessing the predictive value of a genomic sequence feature (phastCons score) for PSI quantification. Correlation between phastCons scores with PSI values in **(A)** an iPSC dataset and **(B)** across all datasets included in the benchmarking analysis. FDR *** < 0.01 ** < 0.05 * < 0.1, n.s.: non-statistically significant.

The phastCons scores were weakly-to-moderately correlated with PSI values for MXE, RI, A5SS, A3SS, AFE, and ALE splicing events (median R values < 0.60). This suggests that phastCons scores had a lower predictive value for splicing event types other than SE. Moreover, except for SE, AFE, and ALE, there was a high variation in the correlation values across the different datasets for the other splicing event types. This further suggests that the correlation between phastCons scores and PSI values in these splicing event types were cell type-dependent, and not universally applicable to other cell types. Therefore, we implemented a splice junction-based approach for computing PSI values, which has been applicable for splicing event types other than SE (Kahles et al., 2018; Schischlik et al., 2019; Song et al., 2017).

Next, we assessed the reproducibility of PSI values computed using the splice junction-based approach compared with PSI values computed using existing software in homogeneous cell populations (Hagemann-Jensen et al., 2020). The PSI values of each cell in a presumed homogeneous cell population are expected to have a high correlation. For SE splicing events, we observed PSI values computed by BRIE mode 2 (prior probability based on mean PSI across a given cell population) to have higher cell-to-cell correlation compared to mode 1 (genomic features as informative prior) and mode 0 (prior probability of 50) (Figure 3.2). This indicates the improvement of PSI value estimation in the latest implementation by BRIE (mode 2). Nevertheless, Expedition and MARVEL demonstrated significantly superior cell-to-cell correlation compared to BRIE (mode 2).

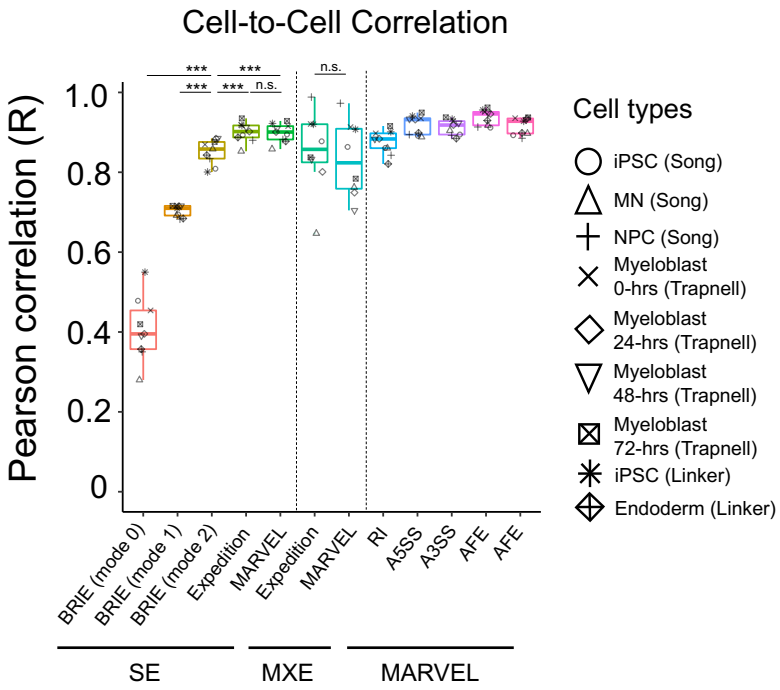


Figure 3.2: Cell-to-cell correlation between PSI values computed using BRIE, Expedition, and MARVEL. Wilcoxon rank-sum test was used here. FDR *** < 0.01 ** < 0.05 * < 0.1, n.s.: non-statistically significant.

There were no significant differences in cell-to-cell correlation between Expedition and MARVEL for SE and MXE splicing events. This is because both software implement similar approach to compute PSI values, i.e., splice junction-

based. In addition to SE and MXE splicing events, MARVEL was able to compute the PSI values for RI, A5SS, A3SS, AFE, and ALE splicing events, all of which demonstrated high cell-to-cell correlation (R > 0.8) and were not available by BRIE or Expedition.

We next compared the cell-to-bulk correlation for BRIE, Expedition, and MARVEL. BRIE (mode 2) demonstrated higher cell-to-bulk correlation compared to BRIE (mode 0/1), and comparable cell-to-bulk correlation with Expedition and MARVEL (Figure 3.3). This reaffirms the approach used by BRIE (mode 2) for inferring PSI values, i.e., using the mean PSI across a given cell population (pseudo-bulk) as the prior probability. There were no significant differences in cell-to-bulk correlation between Expedition and MARVEL for SE and MXE splicing events. Finally, the cell-to-bulk correlation of RI, A5SS, A3SS, AFE, and ALE splicing events computed by MARVEL demonstrated overall high correlation (R > 0.8).

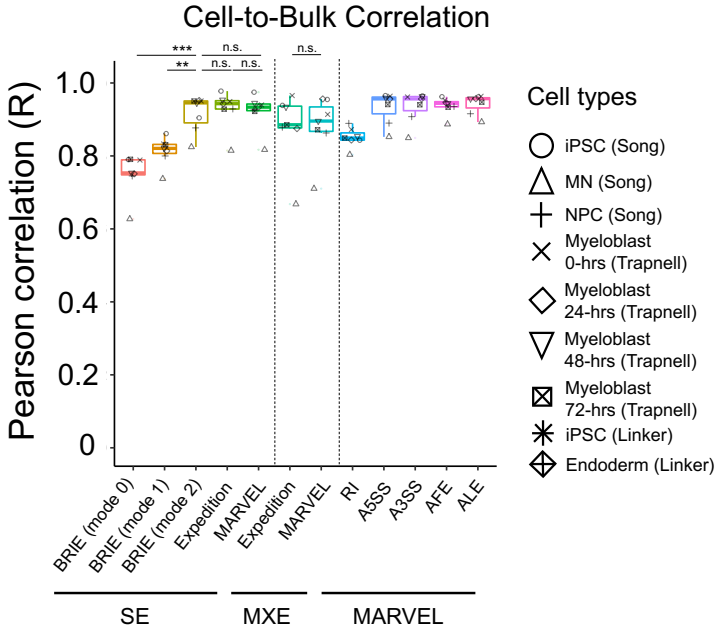


Figure 3.3: Cell-to-bulk correlation between PSI values computed using BRIE, Expedition, and MARVEL. Wilcoxon rank-sum test used here. FDR *** < 0.01 ** < 0.05 * < 0.1, n.s.: non-statistically significant.

Lastly, we compared the computational efficiency of quantifying PSI values for BRIE, Expedition, and MARVEL. We used two measurements for assessing

computational efficiency, namely the time and RAM (random-access memory) required to compute the PSI values for a fixed number of splicing events.

For SE splicing events, we observed MARVEL to require less time to compute the PSI values for 1,000 events compared to all three modes of BRIE (Figure 3.4). Similarly, for SE and MXE splicing events, we observed MARVEL to require less time to compute the PSI values for 1,000 events compared to Expedition. With the exception of RI splicing events, MARVEL required less than one minute to compute the PSI values for all other splicing event types.

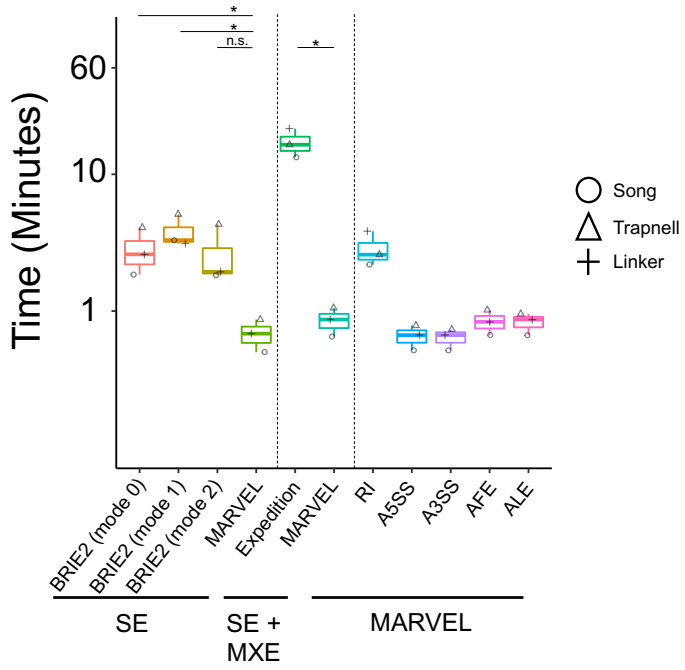


Figure 3.4: Assessing computational efficiency for computing PSI values by BRIE, Expedition, and MARVEL in terms of time required to compute the PSI values for 1,000 splicing events. Wilcoxon rank sum test used here. FDR *** < 0.01 ** < 0.05 * < 0.1, n.s.: non-statistically significant.

The shorter processing time of MARVEL compared to BRIE came at the cost of requiring slightly more RAM (Figure 3.5). On the other hand, MARVEL required slightly less RAM compared to Expedition. Overall, with the exception of RI splicing events, MARVEL required moderate amount of RAM (4-6 GBs) to compute PSI values.

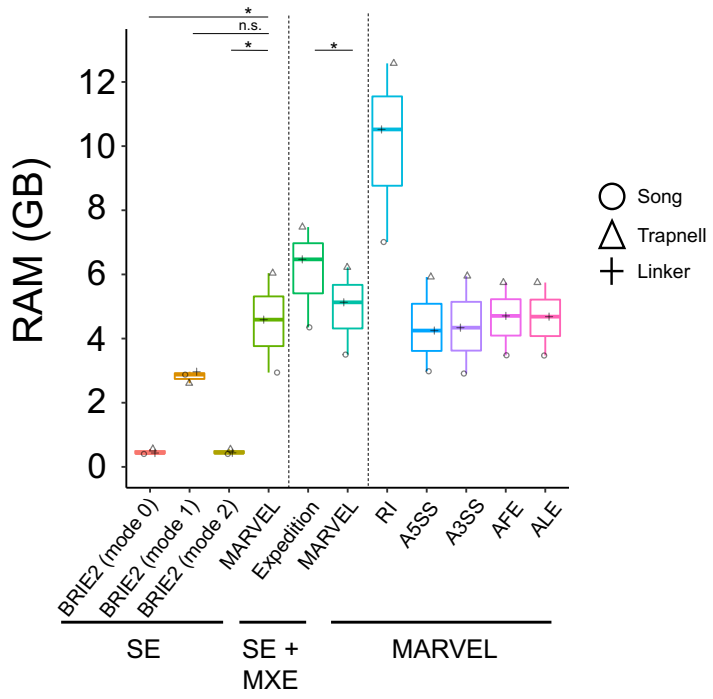


Figure 3.5: Assessing computational efficiency for computing PSI values by BRIE, Expedition, and MARVEL in terms of RAM required to compute the PSI values for 1,000 splicing events. Wilcoxon rank-sum test used here. FDR *** < 0.01 ** < 0.05 * < 0.1, n.s.: non-statistically significant.2

It is noteworthy that BRIE and Expedition were allocated four CPUs (central processing units) each for computing PSI values here whereas, with the exception of computing PSI values for RI splicing events, MARVEL does not support multi-thread processing. Therefore, MARVEL only utilised one CPU to compute the PSI values for non-RI splicing event types. Moreover, prior to computing the PSI values, Expedition required at least 16 CPUs for *de novo* detection of splicing events. This will preclude many prospective users from utilising Expedition because of the lack of access to infrastructures that can provide such large number of CPUs for multi-thread processing. For example, yours truly was not able to run Expedition prior to October 2020 because the computer cluster only had the capacity for four CPUs for multi-thread processing.

Taken together, the splice junction-based approach implemented by MARVEL for computing single-cell PSI values demonstrated better reproducibility compared to

all three modes of BRIE. Furthermore, MARVEL was able to compute the PSI values for splicing event types not provided by either BRIE or Expedition, namely RI, A5SS, A3SS, AFE, and ALE. Finally, MARVEL required the least amount of time to compute PSI values compared to BRIE and Expedition, with minimal trade-off with the amount of RAM required.

3.2 Benchmarking modality assignment

PSI can take any values between 0 and 100, therefore the PSI distribution of a given splicing event may be modelled using the beta distribution and subsequently categorised into modalities. Modalities assigns the PSI distributions into discrete categories, and the original modality classes proposed were included, excluded, bimodal, middle, and multimodal (Song et al., 2017).

However, the original modality assignment algorithm did not consider PCR amplification bias during single-cell library preparation that may lead to inaccurate modality assignment (Buen Abad Najar et al., 2020). This is especially true for bimodal distributions whereby a significant proportion of bimodal distributions was identified as spurious due to PCR amplification bias of the minor isoform. It was recently proposed that limiting splicing analysis to genes with high mRNA counts may mitigate false bimodal classifications (Buen Abad Najar et al., 2020). However, this approach will preclude a substantial number of genes from splicing analysis. To illustrate this point, we demonstrated that ~90% of genes would not been eligible for splicing analysis if a minimum of 10 cells were required to have at least 10 mRNA counts for a given gene for splicing analysis (Figures 3.6).

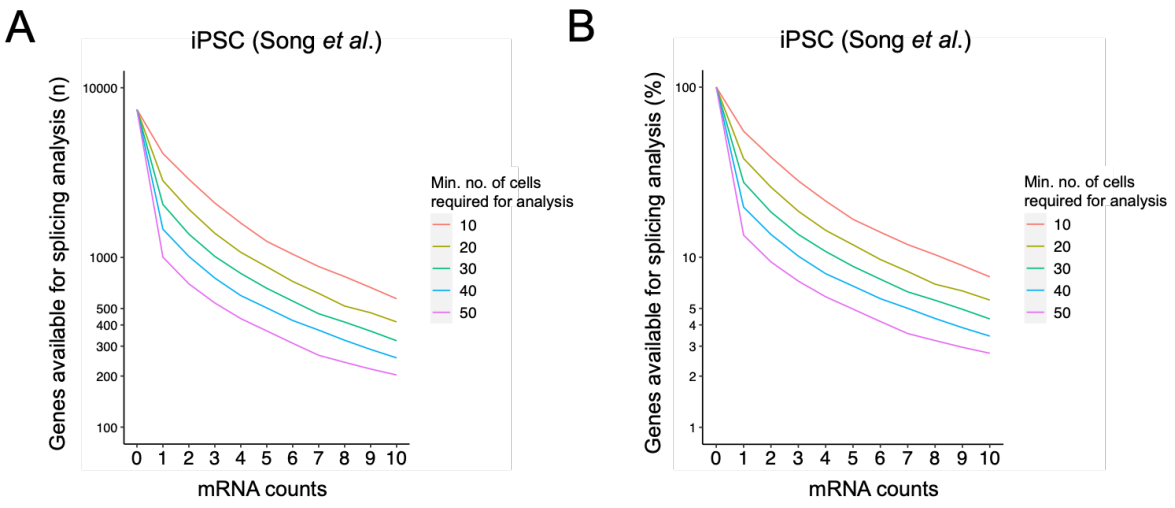


Figure 3.6: Genes available for splicing analysis at different minimum requirement of mRNA counts and cell numbers for an iPSC population as an example (Song et al., 2017). The number of eligible genes for splicing analysis in **(A)** absolute numbers and in **(B)** percentage at different mRNA count and cell number thresholds.

To avoid excluding a substantial number of genes from splicing analysis based on only including genes with high mRNA counts, we proceeded with tabulating a set of alternative splicing events with known true and false bimodal distributions in order to identify distinguishing features between true and false bimodal distributions (Figure 3.7A). We observed the fold differences (ratios) and differences in the percentage of cells with PSI values with < 25 vs > 75 (and vice versa) were able to distinguish true from false bimodal distributions (Figures 3.7B and C). This is consistent with our observation in Figure 3.7A that false bimodal distributions had disproportionated number of cells at one end of the PSI distribution compared to the other end of the PSI distribution. In contrast, bimodal distributions had similar (balanced) number of cells at both ends of the PSI distribution. Therefore, we incorporated a heuristic threshold of <3 and <50 for fold differences (ratios) and differences, respectively, in the percentage of cells with PSI values with < 25 vs > 75 (and vice versa) to distinguish true from false bimodal distributions into MARVEL. Furthermore, we observed true bimodal distributions to have mean PSI values ~50 (Figure 3.7D). Therefore, false bimodal distributions were reclassified as included or excluded modality by MARVEL when the mean PSI values were > 50 or < 50, respectively.

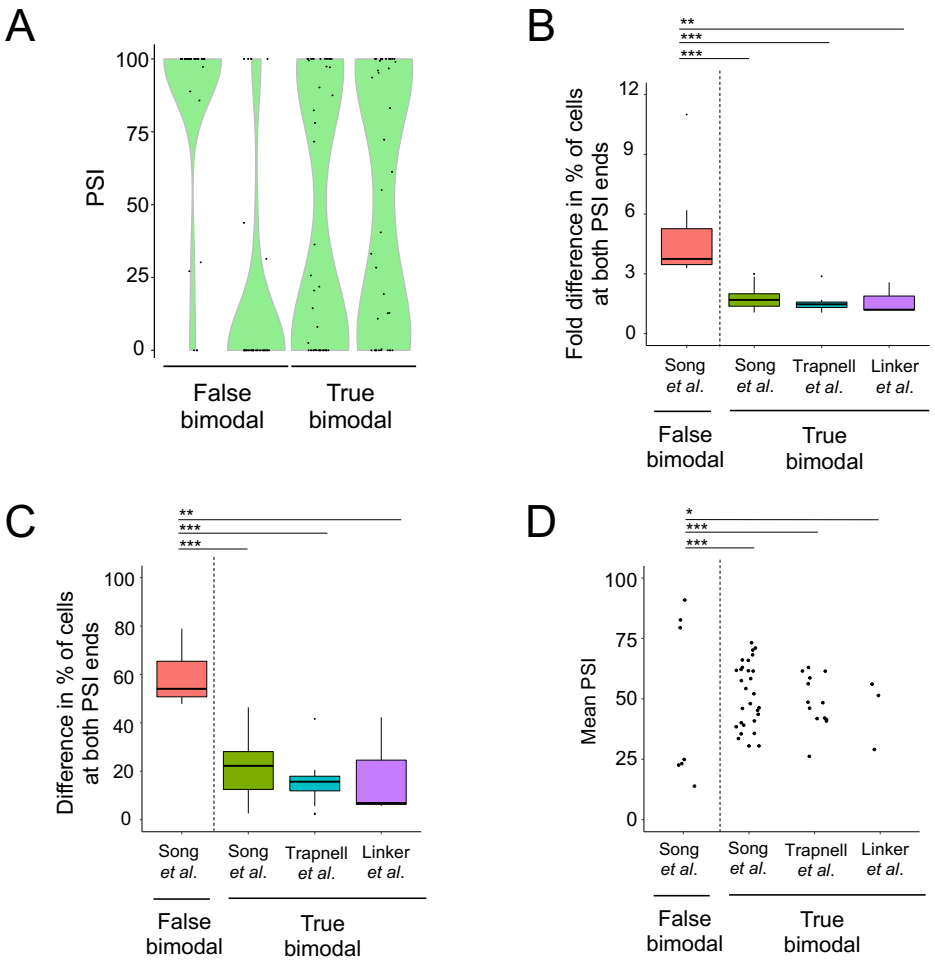


Figure 3.7: Identifying distinguishing features between true and false bimodal distributions. (A) Representative PSI distributions for splicing events with known true and false bimodal classifications. **(B-D)** Distinguishing features assessed included **(B)** fold difference (ratio) and **(C)** difference in proportion of cells with PSI values > 75 vs < 25 (and vice versa), and **(D)** mean PSI values. Wilcoxon rank-sum test was used in (B-C) while Kolmogorov-Smirnov test was used in (D). FDR *** < 0.01 ** < 0.05 * < 0.1.

Next, we assessed if MARVEL was able to distinguish bimodal from non-bimodal distributions using the learned heuristic thresholds. To this end, we tabulated a set of splicing events with known bimodal and non-bimodal distributions (Figure 3.8A). We observed MARVEL and Expedition to have comparable sensitivity, specificity, and negative predictive value (Figure 3.8B). However, Expedition had lower precision compared to MARVEL. This was attributed to the higher number of known

non-bimodal distributions that were classified as bimodal by Expedition (Figures 3.8C and D). This led to higher false positive rates, and by extension, lower precision by Expedition.

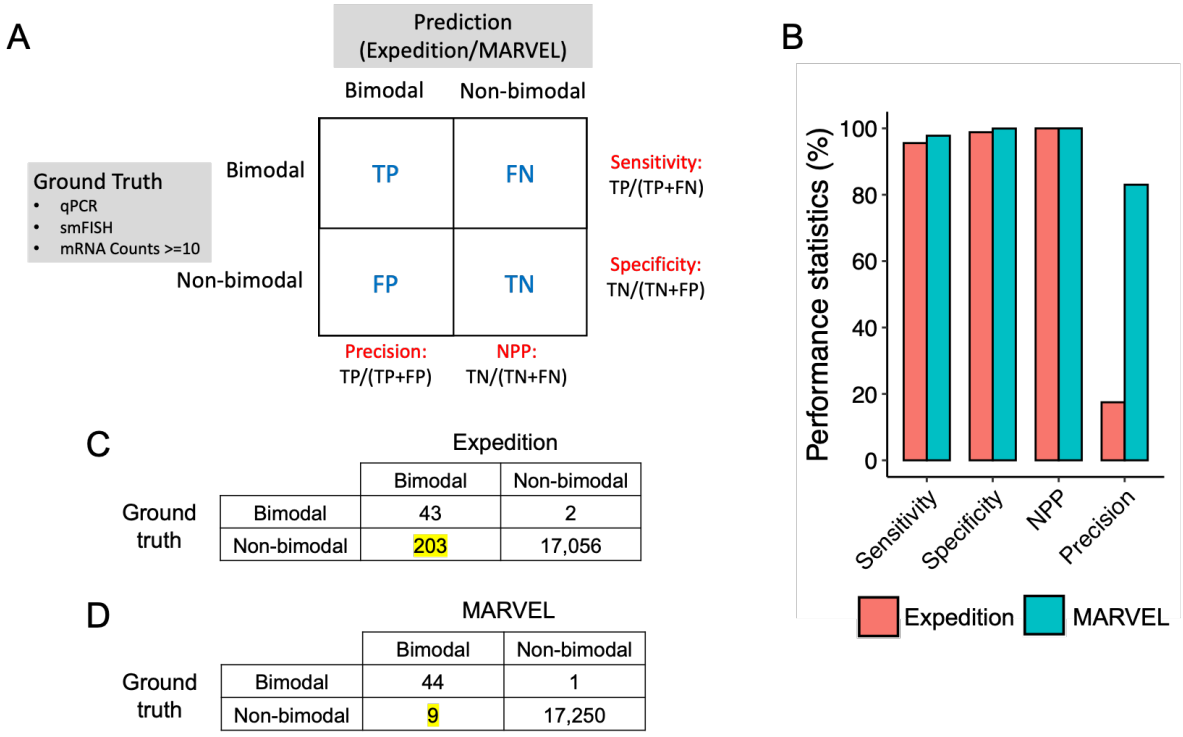


Figure 3.8: Assessing the ability of Expedition and MARVEL in classifying bimodal and non-bimodal distributions. (A) Confusion matrix generated based on known and predicted bimodal and non-bimodal classifications. **(B)** Comparison of performance metrices between Expedition and MARVEL computed from (A). **(C-D)** Values of the confusion matrix computed for **(C)** Expedition and **(D)** MARVEL. FDR *** < 0.01 ** < 0.05 * < 0.1.

Lastly, we assessed if MARVEL was able to reduce the high proportion of bimodal classification by Expedition (Song et al., 2017) to that of when only splicing events with high mRNA counts were included (Buen Abad Najar et al., 2020). To this end, we first compared the percentage of splicing events, irrespective of mRNA counts, that were classified as bimodal by Expedition and MARVEL. We observed MARVEL to classify a smaller proportion of splicing events as bimodal compared to Expedition (median 1.4% vs 7.8%) (Figure 3.9A). However, MARVEL classified a higher proportion of splicing events as bimodal compared to when only splicing events

with high mRNA counts were included (median 1.4% vs 0.2%). Nevertheless, more genes and splicing events were eligible for analysis by MARVEL compared to when only splicing events with high mRNA counts were included. Therefore, MARVEL may identify splicing events with true bimodal distributions that would otherwise be missed when the events were expressed at low-to-moderate levels. Indeed, MARVEL successfully classified a *PKM* MXE splicing event as bimodal (Figure 3.9B). The bimodal distribution of this splicing event in MN cell population has been previously validated using smFISH (Song et al., 2017). Notably, this splicing event in MN cell population had low mRNA counts (median 2.9 mRNA counts per cell) (Figure 3.9C). This is in contrast to *bona fide* housekeeping genes, such as *GAPDH*, that have hundreds of mRNA counts per cell (Figure 3.9D).

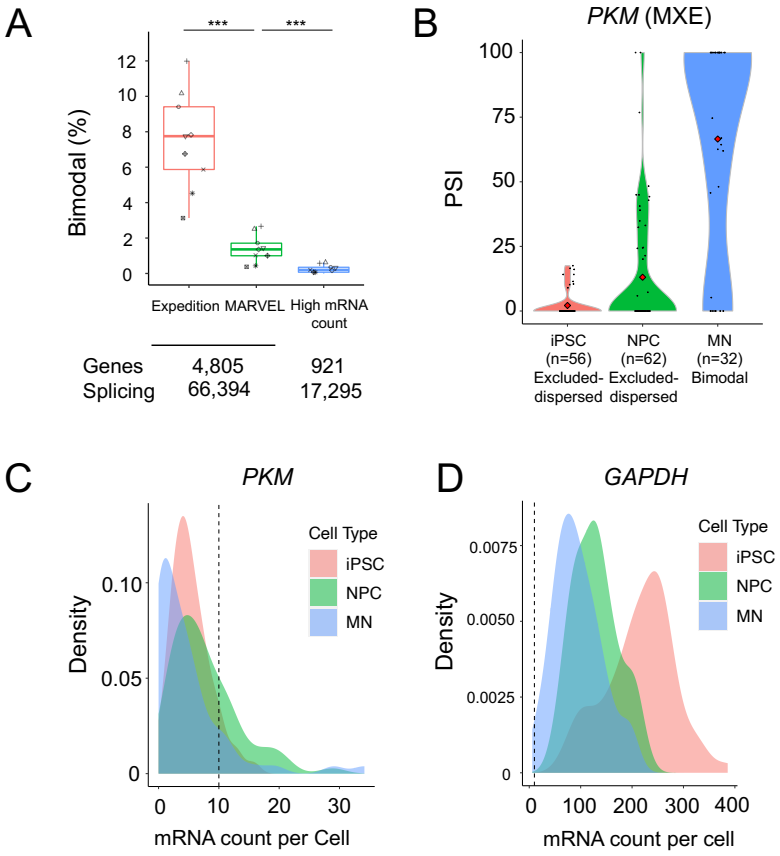


Figure 3.9: Assessing the proportion of splicing events classified as bimodal by Expedition, MARVEL, and when only splicing events with high mRNA counts were included. (A) Proportion of splicing events classified as bimodal. The number of genes and splicing events eligible for analysis shown. Note that genes and splicing

events were included for Expedition and MARVEL irrespective of mRNA counts. **(B)** PSI distributions for a splicing event that has been previously validated using smFISH. **(C-D)** The mRNA count distributions for **(C)** the gene corresponding to the splicing event in (B) and **(D)** a housekeeping gene. Wilcoxon rank-sum test used in (A). PSI: Percent spliced-in. FDR *** < 0.01 ** < 0.05 * < 0.1.

Taken together, MARVEL incorporated learned heuristic thresholds that may aid in distinguishing true from false bimodal distributions, and subsequently lead to more accurate modality assignment, without limiting the splicing analysis to highly expressed genes.

3.3 Benchmarking differential splicing analysis

Current approaches for differential splicing analysis in single cells include comparing only two cells at a time (Huang & Sanguinetti, 2017), detecting changes in modality between two cell populations (Song et al., 2017), or evaluating differences in PSI values between two homogeneous cell populations (Huang & Sanguinetti, 2021). These approaches are implemented in BRIE (mode 0/1), Expedition, and BRIE (mode 2), respectively. The first approach only allows differential splicing analysis within cells of a population but not between two cell populations. The second approach may miss splicing events with significant differences in PSI distribution between two cell populations without any change in modality. The third approach assumes both cell populations consist of homogeneous cell populations, and therefore may not be suitable for comparing heterogeneous cell populations. Therefore, there is a need for novel single-cell differential splicing analysis framework.

We first evaluated four different statistical tests for differential splicing analysis in 0- and 72-hrs myoblast (Trapnell et al., 2014), namely Kolmogorov-Smirnov (KS), Anderson-Darling (AD), D Test Statistics (DTS) (Dowd, 2020), and Wilcoxon rank-sum test. We selected Kolmogorov-Smirnov, Anderson-Darling, and DTS for evaluation because these statistical tests consider the entire PSI distribution between sample groups, instead of only evaluating differences in average PSI values between sample groups. Therefore, these statistical tests will be able to distinguish samples groups with similar average PSI values but with different PSI distributions, such as bimodal, middle, and multimodal. These three modalities have similar average PSI values but

every different PSI distribution. We selected myoblast here because muscle-related gene sets were expected to be differentially spliced in this cell type, and therefore these gene sets may serve as a ground truth for benchmarking our differential splicing analysis framework.

We observed DTS to identify the most number differentially spliced events. Nevertheless, we noticed a substantial number of these differences were driven by a small number of outlier cells in either one of the cell populations. These outlier cells had PSI > 0 or PSI < 100 in cell populations with excluded (Figures 3.10A and B) or included modality (Figures 3.10C and D), respectively. To mitigate differential splicing events driven by these small number of outlier cells, we applied our bimodal-adjust modality assignment algorithm to identify cell populations with excluded-to-excluded or included-to-included modality change. We then further required either one of the two cell populations to have a minimum of 10 cells with PSI > 0 or PSI < 100 for cell populations with excluded (Figures 3.10E and F) or included modality (Figures 3.10G and H), respectively. Using this outlier adjustment approach, we successfully reduced the number of differentially spliced events identified by DTS to that of comparable to other statistical tests (Figure 3.10I).

We next compared the number of differentially spliced events detected by the different statistical tests. We observed AD and DTS captured majority of the differentially spliced events (Figure 3.10J). Therefore, we recommended combing AD and DTS together with our outlier adjustment technique as the default differential splicing test in MARVEL.

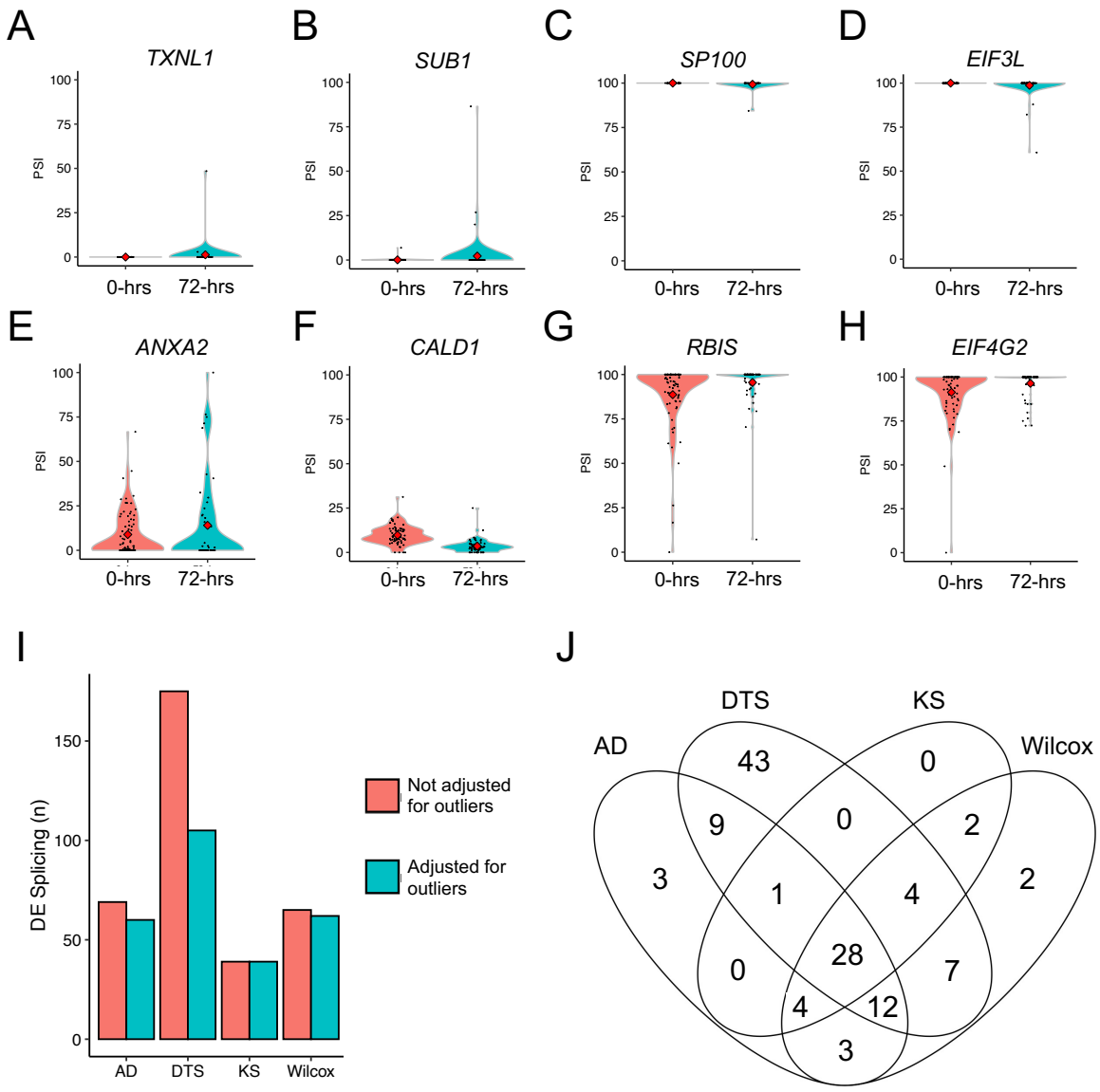


Figure 3.10: Comparing four different statistical tests for differential splicing analysis. (A-D) Representative differentially spliced events detected by DTS that were driven by small number of outlier cells. **(E-H)** Representative differentially spliced events detected by DTS after adjusting for outlier cells. **(I)** Number of differentially spliced events before and after removing events driven by small number of outlier cells. **(J)** Comparing the number of overlapping or statistical test-specific differentially spliced events across the different statistical tests evaluated here. AD: Anderson-Darling; DE: Differential; DTS: D Test Statistics; KS: Kolmogorov-Smirnov.

Next, we benchmarked our differential splicing analysis framework against BRIE (mode 2) using 0- and 72-hrs myoblast (Trapnell et al., 2014). MARVEL

identified 114 differentially spliced events whereas BRIE (mode 2) identified 73 differentially spliced events (Figure 3.11A). Fifty-seven differentially spliced events were detected by both software, and more events were exclusively detected by MARVEL compared to BRIE (mode 2). Notably, differentially spliced genes identified by MARVEL were enriched for muscle-related pathways (Figure 3.11B). This is consistent with the biological pathways expected to be regulated when immature myoblast (0-hrs) developed into more mature myoblast (72-hrs). Moreover, differentially spliced genes identified exclusively by MARVEL were enriched for protein translation pathways whereas no enriched pathways were identified among differentially spliced genes identified exclusively by BRIE (mode 2) (Figure 3.11C).

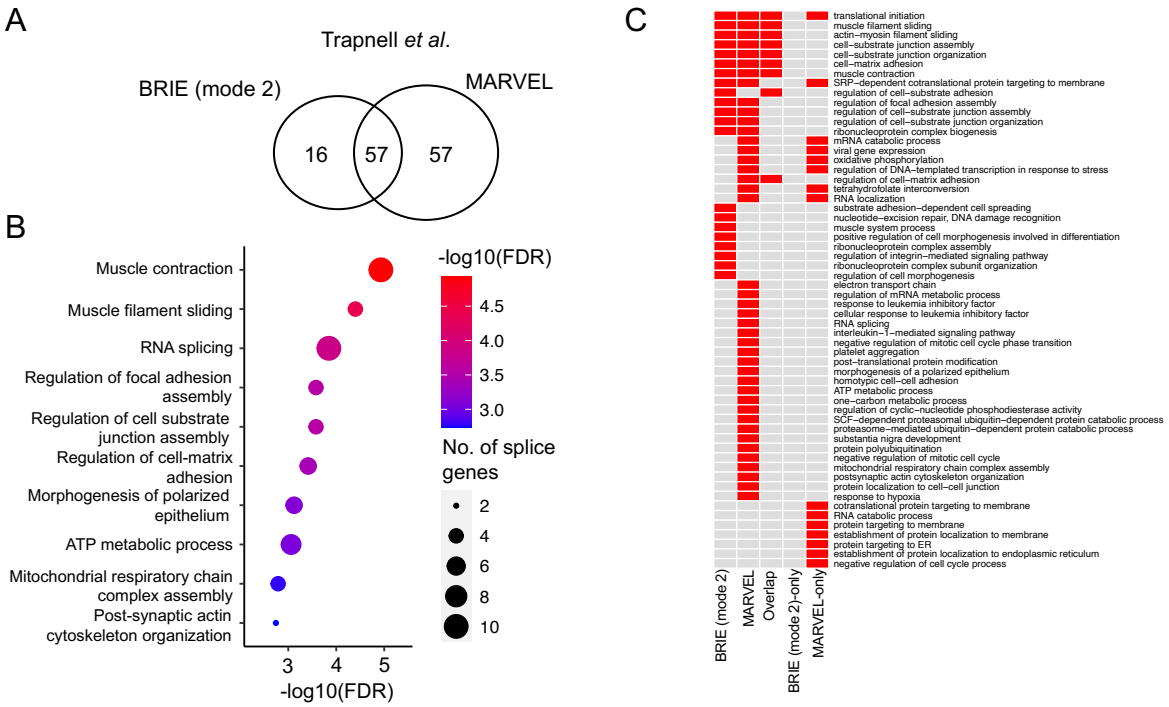


Figure 3.11: Comparing differentially spliced events detected by BRIE (mode 2) and MARVEL in 0- vs 72-hrs myoblasts. (A) Venn diagram revealing the number of differentially spliced events detected by both software (overlap) or exclusively to either software. **(B)** Biological pathways enriched among differentially spliced genes detected by MARVEL. **(C)** Comparison of all biological pathways enriched among differentially spliced genes detected by BRIE (mode 2), MARVEL, both software (overlap), or exclusive to either software.

To assesses the generalisability of our differential splicing analysis benchmarking results, we further benchmarked our differential splicing analysis framework against BRIE (mode 2) using single cells derived from the spinal cords of mice induced with multiple sclerosis (EAE) and healthy control mice (Falcao et al., 2018). MARVEL identified 248 differentially spliced events whereas BRIE (mode 2) identified 238 differentially spliced events (Figure 3.12A). One hundred and eight differentially spliced events were detected by both software, and the number of events exclusive to either BRIE (mode 2) or MARVEL was similar. The smaller number of overlapping differentially spliced events detected by both software in this dataset compared to the previous human myoblast dataset may be attributed to one or more the following reasons: (1) This mouse dataset consists of heterogeneous cell populations which violates the assumption of homogeneous cell population of BRIE (mode 2), (2) this mouse dataset was generated with relatively short reads compared to the human myoblast dataset (50bp SE vs 100bp PE) and also had lower coverage (median sequencing depth <500,000 vs >1,000,000), both of which may present challenges in identifying differentially spliced events.

Differentially spliced genes identified by MARVEL were enriched for biological pathways related to the central nervous system (Figure 3.12B). This is consistent with the biological pathways expected to be dysregulated in multiple sclerosis which is an autoimmune disease that attacks the myelin sheaths of the neurons (Ghasemi, Razavi, & Nikzad, 2017). Moreover, differentially spliced genes identified exclusively by MARVEL were enriched for RNA splicing pathways whereas no enriched pathways were identified among differentially spliced genes identified exclusively by BRIE (mode 2) (Figure 3.12C).

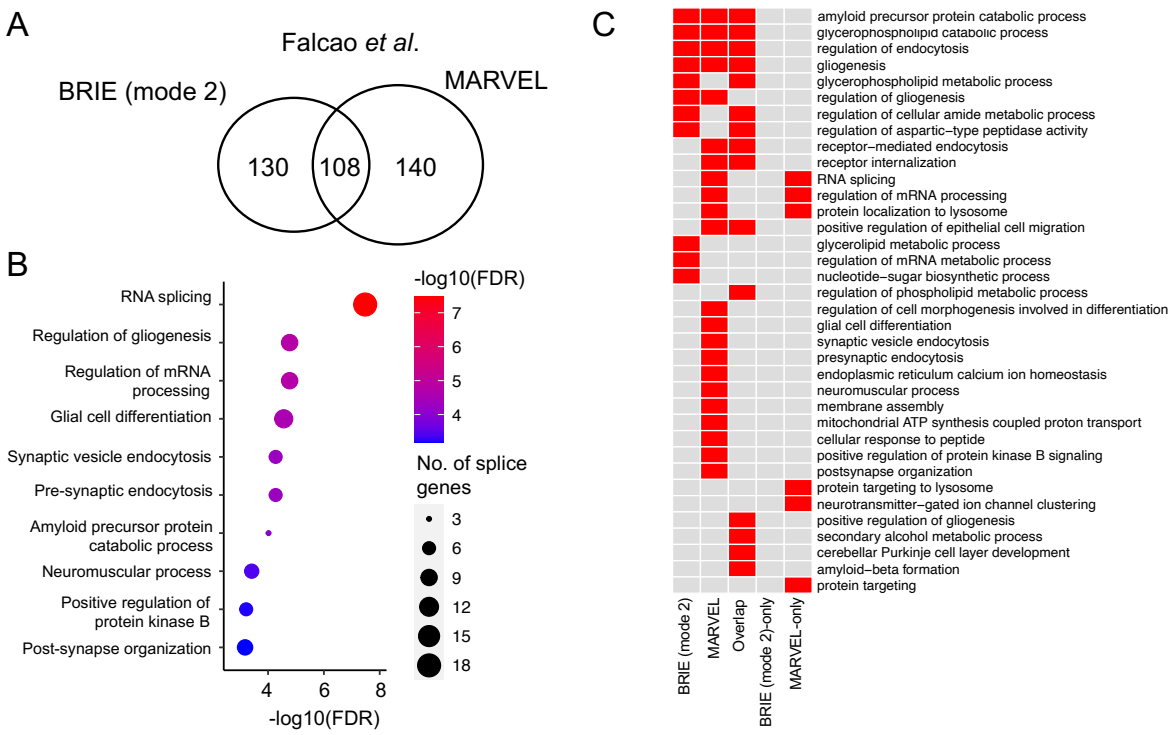


Figure 3.12: Comparing differentially spliced events detected by BRIE (mode 2) and MARVEL in mice with multiple sclerosis vs healthy control mice. (A) Venn diagram revealing the number of differentially spliced events detected by both software (overlap) or exclusively to either software. **(B)** Biological pathways enriched among differentially spliced genes detected by MARVEL. **(C)** Comparison of all biological pathways enriched among differentially spliced genes detected by BRIE (mode 2), MARVEL, both software (overlap), or exclusive to either software.

Taken together, we introduced a novel statistical framework for differential splicing analysis between two cell populations and demonstrated its ability to detect biological relevant differentially spliced genes. Lastly, MARVEL complements existing software, namely BRIE (mode 2), by identifying differentially spliced events otherwise missed by existing software.

3.4 Demonstration on plate-based RNA-seq dataset

To demonstrate the full range of features available by MARVEL for analysing plate-based RNA-seq dataset, we have performed single-cell splicing analysis on induced pluripotent stem cells (iPSCs) and endoderm cells differentiated from iPSCs (Linker et al., 2019). We have chosen this dataset because this dataset represents

two disparate cell populations. Therefore, we expect the splicing profile of these populations to be very distinct from one another, and consequently augment the analytical features of MARVEL such as differential splicing analysis.

We first surveyed the total number of splicing events expressed in each of the cell population and further stratified the splicing events into the different splicing event types. A splicing event was considered to be expressed when the splicing event was supported at least 10 reads in at least 25 cells. We observed 13,125 and 5,308 splicing events to be expressed in iPSCs and endoderm cells, respectively (Figures 3.13A and B). In both cell populations, the dominant splicing event type was skipped-exon (SE) followed by retained intron (RI), alternative first exon (AFE), alternative 3' splice site (A3SS), alternative 5' splice site (A5SS), alternative last exon (ALE), and finally mutually exclusive exons (MXE).

Consistent with previous studies, SE was the most prevalent splicing event type (Pellagatti et al., 2018; Shiozawa et al., 2018). Nevertheless, we have also shown here that splicing event type other than SE constituted more than half of expressed splicing events. However, single-cell splicing studies to date have focused only on SE and MXE event types (Huang & Sanguinetti, 2017, 2021; Song et al., 2017; Warf, Diegel, von Hippel, & Berglund, 2009). Therefore, MARVEL may be able to more comprehensively describe the splicing landscape and generate more novel biological insights from single-cell analysis compared to other published single-cell splicing software.

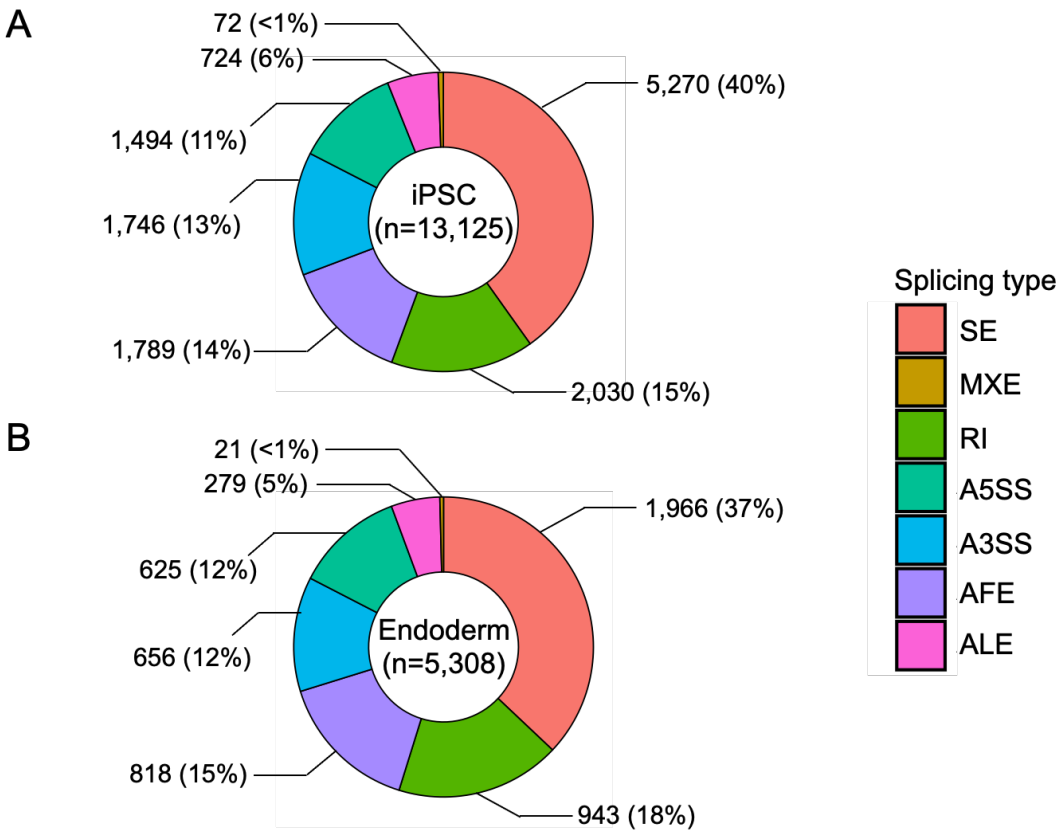


Figure 3.13: Expressed splicing events. The number and proportion of expressed splicing events stratified by splicing event types in **(A)** iPSCs and (**B)** endoderm cells.

Next, we stratified each splicing event into modalities. The different modality classes are able to reveal the overall splicing pattern of a given splicing event across a cell population. Therefore, modality classes are able to reveal whether a presumed homogeneous cell population predominantly expresses one type of isoform (isoforms that include (splice in) or exclude (splice out) the alternative exon) or both types of isoforms for a given splicing event.

We observed included and excluded modalities to be the most prevalent modality classes in both iPSCs and endoderm cells (Figure 3.14A and B). On the other hand, bimodal, middle and multimodal modalities constituted only <5% of all expressed splicing events. This suggests that single cells most commonly express one dominant isoform for a given splicing event, i.e., either the isoform includes (splices in) the alternative exon (included modality) or the isoform excludes (splices out) the alternative exon (excluded modality). This is consistent with previous single-cell

studies that used empirical and simulated RNA-seq datasets to demonstrate the dominance of one isoform for a given splicing event (Westoby et al., 2018).

We further showed that primary and dispersed modalities constituted the included and excluded modalities in roughly equal proportions. The primary included and excluded modalities suggest that single cells either absolutely included (spliced in) or excluded (spliced out) the alternative exon for a given splicing event, respectively. On the other hand, the dispersed included and excluded modalities suggest that while there was a dominant population of cells that absolutely included (spliced in) or excluded (spliced out) the alternative exon, there existed a minor population of cells that expresses both isoforms.

We further surveyed the proportion of the different modality classes in each splicing event type to identify if certain modality classes were more prevalent in certain splicing event types compared to other splicing event types (Figure 3.14C and D). We observed included and excluded modalities to be the dominant modality classes across all splicing events in both iPSCs and endoderm cells. On the other hand, bimodal, middle, and multimodal modalities constitute only <5% of all expressed splicing events across all splicing event types. This suggests that typically only one dominant isoform is expressed (included or excluded) in a given cell, consistently with previously studies (W. Liu & Zhang, 2020; Song et al., 2017).

Nevertheless, there were certain splicing events that were slightly more prevalent in a specific splicing event type compared to other splicing event types. Notably, RI had the highest rate of excluded modality classification. Specifically, primary and dispersed excluded modalities collectively constituted ~75% of all RI splicing events in both iPSCs and endoderm cells. This is consistent with the role of RI in regulating gene expression in physiological and disease states whereby the inclusion (splicing in) of the intron within the isoform creates a premature stop codon (PTC) and subsequently leads to nonsense-mediate decay (NMD) of truncated isoform (Smart et al., 2018).

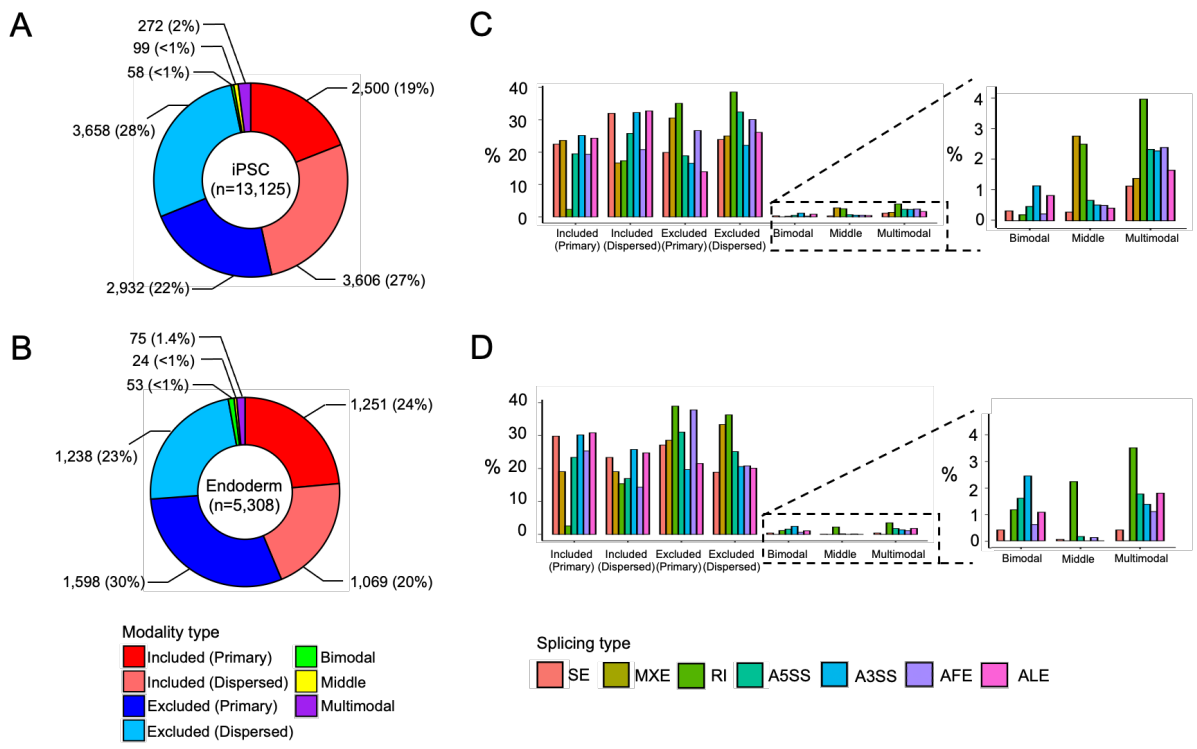


Figure 3.14: Modality classification of expressed splicing events. (A-B) Proportion of each modality classes in **(A)** iPSCs and **(B)** endoderm cells. **(C-D)** Proportion of each modality classes stratified by splicing event type in **(C)** iPSCs and **(D)** endoderm cells.

Using the expressed splicing events identified in both iPSCs and endoderm cells, we assessed whether splicing represents a source of heterogeneity underlying gene expression profile. To this end, we performed differential gene expression analysis between the two cell populations and identified 7,643 differentially expressed genes. Using these differentially expressed genes (FDR < 0.10 and |log2 fold change| > 0.5), we were able to distinguish the two cell populations on the principal component analysis (PCA) space, as expected (Figure 3.15A). However, the non-differentially expressed genes did not distinguish the two cell populations (Figure 3.15B). Interestingly, expressed splicing events of the non-differentially expressed genes successfully separate the two cell populations (Figure 3.15C). The two cell populations remained distinguishable when expressed splicing events for each splicing event type was used for dimension reduction analysis individually (Figures 3.15D-J). This

suggests that splicing represents an additional layer of complexity underlying and invisible at gene expression level.

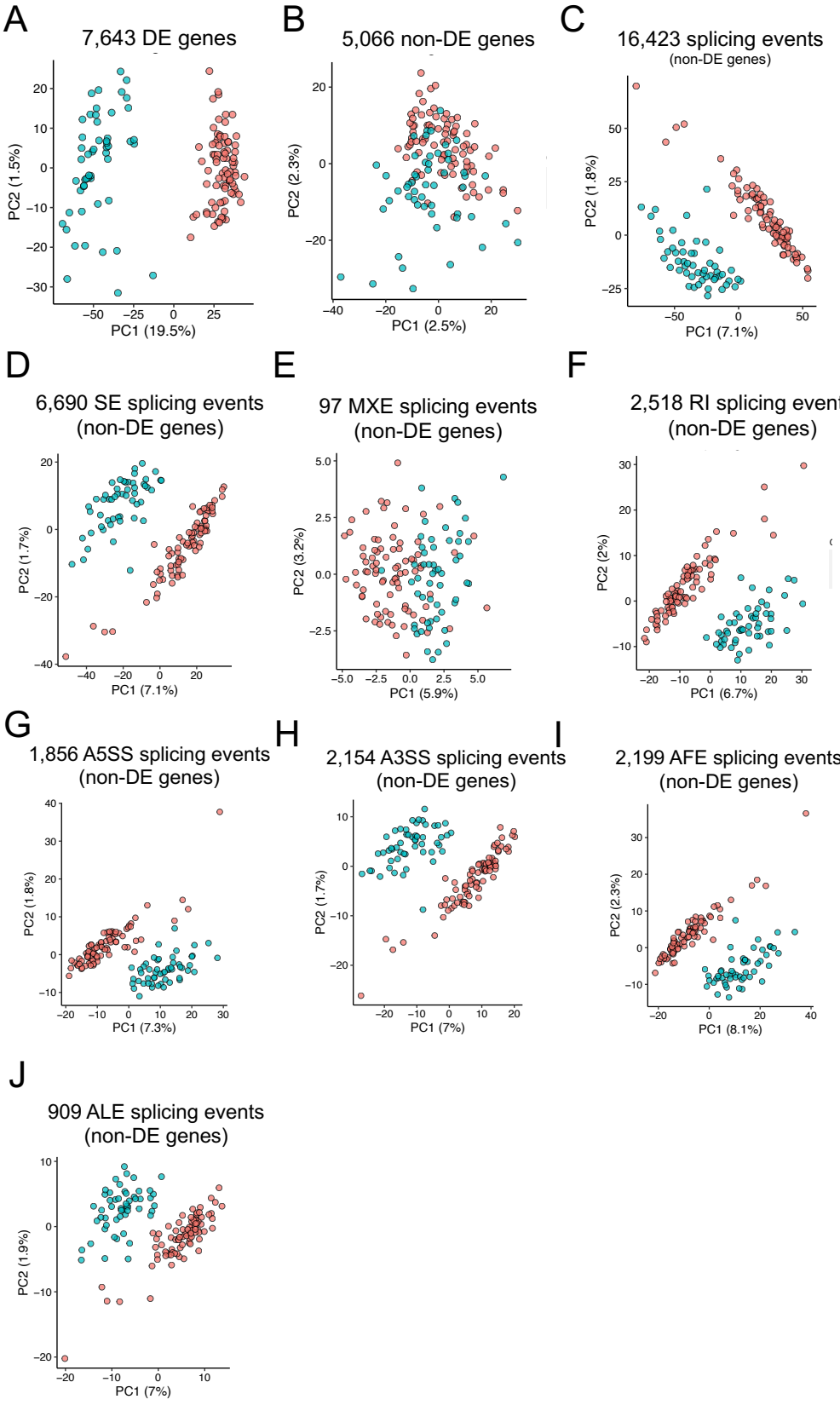


Figure 3.15: PCA using genes and splicing events. (A-B) PCA using **(A)** differentially and **(B)** non-differentially expressed genes. **(C-J)** PCA using **(C)** all expressed splicing events of non-differentially expressed genes and splicing events only from **(D)** SE, **(E)** MXE, **(F)** RI, **(G)** A5SS, **(H)** A3SS, **(I)** AFE, or **(J)** ALE.

Differential expression analysis is the cornerstone of RNA-seq analysis and is the pivotal step to identify candidate biomarkers for downstream functional validation. Differential splicing analysis between iPSCs and endoderm cells identified 1,614 differentially spliced events (FDR < 0.1). Ribosomal genes constituted the top significant splicing events (Figure 3.16A). The top 10 most significant splicing events were *DNAJC15* (RI), *SNRPN* (SE.1, SE.2), *RPL26* (A5SS and AFE), *RPS24* (SE and A3SS), *RPS10-NUDT3* (A5SS and AFE), *RPS14* (AFE).

We stratified the differentially spliced events based on their modality change from iPSCs to endoderm. To this end, we categorised the modality changes into explicit, implicit, and restricted. These categorises represent the extent of splicing pattern changes and therefore enable us to identify splicing events with overt or subtle change in splicing profiles from one cell population to another. Firstly, explicit change involved the change in one of the five original modality classes, i.e., included, excluded, bimodal, middle, and multimodal (Song et al., 2017). Secondly, implicit change involved the change from primary to dispersed, or vice versa. Thirdly, restricted change occurred when the splicing profile between the two cell populations was statistically significant, such as when there was a difference in mean PSI values, but both cell populations retained the same modality classification.

Majority of differentially spliced events underwent restricted modality change from iPSCs to endoderm cells, followed by implicit and explicit (Figure 3.16B). Example of splicing events that demonstrated explicit, implicit, and restricted modality changes were *CNBP* (A5SS), *ABI2* (ALE), and *ACTB* (RI), respectively (Figures 3.16C-E). It is noteworthy that only a small proportion of splicing events underwent overt splicing changes, i.e., explicit modality changes. Therefore, defining differentially spliced events based on splicing events which demonstrated only explicit modality changes between two cell populations, as previously described (Song et al., 2017), would miss a substantial number of potentially biologically relevant splicing events.

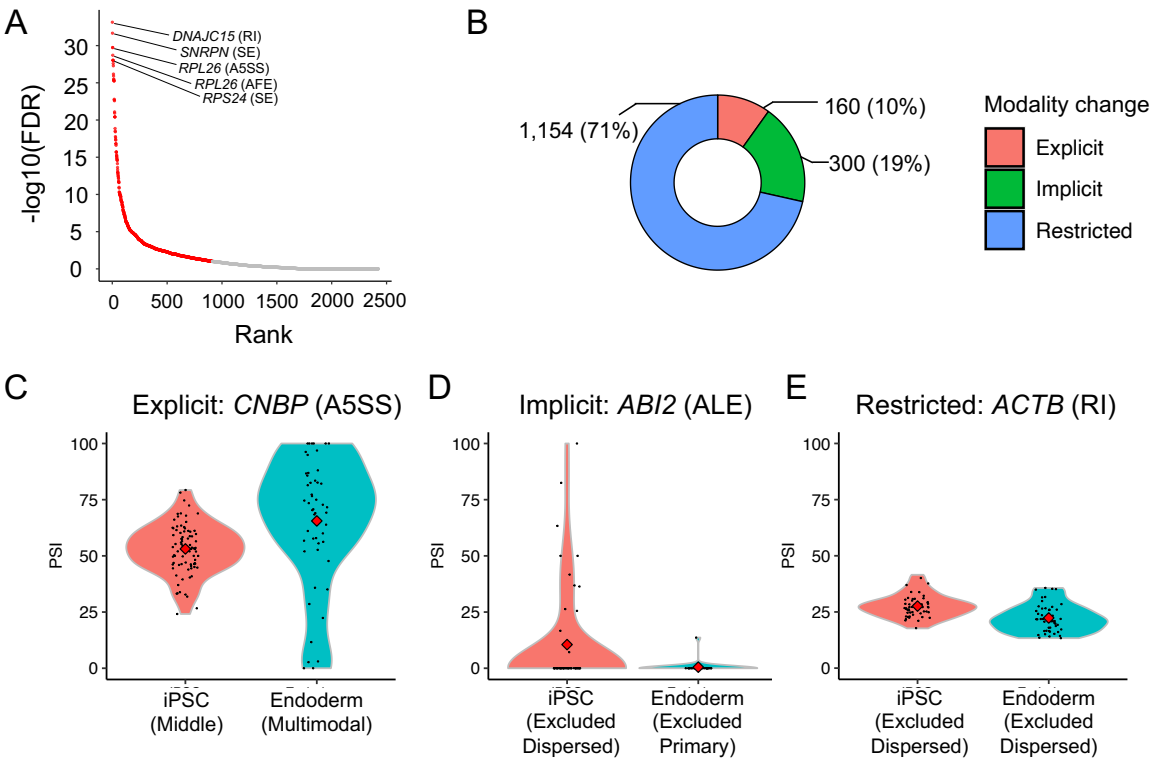


Figure 3.16: Differential splicing analysis between iPSCs and endoderm cells. **(A)** Splicing events ranked from most-to-least statistically significant based on Anderson-Darling test. **(B)** Proportion of differentially spliced events based on type of modality change. **(C-E)** Representative examples of splicing events that underwent **(C)** explicit, **(D)** implicit, and **(E)** restricted modality change from iPSCs to endoderm cells.

Current single-cell splicing software only enable splicing analysis without gene expression profiling. MARVEL is able to perform both differential gene and splicing analysis and therefore provide valuable insights into the gene-splicing relationship between two cell populations.

The 1,614 differentially spliced events identified between iPSCs and endoderm cells constituted 816 genes. Of which, 479 (59%) were differentially expressed (Figure 3.17A).

To understand the gene-splicing relationship, we categorised the change in splicing profile relative to change in gene expression profile between two cell populations into coordinated, opposing, isoform switching, and complex. Coordinated and opposing gene-splicing relationships are defined as the change in mean gene

expression value is in the same or oppositive direction relative to change in mean percent spliced-in (PSI) value, respectively. Isoform-switching is defined as significant change in splicing profile between two cell populations without significant change in gene expression profile (Smith et al., 2019). Lastly, complex gene-splicing relationships are defined as genes with both coordinated, opposing, isoform switching relationships with two or more splicing events. Isoform-switching constituted majority of gene-splicing relationships, followed by coordinated, opposing, and finally complex (Figure 3.17B). Examples of coordinated, opposing, isoform-switching, and complex gene-splicing relationships were *DHX9* (Figures 3.17C and D), *BCLAF1* (Figures 3.17E and F), *CELF1* (Figures 3.17G and H), and *TERF1* (Figures 3.17J-K), respectively.

It is noteworthy that opposing, isoform-switching, and complex gene-splicing relationships may not be inferred directly from differential gene expression analysis. This is because, unlike coordinated gene-splicing relationship, the gene expression profile changes were not in the same direction relative to splicing changes. Therefore, differential splicing analysis may identify differentially regulated genes that would otherwise been missed by differential gene expression analysis alone.

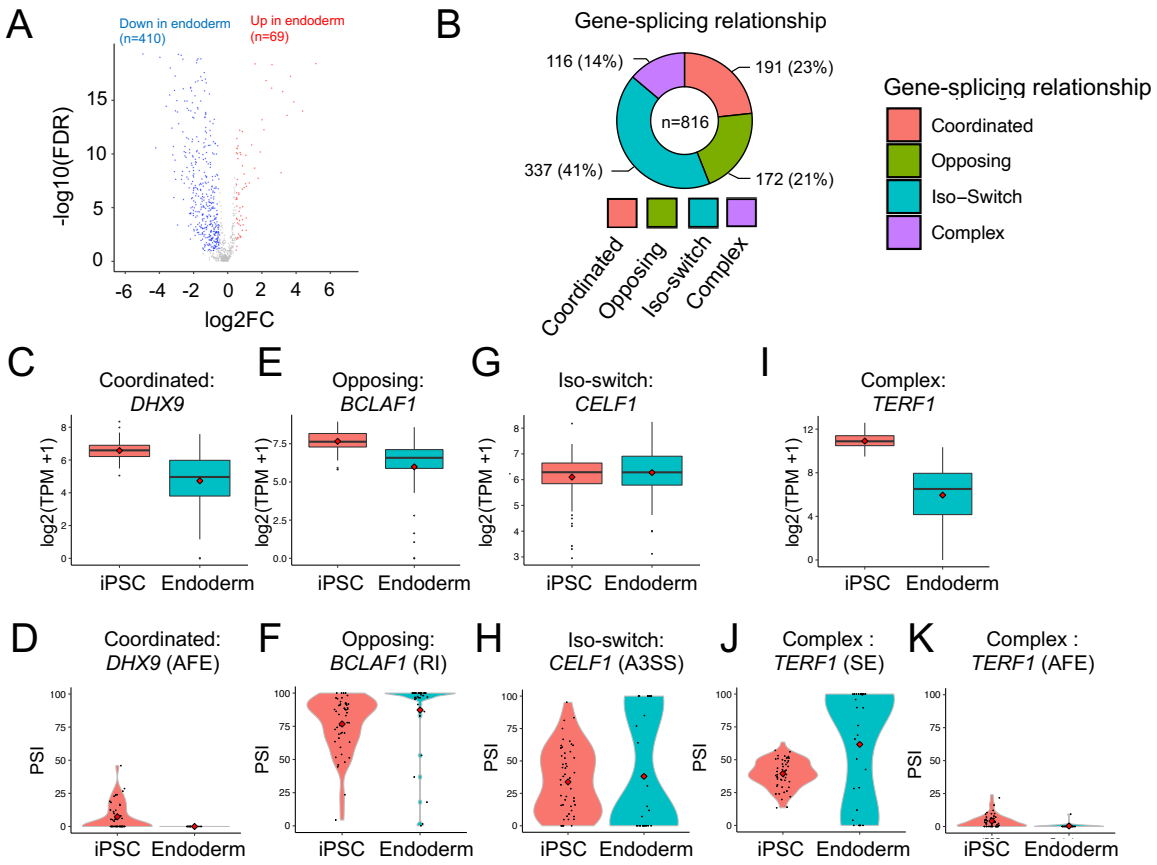


Figure 3.17: Gene-splicing relationship between iPSCs and endoderm cells. (A) Differential gene expression analysis of 816 genes that were differentially spliced between iPSCs and endoderm cells. **(B)** Stratification of gene-splicing relationship into coordinated, opposing, isoform-switching, and complex. **(C-K)** Representative examples of **(C-D)** coordinated, **(E-F)** opposing, **(G-H)** isoform-switching, and **(I-K)** complex gene-splicing relationship.

Following integrated differential gene and splicing analysis, MARVEL provides two features for functional annotation of differentially spliced events, namely gene ontology (GO) analysis and NMD prediction. The former identifies gene sets that are enriched among differentially spliced genes. The latter identifies isoform subjected to NMD when the alternative exon is included (spliced in) into the isoform. Both approaches may enable identification of candidate genes and isoforms for downstream experimental studies.

GO analysis of differentially spliced genes identified 141 biological pathways that were enriched (FDR < 0.05) among the 816 genes that were differentially spliced

between iPSCs and endoderm cells. The top pathways included gene transcription and translation, alternative splicing, and ribosome assembly (Figure 3.18).

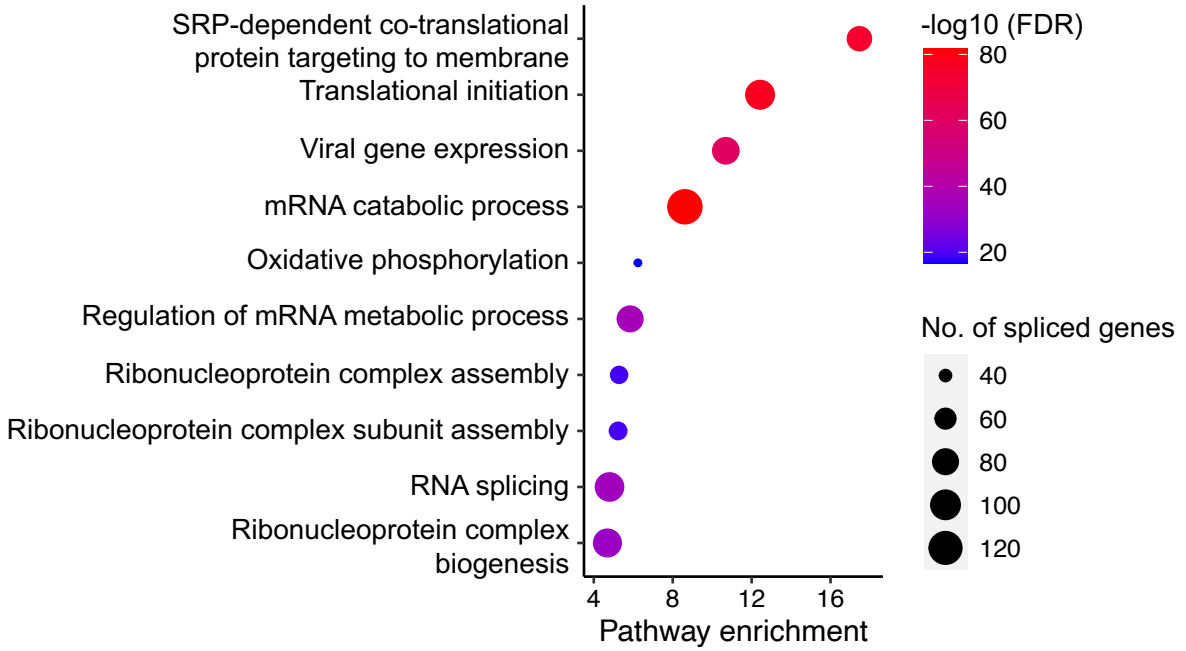


Figure 3.18: Gene ontology analysis of genes that were differentially spliced between iPSCs and endoderm cells.

NMD represents one mechanism by which splicing regulates isoforms, and by extension, gene expression (Smart et al., 2018). For a given differentially spliced event, MARVEL categorised the effect of the inclusion (splicing in) of alternative exon, i.e., alternative exon with significantly higher PSI value in endoderm cells relative to iPSCs, on the corresponding isoforms. The categories were novel splice junction (SJ), absence of coding sequence (CDS), no premature stop codon (PTC) introduced, and PTC created. Firstly, “novel SJs” are defined the splice junctions not yet reported in publicly available isoform database and therefore precluded PTC identification. Here, we used GENCODE v31 as our public isoform database. Secondly, the “no CDS” is referred to the alternative exons falling on isoforms that do not have open reading frames (ORFs), and therefore were not included for PTC identification. Thirdly, “no PTC” is referred to the preservation of the ORF when the alternative exon is introduced (spliced) into the isoform sequence. Lastly, “PTC” is referred to the introduction of PTC following the inclusion (splicing in) of alternative exon into the isoform sequence.

Across the different splicing event types, the inclusion of alternative exons in endoderm cells primarily affected isoforms with no coding sequences (“no CDS”; Figure 3.19A). When only alternative exons that affected isoforms with ORF were analysed, the inclusion (splicing in) of introns (RI) led to the highest rate of PTC creation (Figure 3.19B). This is not surprising given that the average intron length is longer than the average length of SE, A5SS, or A3SS. Therefore, it is conceivable that the probability of PTC creation is higher for RI compared to other splicing event types. Notably, the genes predicted to undergo NMD due to RI were associated with decreased expression in endoderm cells relative to iPSCs, but not for SE, A5SS, and A3SS (Figure 3.19C). This is consistent with a previous study using long-read sequencing technology that found decreased gene expression by RI but not by other splicing event types (A. D. Tang et al., 2020).

Lastly, MARVEL facilitates candidate gene selection for downstream functional studies by highlighting genes subjected to NMD on the volcano plot generated from differential gene expression analysis. Among the genes predicted to undergo splicing-mediated NMD and concurrently demonstrated decreased gene expression were *BUB3*, *HSPA4*, *EIF5*, *RPL22L1*, *DDX39B*, *SRRM1*, and *SRSF10* (Figure 3.19D).

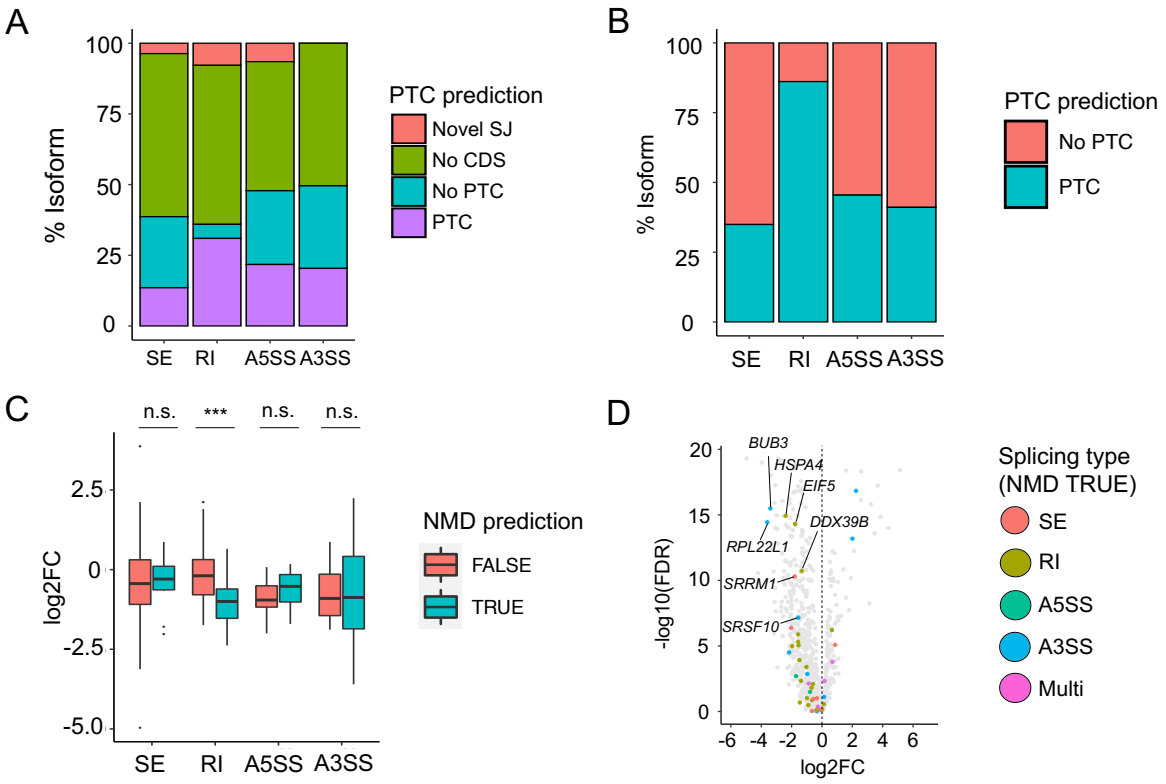


Figure 3.19: Predicting NMD from alternative exon inclusion (splicing in) in endoderm cells relative to iPSCs. **(A)** Identifying isoforms based on whether the inclusion (splicing in) of alternative exons introduced PTC, or not. **(B)** Same figure as (A) but with novel SJs and isoforms with no CDS excluded. **(C)** Association between gene expression change in log2FC in endoderm cells relative to iPSCs when alternative exons led to NMD of genes (TRUE), or not (FALSE). **(D)** Genes predicted to undergo splicing-mediated NMD annotated on the volcano plot generated from differential gene expression analysis (see Figure 3.17A). CDS: Coding sequence; FC: Fold change; NMD: Nonsense-mediated decay; PTC: Premature stop codon; SJ: Splice junction.

Taken together, MARVEL provides end-to-end support for single-cell splicing analysis of RNA-seq data generated from plate-based methods, beginning with splicing event detection and validation, PSI quantification, dimension reduction analysis, modality assignment, integrated differential gene and splicing analysis, and finally functional annotation of differentially spliced genes to enable candidate biomarker selection for downstream experimental validation.

3.5 Demonstration on droplet-based RNA-seq dataset

MARVEL was initially developed for splicing analysis of scRNA-seq data generated from plate-based platforms. This is because plate-based methods, such as Smart-seq, were pioneers in generating single-cell libraries for RNA-seq (Ramskold et al., 2012). We were also initially interested in characterising the splicing landscape in scRNA-seq generated using plate-based platforms, e.g., TARGET-seq (Rodriguez-Meira et al., 2019), from myeloid neoplasm patients with spliceosome mutations. Nevertheless, droplet-based library preparation methods have recently gained more popularity over plate-based methods because droplet-based methods were more-automated and high-throughput, i.e., thousands to tens of thousands of cells may be prepared for a given RNA-seq run with minimal hands-on during sample and library preparation (Zheng et al., 2017). Given the increasing application of droplet-based library preparation for generating scRNA-seq data, we extended MARVEL's functionalities to enable splicing analysis of RNA-seq data generated from droplet-based methods.

To demonstrate the full range of features available by MARVEL for analysing droplet-based RNA-seq dataset, we performed single-cell splicing analysis on induced pluripotent stem cells (iPSCs) and day-10 cardiomyocytes differentiated from iPSCs (Ou et al., 2021). We have chosen this dataset because the biological pathways involved in cardiomyocyte development, such as muscle-related genes, have been well characterised (Grancharova et al., 2021). Therefore, these reported biological pathways may serve as ground truth for our splicing analysis.

Differential splicing analysis between iPSCs and cardiomyocytes identified 818 differentially spliced events (FDR < 0.05 and |ΔPSI| > 5 and mean normalised log2 gene expression > 1.0), of which 575 splice junctions were significantly up-regulated in cardiomyocytes and 243 splice junctions were significantly up-regulated in iPSCs (Figure 3.20A). Examples of differentially spliced events included splice junctions of muscle-related genes *MYH10*, *ATP5F1C*, and *CBX1* (Figures 3.20B-E). Splice junctions of *MYH10* and *ATP5F1C* were significantly up-regulated in cardiomyocytes whereas splice junction of *CBX1* was significantly up-regulated in iPSCs. Pathway enrichment analysis of differentially spliced genes revealed muscle-, cardio- and nerve-related genes to be enriched among differentially spliced genes (Figure 3.20F).

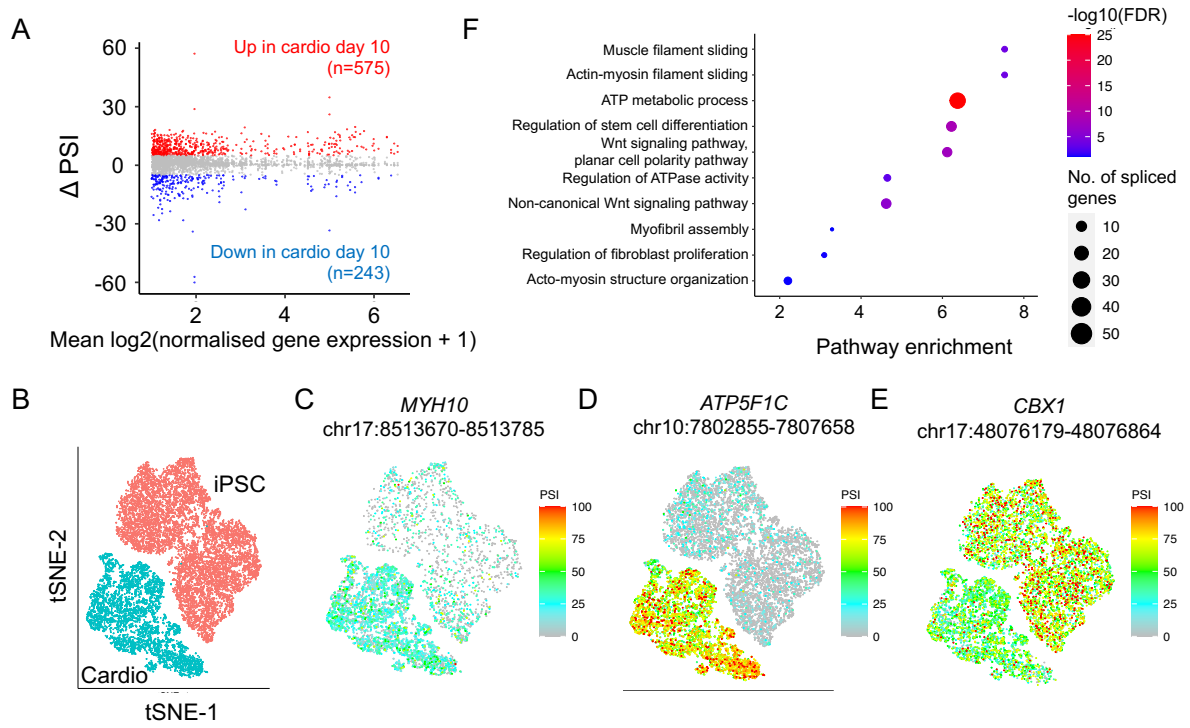


Figure 3.20: Differential splicing analysis between iPSCs and cardiomyocytes. **(A)** Volcano plot of differential splicing analysis results. **(B-E)** Representative examples of differentially spliced junctions **(F)** Gene ontology analysis of differentially spliced genes.

Next, we investigated the gene-splicing relationship between iPSCs and cardiomyocytes. To this end, we tabulated the change in gene expression profile relative to change in splice junction usage when iPSCs differentiated into cardiomyocytes.

In total, 539 genes had at least one splice junction that were differentially spliced between iPSCs and cardiomyocytes. Majority of differentially spliced genes (59%) underwent isoform-switching whereby there was no change in gene expression profile between iPSCs and cardiomyocytes but there was at least one splice junction that were differentially spliced between the two cell populations (Figure 3.21A). A small proportion of differentially spliced genes (19%) were differentially expressed between iPSCs and cardiomyocytes and the change in gene expression profile was in the same direction relative to the change in splice junction usage, i.e., coordinated gene-splicing relationship. A similar proportion of differentially spliced genes (15%) were differentially expressed between iPSCs and cardiomyocytes but the change in gene expression profile was in the opposite direction relative to the change in splice junction usage, i.e., gene-splicing relationship. Lastly, complex gene-splicing relationship constituted the smallest proportion (7%) whereby a given gene was differentially expressed between iPSCs and cardiomyocytes but its relationship with splice junction usage was a combination of coordinated, opposing, and/or isoform switching.

Representative examples of coordinated, opposing, isoform-switching, and complex gene-splicing relationships were *VIM* (Figures 3.21B and C), *UQCRH* (Figures 3.21D and E), *RBM39* (Figures 3.21F-H), and *TPM1 and TPM2* (Figures 3.21I-N), respectively.

It is noteworthy that majority of the differentially spliced genes did not occur in the same direction as their corresponding splice junctions from iPSCs to cardiomyocytes, i.e., coordinated gene-splicing relationship. Therefore, differential splicing analysis may identify differentially regulated genes that would otherwise been missed by differential gene expression analysis alone.

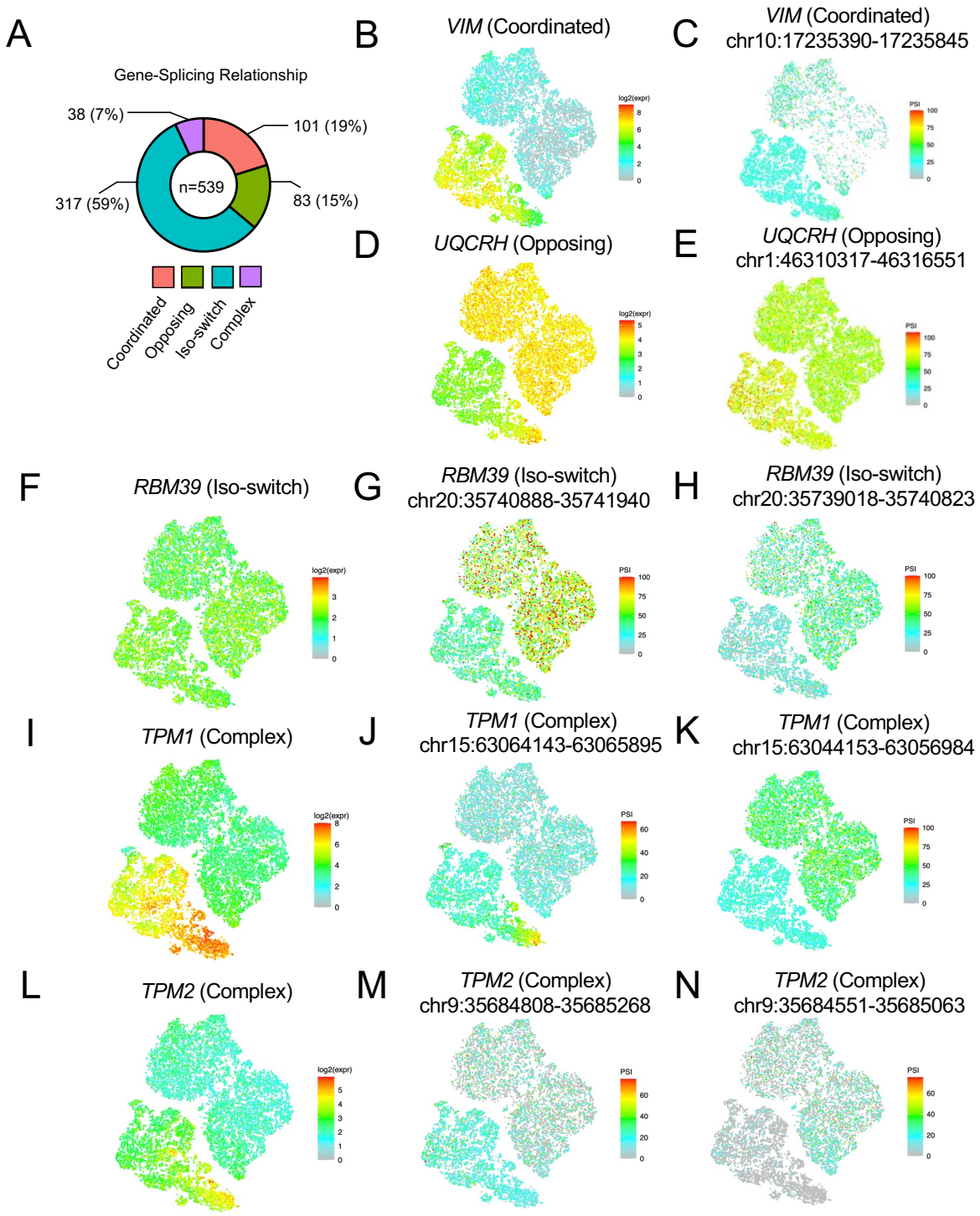


Figure 3.21: Gene-splicing relationship between iPSCs and cardiomyocytes. (A) Stratification of gene-splicing relationship into coordinated, opposing, isoform-switching, and complex. **(B-K)** Representative examples of **(B-C)** coordinated, **(D-E)** opposing, **(F-H)** isoform-switching, and **(I-N)** complex gene-splicing relationship.

To illustrate and fully appreciate the intricate relationship between gene expression and splicing, we have scrutinised and traced the relationship between *TPM2* gene expression and its corresponding usage of all expressed splice junctions from iPSCs to cardiomyocytes at day-2, -4, and -10. We chose *TPM2* for illustration here because this gene demonstrated a complex relationship with its splice junctions (Figures 3.21L-N).

We observed progressive increased in *TPM2* gene expression and increased in number of cells expressing this gene from iPSCs to later stages of cardiomyocyte development (Figure 3.22A). In total, 10 splice junctions were expressed in least one cell population and their usage did not always change in the same direction as the corresponding gene expression (Figure 3.22B). For example, SJ-1 was more highly expressed in iPSCs and earlier stages of cardiomyocytes compared to day-10 cardiomyocytes (Figure 3.22C). On the other hand, SJ-2 was not differentially spliced across the four different developmental stages (Figure 3.22D). In contrast, SJ-3 showed progressive increased in usage from iPSCs to later stages of cardiomyocytes, consistent with the changes in gene expression profile (Figure 3.22E).

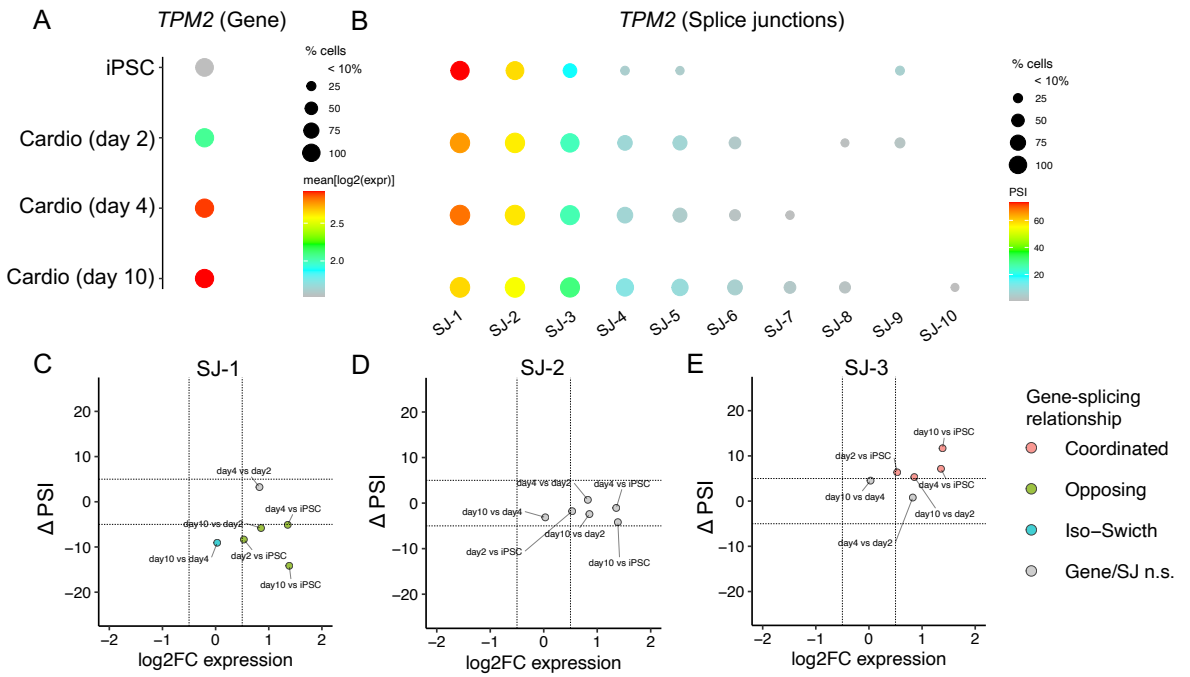


Figure 3.22: Gene-splicing relationship of *TPM2* across the different cardiomyocyte developmental stages. (A-B) Changes in **(A)** gene expression and **(B)** splice junction usage from iPSCs to cardiomyocytes at day-2, -4, and -10. **(C-E)**

Relative change of splice junction usage change against gene expression change for all pair-wise comparison of the different cell populations for **(C)** SJ-1, **(D)** SJ-2, and **(E)** SJ-3.

One main difference between the RNA-seq data generated from plate-based compared to droplet-based methods is the pervasive 3'/5'-bias of coverage distribution in the latter (Zheng et al., 2017). We observed SJ-1, SJ-2, and SJ-3 to have relatively higher expression compared to the other splice junctions (Figures 3.22B). Moreover, the overall expression of SJ-1 was higher compared to SJ-2, followed by SJ-3. This hints at 3'-bias coverage.

To confirm this 3'-bias coverage, MARVEL plotted the positions of user-specified splice junctions relative to all corresponding isoforms of the gene. Here, we observed SJ-1, which had the highest overall expression, to be located at the most 3'-end of the isoforms (Figure 3.23). SJ-2 and SJ-3, which had the 2nd and 3rd highest overall expression, respectively, were located at the 2nd and 3rd most 3'-end of the isoforms.

Therefore, several considerations would need to be taken account for splicing analysis of RNA-seq data generated from droplet-based sequencing. Firstly, splicing analysis may only be feasible at the 3'/5'-end of the isoforms. Secondly, due to progressive decline in splice junction expression from the ends of the isoforms, it is only meaningful to compare the splice junction usage for the same splice junction across different cell populations, SJ-1 in iPSCs vs SJ-1 in cardiomyocytes. However, it is not meaningful to compare the splice junction usage of different splice junction within the same cell population or across different cell populations, e.g., SJ-1 vs SJ-2. This is because the splice junction usage near to the end of the isoform will likely be consistently higher compared to splice junction usage away from the end of the isoform.

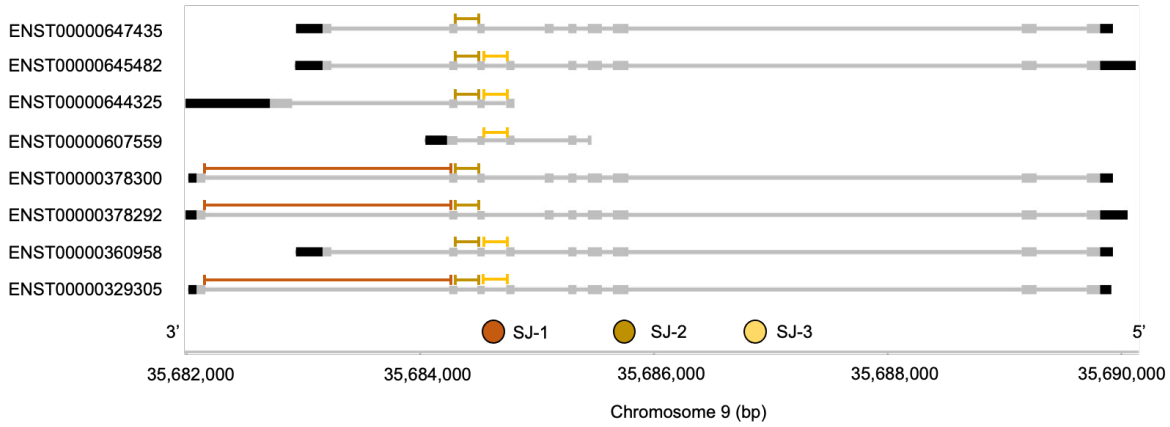


Figure 3.23: Location of SJ-1, SJ-2, and SJ-3 relative to all isoforms of *TPM2* reported in GENCODE v31.

Finally, we assessed the MARVEL's ability to scale to larger 10x Genomics dataset. To this end, we performed differential splicing analysis on >100k single cells derived from brain tissues of autism spectrum disorder (ASD) patients and healthy controls (Velmeshev et al., 2019). This dataset consists of 11 neuronal cell populations (L2/3, L4, L5/6, L5/6-CC, IN-PV, IN-SST, IN-VIP, IN-SV2C, Neu-NRGN-I, Neu-NRGN-II, Neu-mat) and 6 non-neuronal cell populations (AST-PP, AST-FB, microglia, oligodendrocytes, OPC, endothelial) (Figure 3.24).

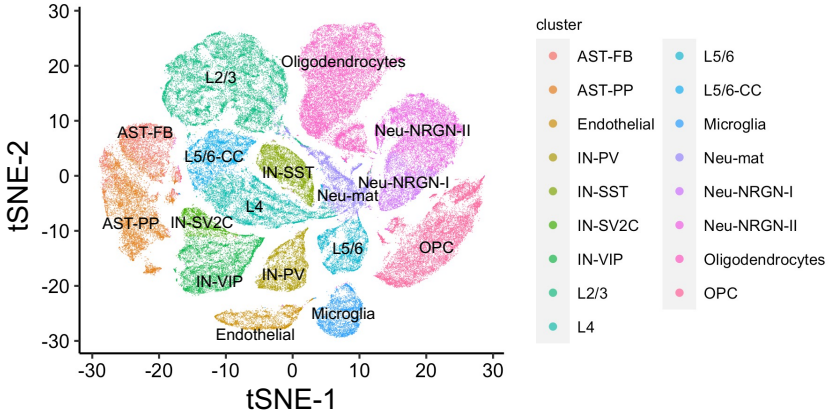


Figure 3.24: Seventeen cell populations included in this analysis. tSNE coordinates and cell type annotation were derived from Supplementary Table 2 of the original study.

Differential splicing analysis identified 691 and 1,903 splice junctions to be up- and down-regulated, respectively, among neuronal cell types of ASD relative to

healthy controls (Figure 3.25A). Furthermore, 297 and 173 splice junctions were up- and down-regulated, respectively, among non-neuronal cell types of ASD relative to healthy controls (Figure 3.25B). Notably, *SYT1*, which a canonical marker gene for excitatory neurons, was differentially spliced. Specifically, the splice junction chr12:78865110:78977798 was significantly spliced out among ASD patients relative to controls (Figure 3.25C and D).

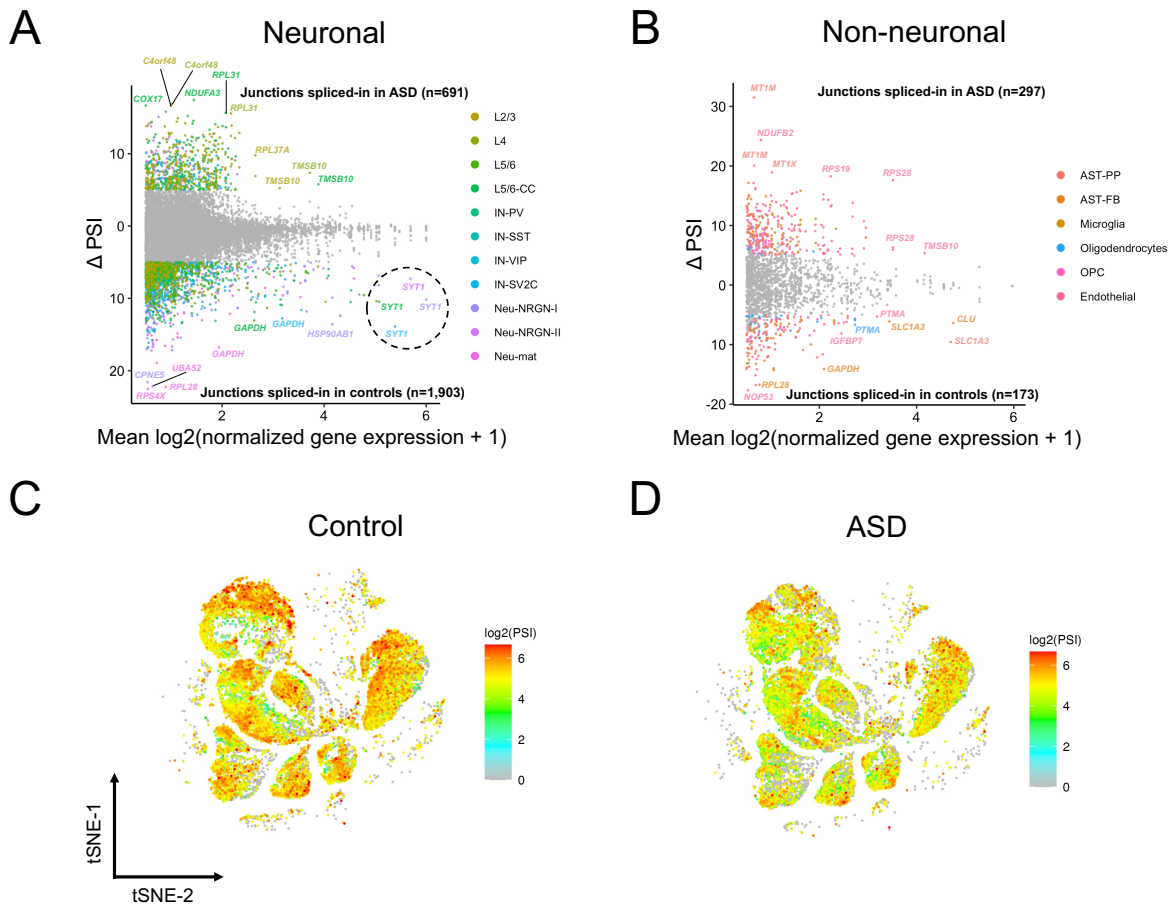


Figure 3.25: Differential splicing analysis between ASD patients and healthy controls. (A-B) Volcano plot of differential splicing analysis among **(A)** neuronal and **(B)** non-neuronal cell types. **(C-D)** *SYT1* chr12:78865110:78977798 splice junction expression in **(C)** controls and **(D)** ASD patients.

Next, we assessed if the list of differentially spliced genes identified by MARVEL were relevant to ASD. To this end, we assessed the overlap of differentially spliced genes identified by MARVEL with that of ASD-related genes reported by

Simons Foundation Autism Research Initiative (SFARI) database (Abrahams et al., 2013). In total, 609 genes were identified as differentially spliced by MARVEL, of which, 49 overlapped with SFARI (*P* = 5.6e-24) (Figure 3.26A). Biological pathways enriched among differentially spliced genes were RNA splicing, pathways associated with synapses, axons, and dendrites, and tau-protein kinase activity (Figure 3.26B).

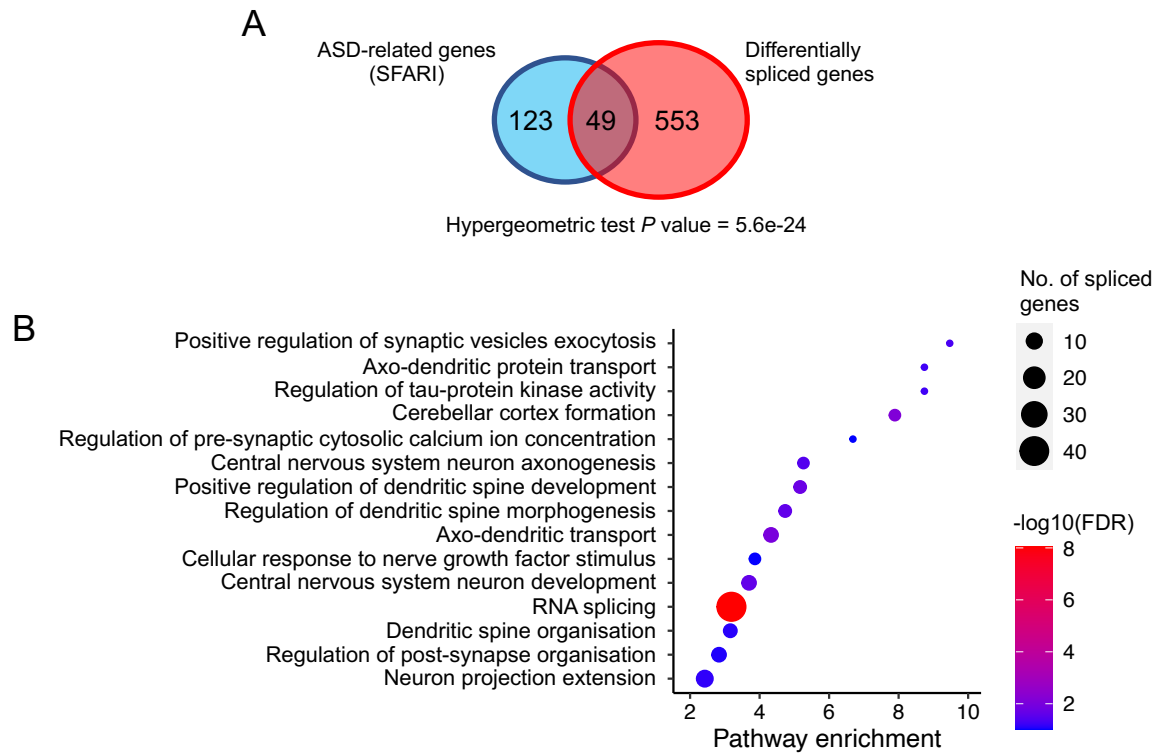


Figure 3.26: Assessing the enrichment of ASD-related genes and pathways among differentially spliced genes identified by MARVEL. (A) Overlap between ASD-related genes and differentially spliced genes. **(B)** Pathway enrichment analysis of differentially spliced genes.

Lastly, we assessed the computational efficiency of MARVEL for differential splicing analysis of this large 10x Genomics dataset. Overall, across the 17 cell types included for analysis, L5/6-CC demonstrated the highest number of differentially spliced junctions (Figure 3.27A). The median time taken to complete differential splicing analysis for a given cell population was 1 minute 34 seconds, and the total time taken to complete differential splicing analysis for all cell populations was 36.1 minutes (Figure 3.27B).

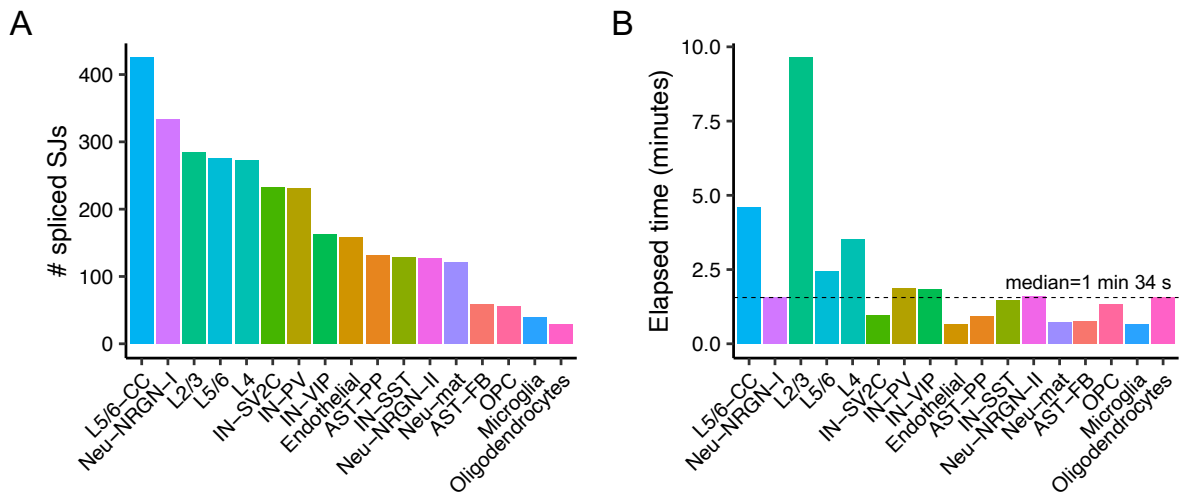


Figure 3.27: Computational efficiency evaluation of differential splicing analysis by MARVEL. (A) Number of differentially spliced junctions identified for each cell population. **(B)** Total time taken to complete differentially spliced junctions identified for each cell population.

Taken together, MARVEL provides end-to-end support for single-cell splicing analysis of RNA-seq data generated from droplet-based methods, beginning with splicing junction validation, splice junction usage quantification, integrated differential gene and splicing analysis, and finally functional annotation of differentially spliced genes to enable candidate biomarker selection for downstream experimental validation. We also demonstrated that MARVEL is able to scale to large 10x Genomics dataset (>100k cells).

Nevertheless, there are additional features that may be incorporated into MARVEL to enhance single-cell alternative splicing analysis. For example, MARVEL may integrate gene- and splicing-level information for dimension reduction analysis. Additionally, while we demonstrated the utility of MARVEL on ~100k cells, it would be of particular interest to assess MARVEL's ability to scale to millions of cells. Lastly, MARVEL was developed for analysing individual exon-level splicing events from short-read scRNA-seq. It may be necessary to follow-up with long-read scRNA-seq to deconvolute the combinatorial patterns of isoform usage for differentially spliced genes identified from MARVEL.

4 VALERIE: A novel computational tool for visual validation of alternative splicing events at single-cell resolution

4.1 Validation of previously reported *PKM* splicing event

To benchmark VALERIE, we have a selected an alternative splicing event that have been identified in scRNA-seq and validated using single-molecule fluorescence *in situ* hybridization (smFISH) (Song et al., 2017). We selected this dataset also because it included both single cells and matched-bulk RNA-seq data. This will allow us to compare the visualisation of alternative splicing events at the bulk and single-cell level. In this dataset, *PKM* exon 9 and 10 were mutually exclusive exons that demonstrated differential exon usage during induced pluripotent stem cells (iPSCs) differentiation into neural progenitor cells (NPCs) or motor neurons (MNs).

We first demonstrated the limitations of using a current genome browser (Thorvaldsdottir, Robinson, & Mesirov, 2013) to visualise alternative splicing events at the bulk level (Figure 4.1A). We observed iPSCs to have higher coverage at exon 10 compared to exon 9. Similarly, we observed NPCs to have higher coverage at exon 10 compared to exon 9. On the other hand, MN bulk samples corresponding to rows 1 and 3 had higher coverage at exon 9 but MN bulk sample corresponding to row 2 had higher coverage at exon 10. This hinted the presence of heterogeneous cell populations underlying MN bulk samples that may not be resolved visually using bulk RNA-seq.

Next, we demonstrated the limitations of using a current genome browser (Thorvaldsdottir et al., 2013) to visualise alternative splicing events at the single-cell level (Figure 4.1B). We selected five representative single cells from each of the three cell populations for visualising on the genome browser. We observed iPSCs had higher coverage at exon 10 compared to exon 9 across all the five single cells while there was virtually no coverage at exon 9. This suggests iPSCs exclusively expressed exon 10. Similarly, we observed NPCs to have higher coverage at exon 10 compared to exon 9 across all the five single cells while there was virtually no coverage at exon 9, with the exception of NPC at row number 10. This suggests that NPC primarily expressed exon 10. We observed three MNs to have coverage at exon 10 but virtually no coverage at exon 9, whereas one MN had coverage at exon 9 but virtually no coverage at exon 10. This suggests that MN potentially consist of two disparate cell populations, in which one primarily expressed exon 10 while another primarily

expressed exon 9. It is also noteworthy that one cell did not have any coverage in *PKM*, thus precluding any inference to be drawn on differential exon usage.

Therefore, current genome browsers limit the number of single cells in which a user can visually inspect for a given alternative splicing event of interest. Moreover, the cells selected for visual inspection may not always be representative of their corresponding cell populations. As a consequence, the comprehensive alternative splicing profile across all cells may not be captured using this approach and the extent of dropouts, i.e., cells with low coverage and hence represent missing data points, may also not be revealed.

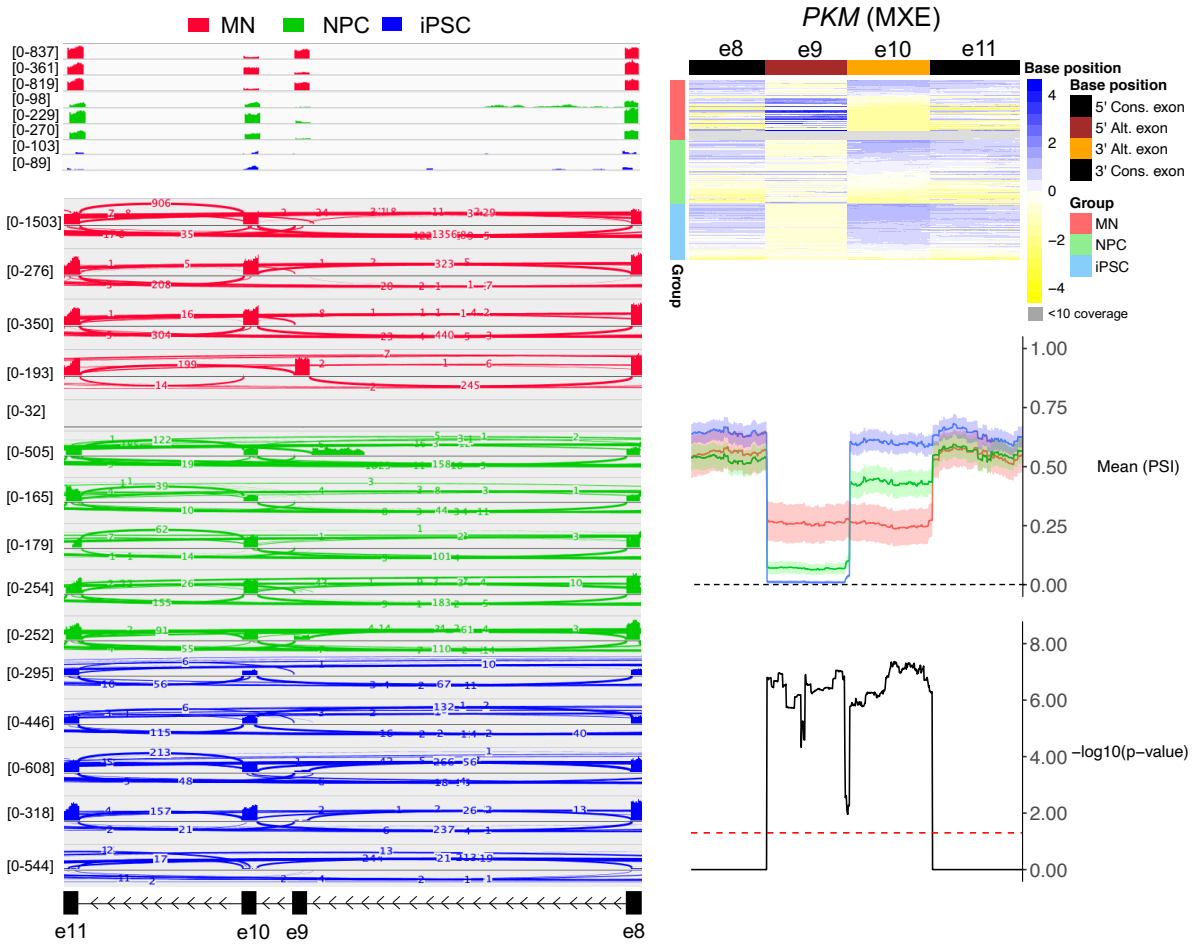


Figure 4.1: Comparison of visualising *PKM* mutually exclusive exons 9 and 10 using IGV and VALERIE. (A) Visualising alternative splicing profile using coverage information of bulk sample in IGV. **(B)** Visualising alternative splicing profile using coverage information of selected single cells in IGV. **(C-E)** Visualising alternative splicing profile using PSI information of all single cells in VALERIE. **(C)** PSI values

displayed for all base positions (columns) corresponding to the alternative splicing event across all single cells (rows). **(D)** PSI values aggregated by cell type. **(E)** *P* values derived from assessing the differences in PSI values across the different cell populations. Differences were assessed using Kruskal-Wallis test and *P* values were adjusted for multiple testing across the base positions using Bonferroni correction. IGV: Integrative Genome Browser.

To address the limitations of current genome browsers on visualising single-cell alternative splicing events, we developed VALERIE to enable us to fully capture and appreciate the alternative splicing heterogeneity across all single cells included in a study (Figure 4.1C).

Firstly, MARVEL presents the percent spliced-in (PSI) values for each base position straddling both constitutive and alternative exons for each single cell included in the original study. The degree of splicing is indicated with the yellow-blue colour scheme whereby higher intensity of yellow represents lower PSI values while higher intensity of blue represents higher PSI values. Here, we observed iPSCs to almost exclusively express exon 10. Approximately 50% of NPCs were observed to express exon 10 whereas a smaller proportion of NPCs expressed 9 or neither exon 9 nor 10. On the other hand, MNs predominantly expressed exon 9. Notably, approximately 20% of MNs had no coverage at this region. This may be due to the cells not expressing the *PKM* gene or technical dropouts that led to insufficient coverage across this genomic locus.

Secondly, MARVEL presented the aggregated PSI values for each cell population, i.e., pseudo-bulk. Using this approach, we observed progressive increased in exon 9 expression from iPSCs to NPCs and then to MNs. Conversely, we observed progressive decreased in exon 10 expression from iPSCs to NPCs and then to MNs. This inverse relationship between exon 9 and 10 is a hallmark feature of mutually exclusive exons.

Thirdly, MARVEL assessed the differences in per-base PSI values across the three cell populations. We observed the PSI values to be significantly different across the cell populations at base positions corresponding to the alternative exons. On the other hand, there were no significant differences in PSI values at base positions corresponding to the constitutive exons, as expected. This statistical assessment may

provide an objective interpretation on whether the alternative exons are differentially spliced across the different cell populations.

4.2 Validation of previously reported *Mbp* splicing event

Another alternative splicing event that has been previously identified and experimental validated is *Mbp* exon 2 (Falcao et al., 2018). This exon has been found to be differentially spliced in mice induced with experimental autoimmune encephalomyelitis (EAE) compared to control mice, and this splicing event has been validated using quantitative polymerase chain reaction (qPCR). VALERIE profiling of this *Mbp* exon 2 showed this exon to be differentially spliced between clusters MOL1/2 from EAE mice and cluster MOL2 Cntrl-A from control mice (Figure 4.2).

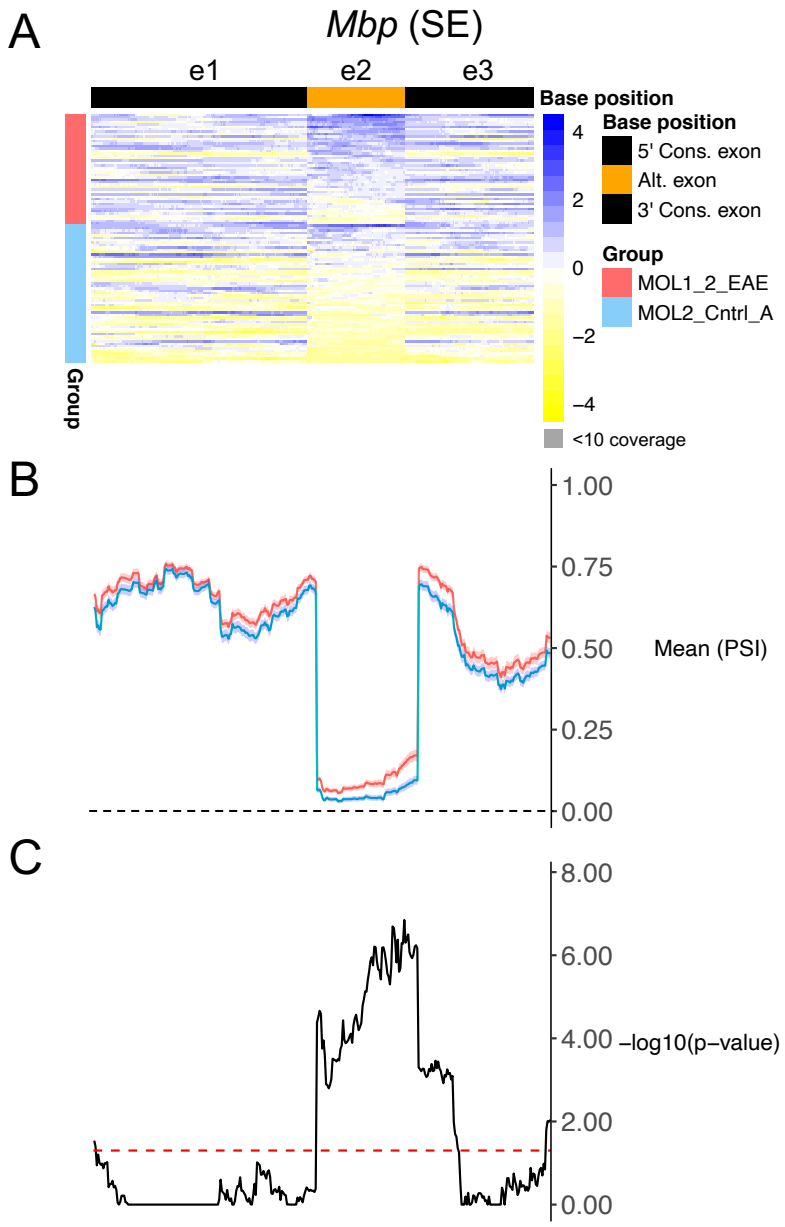


Figure 4.2: Visualising *Mbp* exon 2 alternative splicing profile in VALERIE. (A) PSI values displayed for all bases corresponding to the alternative splicing event across all single cells. **(B)** PSI values aggregated by cell type. **(C)** *P* values derived from assessing the differences in PSI values across the different cell populations. Differences were assessed using Wilcoxon rank-sum test and *P* values were adjusted for multiple testing across the base positions using Bonferroni correction. MOL: Mature oligodendrocytes.

It is noteworthy that the PSI profile in this example (Figure 4.2) appears scant or sparse on the heatmap compared to that in Figure 4.1. This is because the RNA-seq data generated from this study was relatively short, i.e., 50bp in single-end (SE) mode, while the RNA-seq data in previous analysis in Figure 4.1 was generated using relatively longer sequencing reads, i.e., 150bp in paired-end (PE) mode. Therefore, these very short reads do not always span the entire length of the exons nor do the reads always span the exon-exon junctions, i.e., splice junctions. As a consequence, inferring PSI values from coverage information derived from splice junctions alone from current genome browser may be insufficient. VALERIE estimates and displays the PSI values across all the base positions corresponding to the alternative splicing event and therefore is able to visualise and assess the differences in PSI values across the difference cell populations beyond the splice junctions, i.e., across the exon bodies.

4.3 Demonstration on novel splicing events

We demonstrated the ability of VALERIE to visualise and validate previously reported alternative splicing events, namely *PKM* mutually exclusive exons (MXEs) splicing event in human (Song et al., 2017) and *Mbp* skipped-exon (SE) splicing event in mouse model (Falcao et al., 2018). Here, we demonstrate the application of VALERIE on visually validating alternative splicing event detected when induced pluripotent stem cells (iPSCs) differentiated into endoderm cells in Section 3.4. Furthermore, aside from visually validating SE and MXE as in our benchmarking analysis for *Mbp* and *PKM*, respectively, we will demonstrate the visual validation of retained intron (RI), and alternative 5' and 3' splice sites (A5SS, A3SS).

A *SNRPN* SE splicing event was observed to have significantly decreased percent spliced-in (PSI) values in endoderm cells relative to iPSCs (Figure 4.3A-B). The modality change was implicit whereby the PSI distribution transformed from excluded dispersed to primary from iPSCs to endoderm cells. Visual inspection of this alternative splicing event using VALERIE similarly showed decreased in alternative exon usage in endoderm cells relative to iPSCs (Figure 4.3C). Therefore, this alternative splicing was successfully visually validated by VALERIE and may be considered as a true differentially spliced event (true positive) detected by MARVEL. *Snrpn* has been shown to be an imprinted gene in embryonic stem cells, specifically

this gene is methylated at the maternal alleles (Wianny et al., 2016). This may have implication during establishment of stem cell lines.

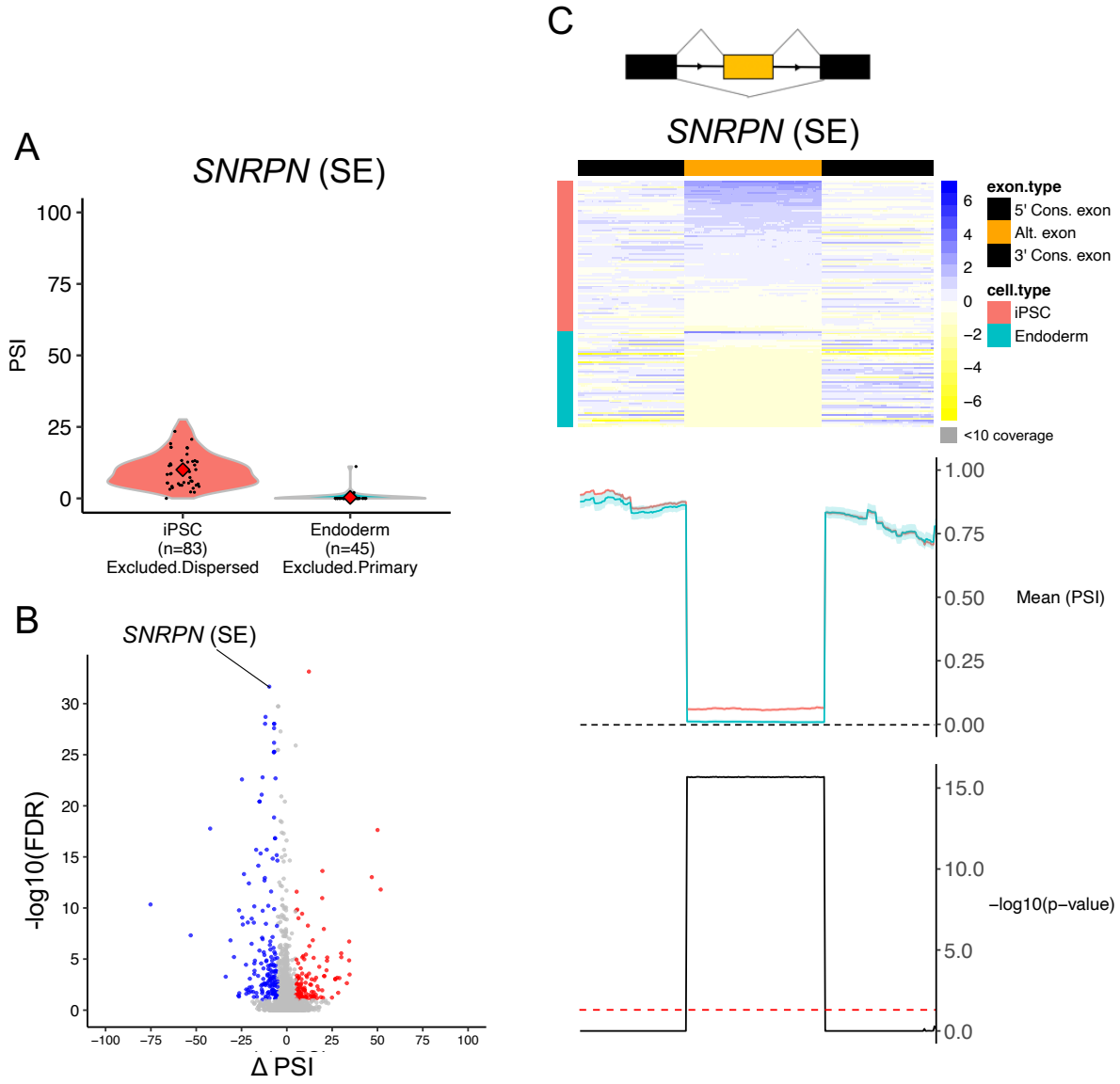


Figure 4.3: *SNRPN* SE splicing event (chr15:24962114:24962209:+@chr15:24967029:24967152:+@chr15:24967932:24968082) identified by MARVEL and visually validated using VALERIE. (A) PSI distributions and modality change of the splicing event from iPSCs to endoderm cells. **(B)** Splicing event annotated on the volcano plot that includes all splicing event included for differential splicing analysis. Blue represents splicing events with changes of PSI values of < -5 and FDR < 0.10 in endoderm cells relative to iPSC. Red represents splicing events with changes of PSI values of > 5 and FDR < 0.10 in

endoderm cells relative to iPSC. **(C)** Visual inspection of alternative splicing event using VALERIE. Differences in PSI values across the two cell populations were assessed using Kolmogorov-Smirnov test and adjusted for multiple testing across the base positions using Bonferroni correction.

A *TPM2* MXE splicing event was observed to have significantly decreased PSI values of its 5' alternative exon in endoderm cells relative to iPSCs (Figure 4.4A-B). The modality change was explicit whereby the PSI distribution transformed from middle to excluded dispersed from iPSCs to endoderm cells. Visual inspection of this alternative splicing event using VALERIE similarly showed decreased in the 5' alternative exon usage in endoderm cells relative to iPSCs (Figure 4.4C). Therefore, this alternative splicing was successfully visually validated by VALERIE and may be considered as a true differentially spliced event (true positive) detected by MARVEL. Genetic variants have been identified in iPSCs derived from patients with rare muscular disorders (Ma et al., 2019). This is not surprisingly given TPM2 role in muscle movements.

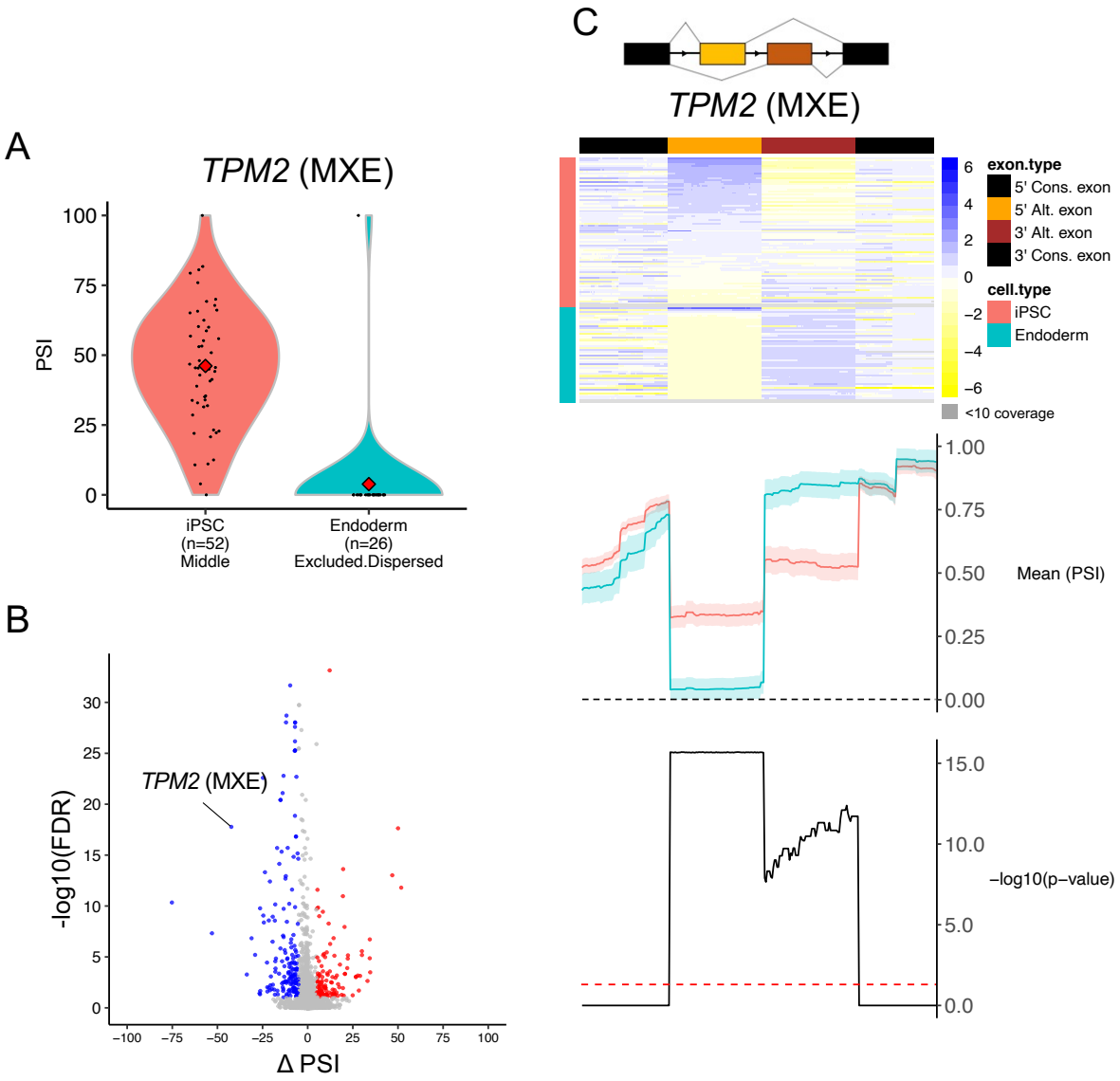


Figure 4.4: *TPM2* MXE splicing event (chr9:35685269:35685339:-@chr9:35685064:35685139:-@chr9:35684732:35684807:-@chr9:35684488:35684550) identified by MARVEL and visually validated using VALERIE. (A) PSI distributions and modality change of the splicing event from iPSCs to endoderm cells. **(B)** Splicing event annotated on the volcano plot that includes all splicing event included for differential splicing analysis. Blue represents splicing events with changes of PSI values of < -5 and FDR < 0.10 in endoderm cells relative to iPSC. Red represents splicing events with changes of PSI values of > 5 and FDR < 0.10 in endoderm cells relative to iPSC. **(C)** Visual inspection of alternative splicing event using VALERIE. Differences in PSI values across the two cell populations were assessed using Kolmogorov-Smirnov test and adjusted for multiple testing across the base positions using Bonferroni correction.

It is noteworthy that while the 5' alternative exon usage was decreased in endoderm cells relative to iPSCs, the 3' alternative exon usage was increased in endoderm cells relative to iPSCs. This inverse relationship between 5' and 3' alternative exon is a hallmark feature of MXE splicing events.

A *RPL30* RI splicing event was observed to have significantly decreased PSI values in endoderm cells relative to iPSCs (Figure 4.5A-B). The modality change was implicit whereby the PSI distribution transformed from excluded dispersed to primary from iPSCs to endoderm cells. Visual inspection of this alternative splicing event using VALERIE similarly showed decreased in intron retention in endoderm cells relative to iPSCs (Figure 4.5C). Therefore, this alternative splicing was successfully visually validated by VALERIE and may be considered as a true differentially spliced event (true positive) detected by MARVEL. RPL30 itself is a splicing factor and is involved in spliceosome assembly (Bragulat et al., 2010).

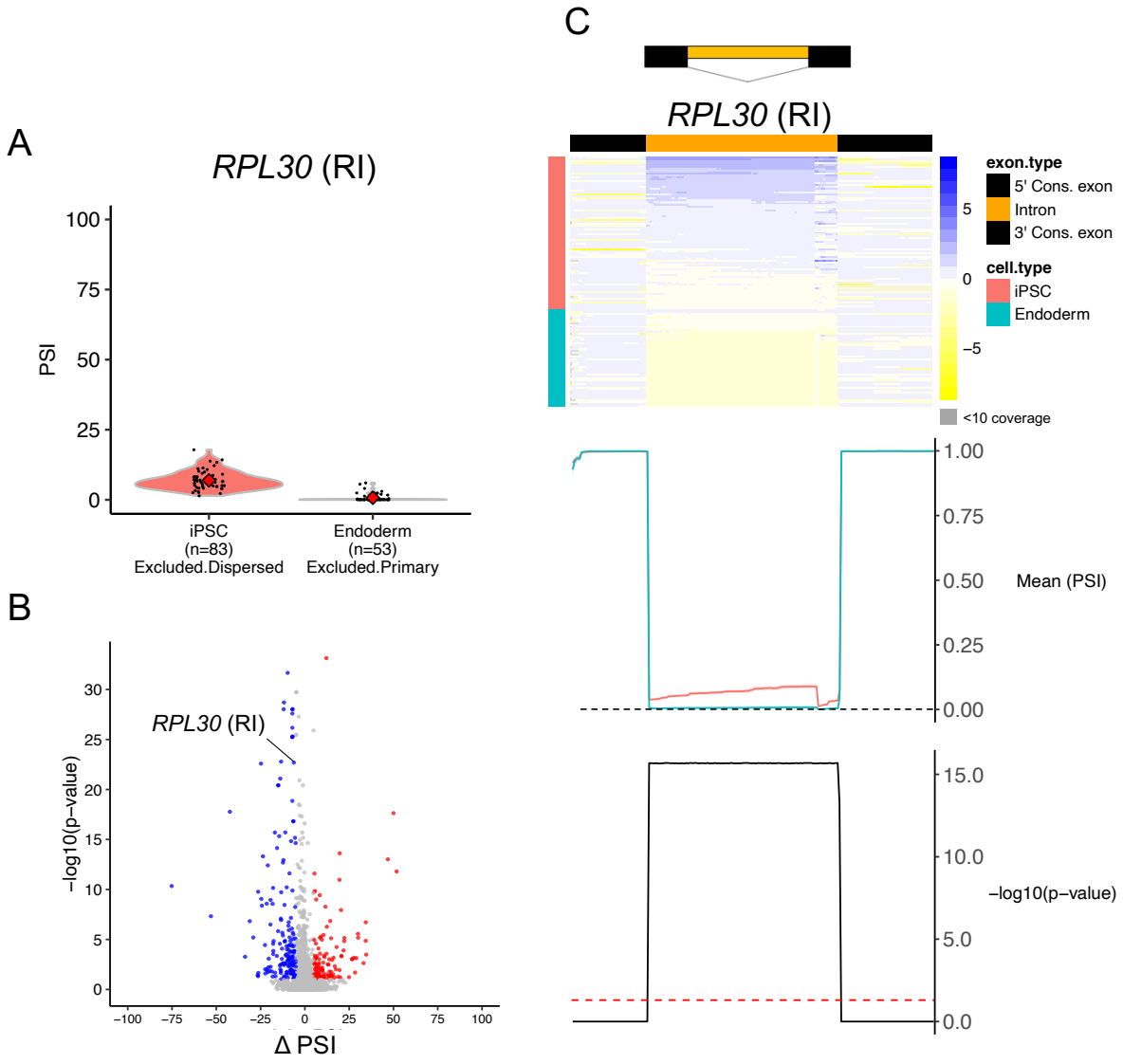


Figure 4.5: *RPL30* RI splicing event (chr8:98045550:98045508:-@chr8:98045399:98045347) identified by MARVEL and visually validated using VALERIE. (A) PSI distributions and modality change of the splicing event from iPSCs to endoderm cells. **(B)** Splicing event annotated on the volcano plot that includes all splicing event included for differential splicing analysis. Blue represents splicing events with changes of PSI values of < -5 and FDR < 0.10 in endoderm cells relative to iPSC. Red represents splicing events with changes of PSI values of > 5 and FDR < 0.10 in endoderm cells relative to iPSC. **(C)** Visual inspection of alternative splicing event using VALERIE. Differences in PSI values across the two cell populations were assessed using Kolmogorov-Smirnov test and adjusted for multiple testing across the base positions using Bonferroni correction.

There are two features of this alternative splicing event revealed by VALERIE that are noteworthy. Firstly, we noticed a potential 3' alternative splice site at the end of the intron. This cryptic A3SS has small, but significantly, higher PSI values in iPSCs compared to endoderm cells. This suggests that this cryptic A3SS was differentially spliced between the two cell populations. The presence of two simultaneous splicing event type in the same genomic locus underscores the complexity of alternative splicing. Secondly, the PSI values along the intron decreased from the 3'- to 5'-end. This may reflect a decreased in coverage from the 3'- to 5'-end of the transcript. This may be due to the 3'-bias inherently present in single-cell library preparation methods using polyA tail capturing of mRNA molecules.

A *RPL8* A5SS splicing event was observed to have significantly decreased PSI values in endoderm cells relative to iPSCs (Figure 4.6A-B). The modality change was implicit whereby the PSI distribution transformed from excluded dispersed to primary from iPSCs to endoderm cells. Visual inspection of this alternative splicing event using VALERIE similarly showed decreased A5SS in endoderm cells relative to iPSCs (Figure 4.6C). Therefore, this alternative splicing was successfully visually validated by VALERIE and may be considered as a true differentially spliced event (true positive) detected by MARVEL. A role for RPL8 in regulating telomere length has been implicated recently (van der Spek et al., 2020).

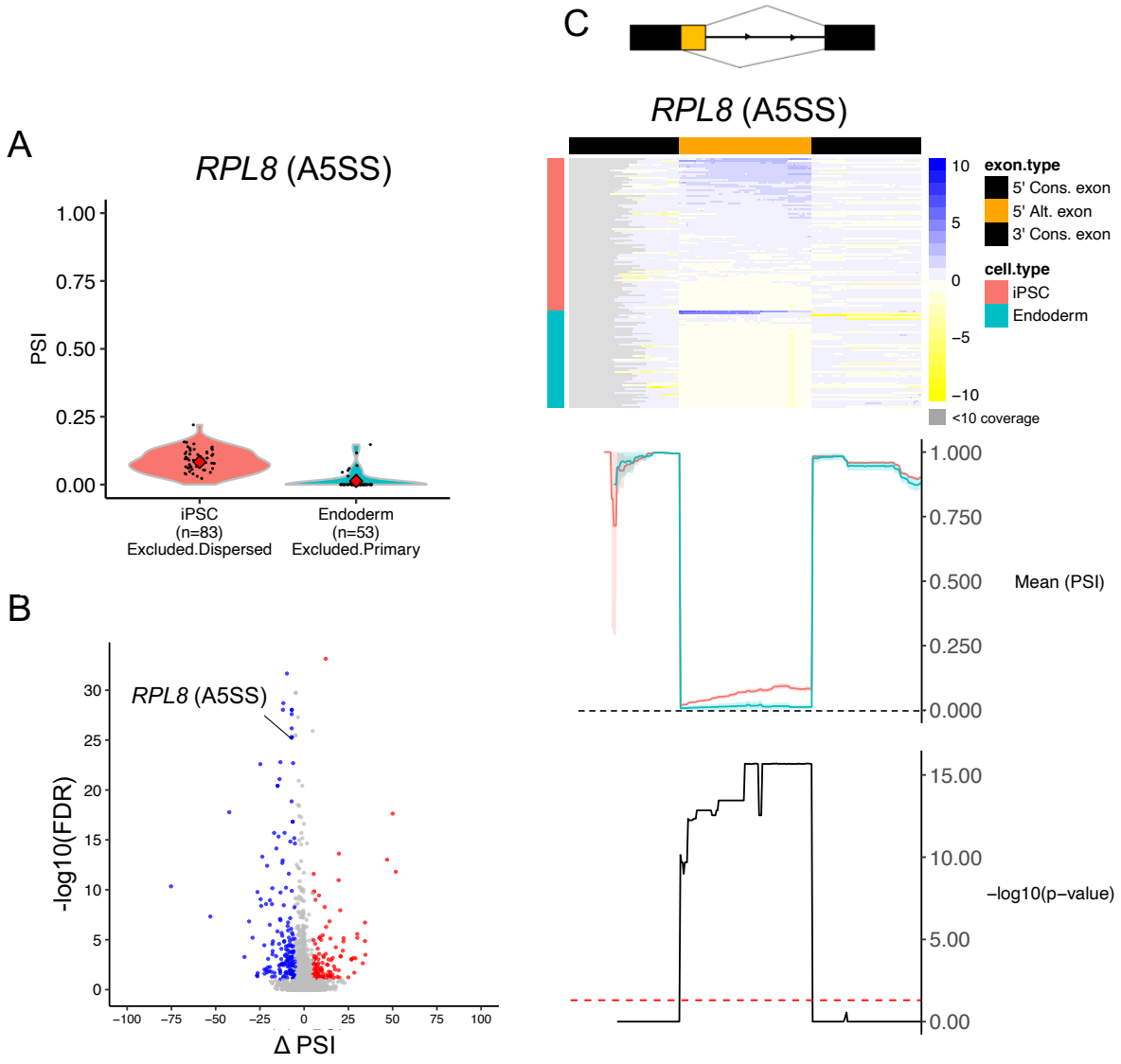


Figure 4.6: *RPL8* A5SS splicing event (chr8:144792587:144792245|144792366:-@chr8:144791992:144792140) identified by MARVEL and visually validated using VALERIE. (A) PSI distributions and modality change of the splicing event from iPSCs to endoderm cells. **(B)** Splicing event annotated on the volcano plot that includes all splicing event included for differential splicing analysis. Blue represents splicing events with changes of PSI values of < -5 and FDR < 0.10 in endoderm cells relative to iPSC. Red represents splicing events with changes of PSI values of > 5 and FDR < 0.10 in endoderm cells relative to iPSC. **(C)** Visual inspection of alternative splicing event using VALERIE. Differences in PSI values across the two cell populations were assessed using Kolmogorov-Smirnov test and adjusted for multiple testing across the base positions using Bonferroni correction.

There are two features of this alternative splicing event revealed by VALERIE that are noteworthy. Firstly, the PSI values along the alternative exon decreased from the 3'- to 5'-end. This may reflect a decreased in coverage from the 3'- to 5'-end of the transcript, not unlike that observed early in Figure 4.5C. Secondly, the lack of coverage (grey regions) and uneven coverage of the 5'-end of the transcript suggests difficulty in uniformly capturing the 5'-end of the transcripts during single-cell library preparation. This phenomenon has also been observed in RNA-seq data generated from long-read sequencing using Nanopore and PacBio (Byrne et al., 2017; Gupta et al., 2018; Legnini, Alles, Karaiskos, Ayoub, & Rajewsky, 2019).

A *PRRC2C* A3SS splicing event was observed to have significantly increased PSI values in endoderm cells relative to iPSCs (Figure 4.7A-B). The modality change was restricted whereby the PSI distribution remained included dispersed from iPSCs to endoderm cells. Visual inspection of this alternative splicing event using VALERIE similarly showed increased A3SS in endoderm cells relative to iPSCs (Figure 4.7C). Therefore, this alternative splicing was successfully visually validated by VALERIE and may be considered as a true differentially spliced event (true positive) detected by MARVEL. Prrc2c has shown to be involved in recognition or processing of poly-A tails and cap-dependent translation by interacting with Nat1 (Sugiyama et al., 2017).

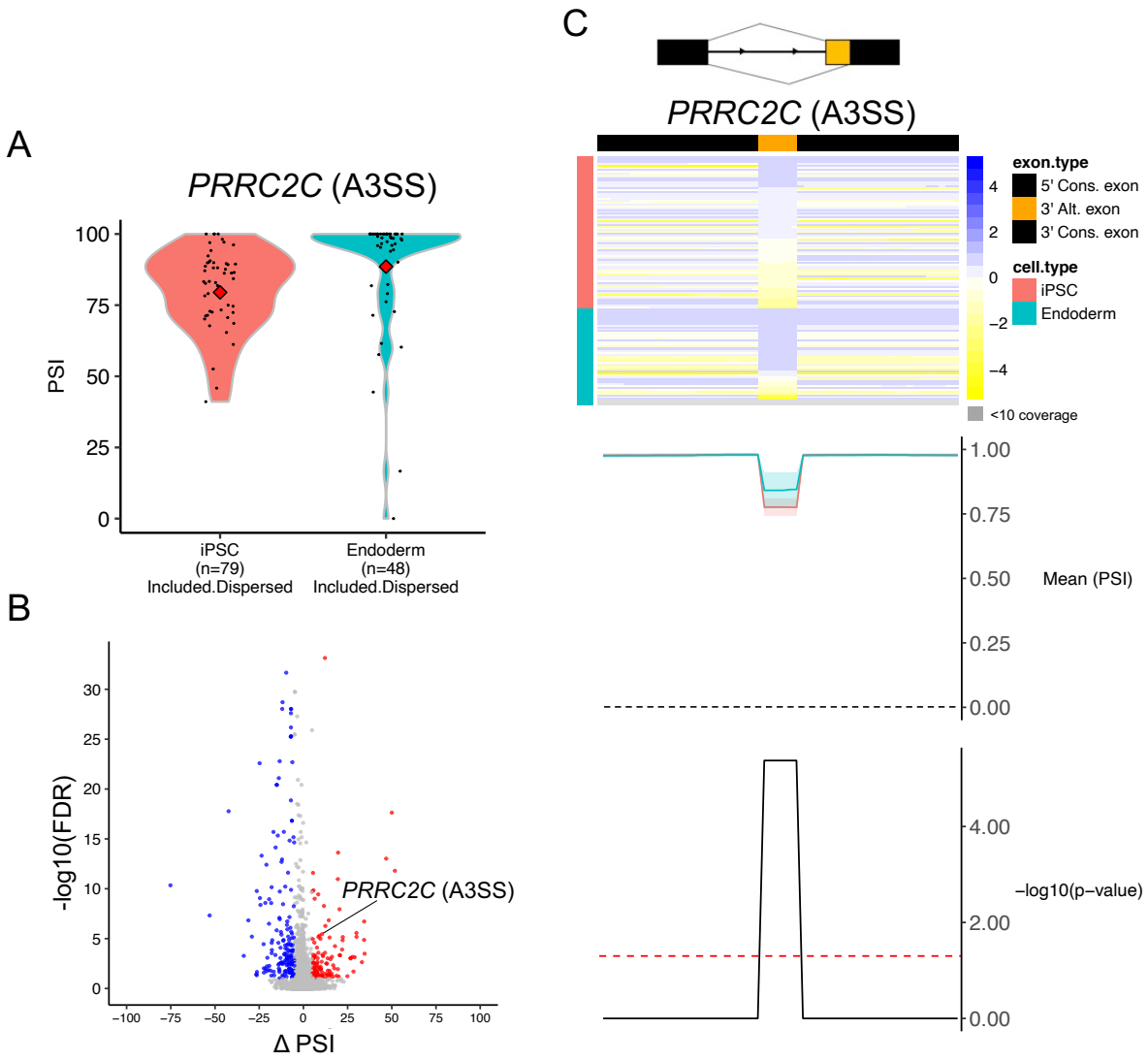


Figure 4.7: *PRRC2C* A5SS splicing event (chr1:171512032:171512200:+@chr1:171512995|171513001:171513172) identified by MARVEL and visually validated using VALERIE. (A) PSI distributions and modality change of the splicing event from iPSCs to endoderm cells. **(B)** Splicing event annotated on the volcano plot that includes all splicing event included for differential splicing analysis. Blue represents splicing events with changes of PSI values of < -5 and FDR < 0.10 in endoderm cells relative to iPSC. Red represents splicing events with changes of PSI values of > 5 and FDR < 0.10 in endoderm cells relative to iPSC. **(C)** Visual inspection of alternative splicing event using VALERIE. Differences in PSI values across the two cell populations were assessed using Kolmogorov-Smirnov test and adjusted for multiple testing across the base positions using Bonferroni correction.

4.4 Summary of visualisation features

Here we have shown that VALERIE successfully visually validated previously experimentally validated alternative splicing events, namely *PKM* (Song et al., 2017) and *Mbp* splicing events (Falcao et al., 2018). We further demonstrated the application of VALERIE on novel splicing events detected during iPSCs differentiation into endoderm cells from Section 3.4. These splicing events constituted SE, MXE, RI, A5SS, and A3SS.

It is noteworthy that users may only investigate one splicing event at a time using VALERIE. A more interactive browser to enable quick visual validation of multiple splicing events will be desirable (Hentges et al., 2022). Nevertheless VALERIE complements existing genome browsers for visualising alternative splicing events with the following features:

a. PSI value display. Current genome browsers only display coverage information. While coverage is informative for visualising gene expression profile (Ozaki et al., 2020), PSI is more informative for visualising alternative splicing events. Afterall, differences in PSI values across cell populations may be invisible at the gene expression level (Ntranos et al., 2019).
b. Per-base PSI value display. User may indirectly infer the PSI values in current genome browsers by using the coverage information at splice junctions (Figure 4.8A). Nevertheless, noise arising from ambiguous sequencing read alignment may make PSI calculation difficult (Figure 4.8B). Moreover, we may check for uniformity of PSI values across the entire exon body from the per-base PSI values display to reveal any cryptic alternative splicing events (Figure 4.5C) or 3'/5'-bias in sequencing coverage (Figures 4.5C and 4.6C).

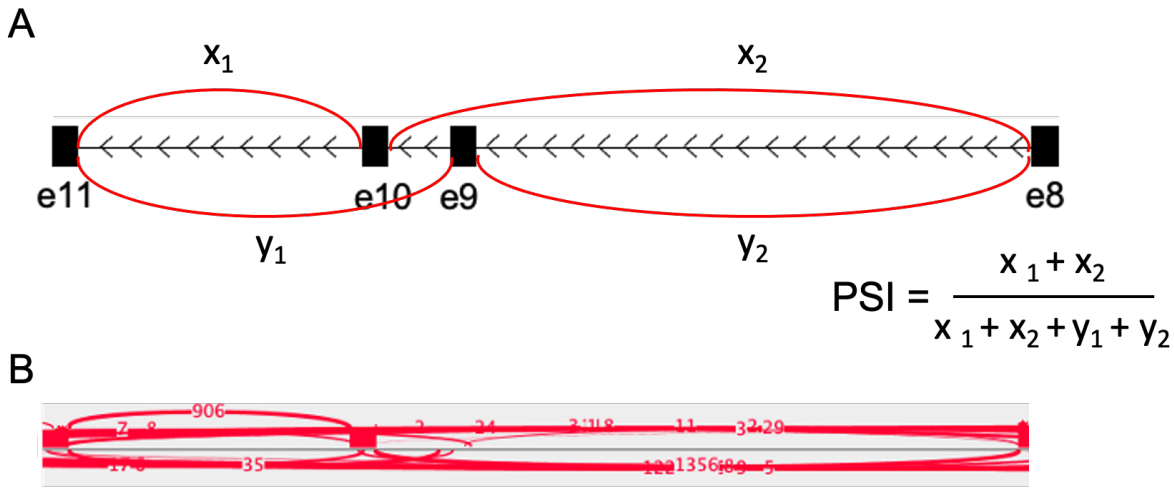


Figure 4.8: Inferring PSI values from splice junction reads from IGV. (A) Simplified diagram for computing PSI values for exon 10. **(B)** Splice junction reads shown for a representative cell in IGV. Noise from ambiguous sequencing read alignment makes PSI inference anything but straightforward.

c. All cells included in display. Current genome browsers only enable a few cells to be displayed for practical reasons, and these few selected cells may not always be representative of their cell population of origin. VALERIE enables all cells included in the study to be displayed. This feature may reveal heterogeneity in alternative splicing profile across a presumed homogeneous cell population and the proportion of cells with dropouts for the genomic locus corresponding to the alternative splicing event.
d. Aggregating PSI values by cell groups. This feature enables users to check for overall differences in alternative splicing profile across the different cell populations in a pseudo-bulk manner.
e. Statistical test for PSI values. This feature enables users to objectively assess the differences in PSI values across the different cell populations for the base positions corresponding to the constitutive and alternative exons.
f. Omitting non-informative intronic regions. This feature enables emphasis of coding exons by censoring long intronic sequences. A notable exception is RI splicing event whereby VALERIE will display the PSI values of the intronic region.
g. Standardising the display of constitutive and alternative exons in the 5' to 3' transcription direction.

5 IMPACT: An integrated myeloid neoplasm platform for alternative splicing candidate prioritisation

5.1 Validation of previously reported aberrant splicing events

To benchmark IMPACT, we first investigated if we were able to recapitulate previously identified aberrant splicing events related to genetic variants in splicing factors *SF3B1*, *SRSF2*, and *U2AF1* (Table 1.1). These splicing events were primarily identified from high-throughput next-generation sequencing and validated using polymerase chain reaction (PCR). The ability to reproduce previously validated splicing events ensures that our data processing including clinical data and genotype tabulation, and splicing quantification, was performed correctly. This is also to ensure that the BeatAML and The Cancer Genome Altas (TCGA) AML cohorts (Tyner et al., 2018) will be relevant for identifying and validating novel splicing events from our own studies.

In total, we tabulated 55 previously reported aberrant splicing events related to *SF3B1*^{K700}, *SRSF2*^{P95}, and *U2AF1*^{S34} (Table 1.1; Figure 5.1). Of which, 28, 9 and 16 splicing events were unique to *SF3B1*^{K700}, *SRSF2*^{P95}, and *U2AF1*^{S34}, respectively. While only two splicing events were aberrantly spliced by both *SRSF2*^{P95} and *U2AF1*^{S34}. This suggests that majority of aberrant splicing events are specific to a given splicing factor.

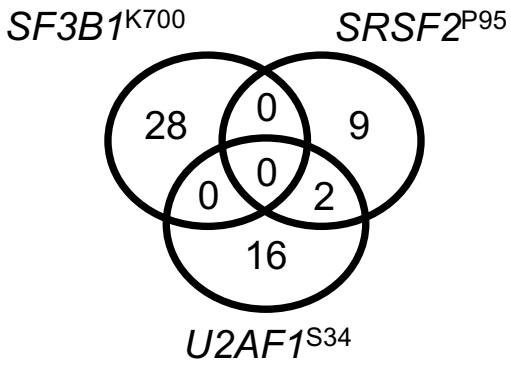


Figure 5.1: Number of previously reported aberrant splicing events related with *SF3B1*^{K700}, *SRSF2*^{P95}, and *U2AF1*^{S34} included in our benchmarking analysis. Note that majority of splicing events are splicing factor-specific.

Of the 28 previously reported *SF3B1*^{K700}-related splicing events, 18 were found to be expressed in the BeatAML (Figure 5.2A). TCGA only consisted of 2 *SF3B1*^{K700} patients and therefore were not included for assessing *SF3B1*^{K700}-related splicing events here.

Of the 18 previously reported *SF3B1*^{K700}- related splicing events, 17 were also differentially spliced in BeatAML *SF3B1*^{K700} patients (Figures 5.2B-R). Specifically, all reported splicing event types related to *SF3B1*^{K700} were that of alternative 3' splice site (A3SS), and these A3SSs were more spliced-in in *SF3B1*^{K700} patients, i.e., these patients had higher percent spliced-in (PSI) values of A3SS. Only one previously reported *SF3B1*^{K700}-related splicing event was not recapitulated in our BeatAML analysis (Figure 5.2S).

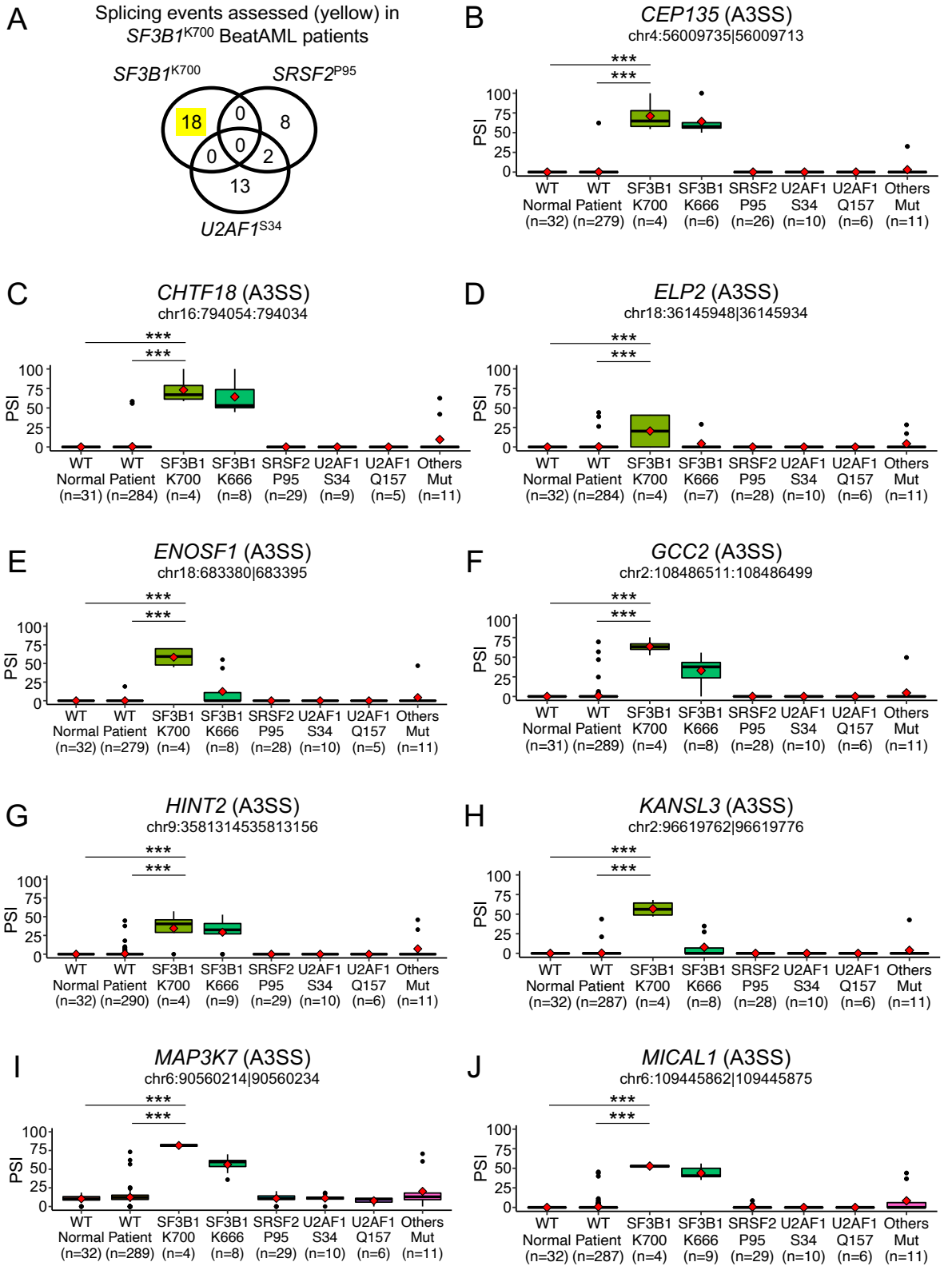
Previous studies have focused on investigating *SF3B1*^{K700}-related splicing events, but not splicing events related to other *SF3B1* hotspot variants. Our BeatAML analysis included *SF3B1*^{K666} hotspot variant and therefore enabled us to compare the splicing profile of *SF3B1*^{K666} to that of *SF3B1*^{K700}. Certain splicing events were differentially spliced to the same degree for both *SF3B1*^{K700} and *SF3B1*^{K666}. For example, A3SS of *CEP135* (Figure 5.2B*), CHTF18* (Figure 5.2C)*, HINT2* (Figure 5.2G)*, MICAL1* (Figure 5.2J)*,* and *SEPTIN6* (Figure 5.2N). This suggests that *SF3B1* hotspot variants may converge on overlapping target genes.

Certain splicing events were differentially spliced for both hotspot variants but to a lesser degree for *SF3B1*^{K666} compared to *SF3B1*^{K700}. For example, A3SS of *ENOSF1* (Figure 5.2E), *GCC2* (Figure 5.2F), *KANSL3* (Figure 5.2H), *MAP3K7* (Figure 5.2I), *PPP2R5A* (Figure 5.2K), *RNF2* (Figure 5.2L), *SEPTIN2* (Figure 5.2M), *SMURF2* (Figure 5.2P), *TMEM14C* (Figure 5.2Q), and *TTI1* (Figure 5.2R). This suggests differences in specificity for overlapping target genes.

On the other hand, certain splicing events were exclusively differentially spliced in *SF3B1*^{K700}, but not *SF3B1*^{K666}. For example, A3SS of *ELP2* (Figure 5.2D) and *SF3B1* (Figure 5.2O). This suggests that different hotspot variants of the same splicing factor may have some mutually exclusive target genes.

Taken together, majority of previously reported splicing events related to *SF3B1*^{K700}, specifically 94% (17/18), were successfully reproduced in our analysis, and therefore we may be confident in using our BeatAML analysis for validation of novel splicing related to *SF3B1* hotspot variants identified from our own studies.

Furthermore, the inclusion of patients with *SF3B1*^{K666} hotspot variant will enable us to investigate splicing events related to this understudied genotype.



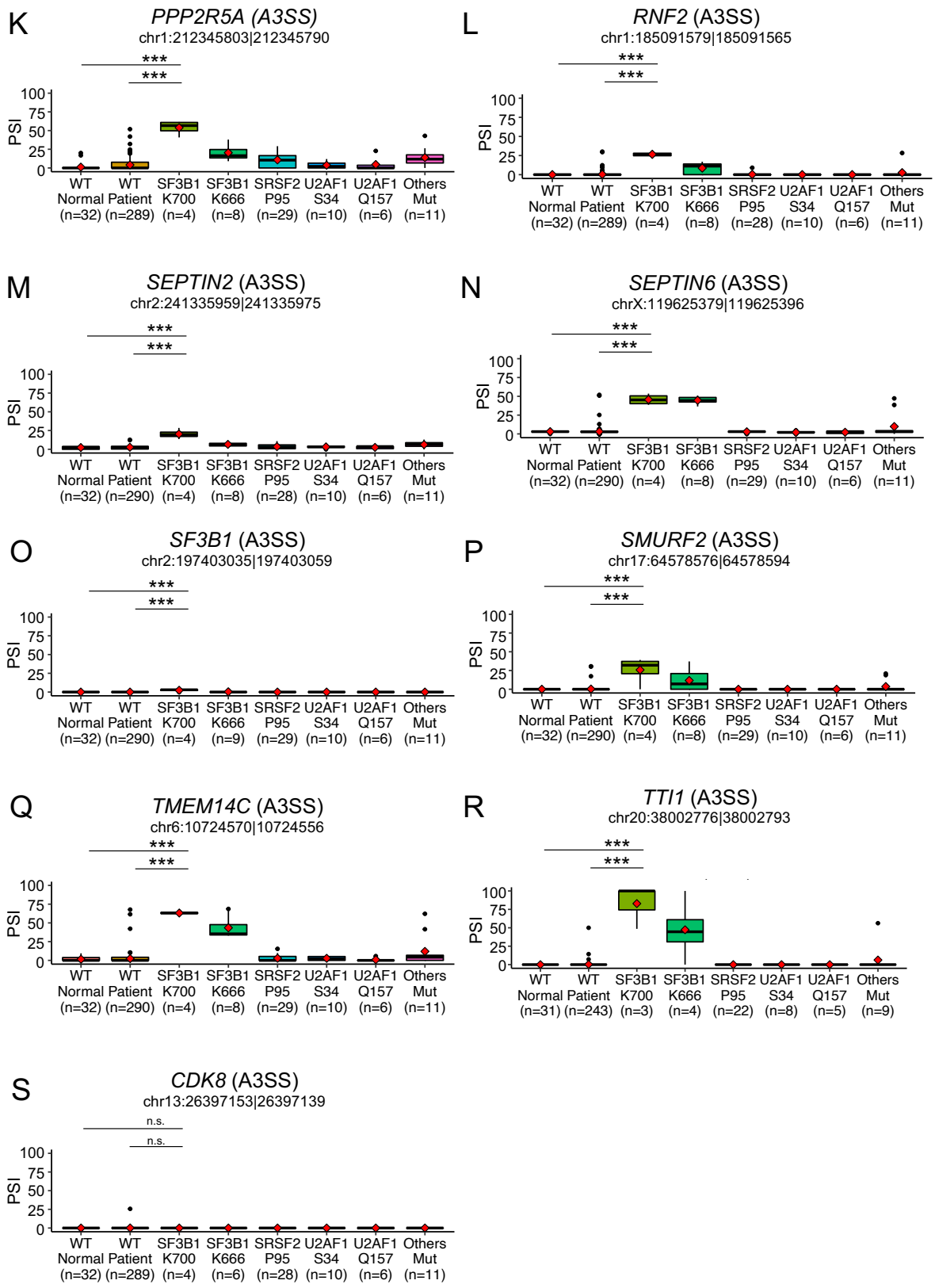


Figure 5.2: Splicing events previously reported to be aberrantly spliced by *SF3B1*^{K700}. (A) Number of previously reported *SF3B1*^{K700}-related splicing events that were expressed and therefore assessed in our BeatAML analysis highlighted in yellow.

(B-R) Previously reported *SF3B1*^{K700}-related splicing events that were confirmed to be differentially spliced in our BeatAML analysis. **(S)** Previously reported *SF3B1*^{K700}-related alternative splicing event that was not confirmed to be differentially spliced in our BeatAML analysis. Wilcoxon rank-sum test used here. WT: Wildtype. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

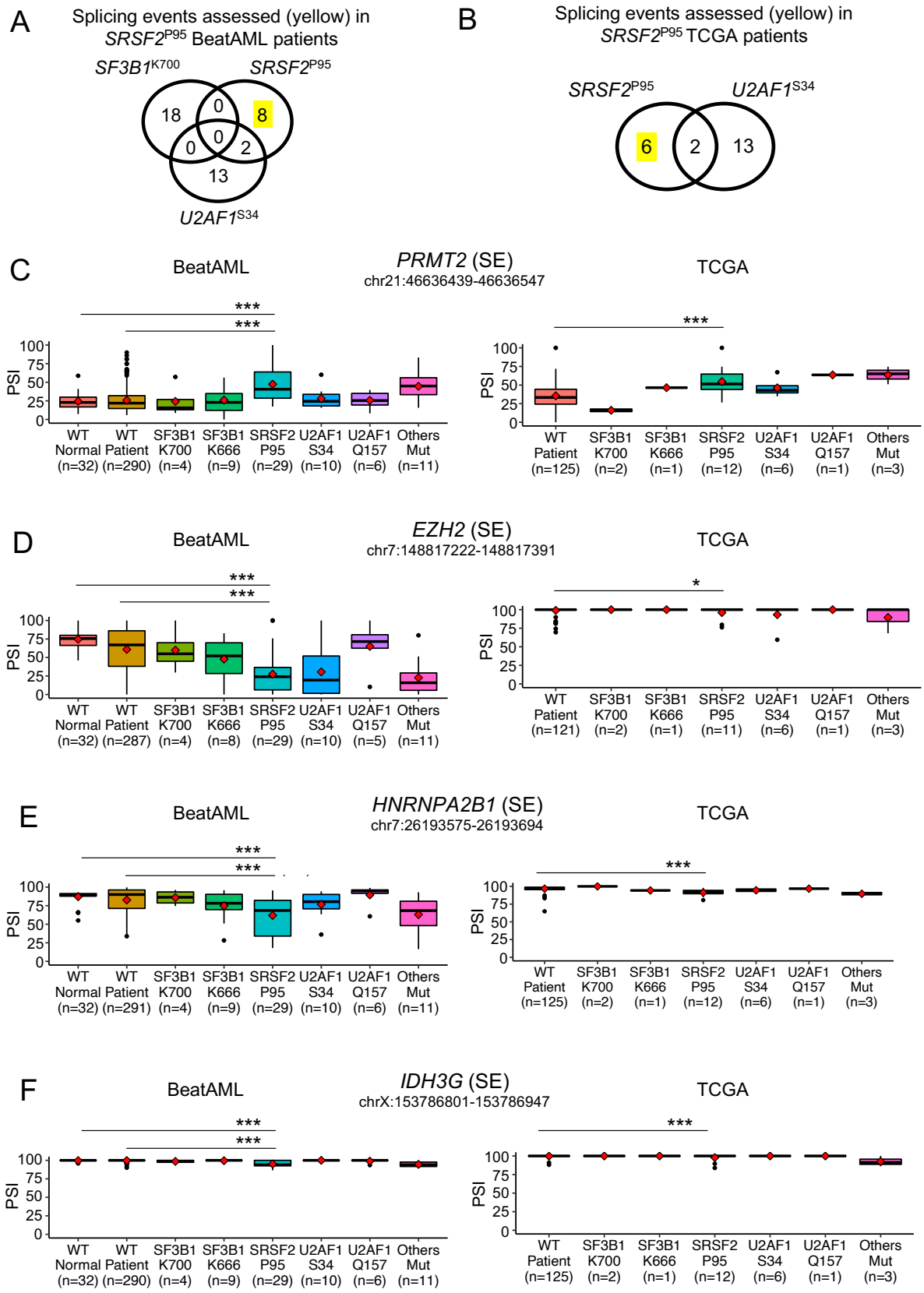
There was sufficient number of *SRSF2*^{P95} samples identified in BeatAML and TCGA, and therefore we were able to assess if previously reported *SRSF2*^{P95}-related splicing events can be recapitulated here in both cohorts. Specifically, 29 and 9 *SRSF2*^{P95} patients were identified in BeatAML and TCGA, respectively.

Of the nine previously reported *SRSF2*^{P95}-specific splicing events, eight were found to be expressed in BeatAML while six were found to be expressed in TCGA AML cohort (Figures 5.3A and B). Half of the previously reported *SRSF2*^{P95}-related splicing events were differentially spliced in *SRSF2*^{P95} patients from both BeatAML and TCGA), namely skipped-exon (SE) of *PRMT2*, *EZH2*, *HNRNPA2B1*, and *IDH3G* (Figures 5.3C-F).

One previously reported *SRSF2*^{P95}-related splicing event was differentially spliced in TCGA, but not BeatAML, namely SE of *AKAP8*. (Figure 5.3G). Furthermore, one previously reported *SRSF2*^{P95}-related splicing event was not differentially spliced in *SRSF2*^{P95} patients from either BeatAML or TCGA, namely SE of *CASP8* (Figure 5.3H). Finally, two previously reported *SRSF2*^{P95}-related splicing events were not differentially spliced in BeatAML cohort and were not expressed in TCGA, namely SE of *FAXDC2* and *IDH3G* (Figures 5.3I-J)

It is noteworthy that *SRSF2*^{P95} preferred splicing event type is SE, and that *SRSF2*^{P95} may either lead to more inclusion (splicing in) or exclusion (splicing out) of the alternative exon. For example, *SRSF2*^{P95} was associated with increased alternative exon inclusion of *PRMT2* (Figure 5.3C) and *AKAP8* (Figure 5.3G) whereas *SRSF2*^{P95} was associated with increased alternative exon exclusion of *EZH2*, *HNRNPA2B1*, and *IDH3G* (Figure 5.3D-F). This is in contrast to *SF3B1* hotspot variants as we have shown earlier that were associated with A3SS splicing event and were associated solely with increased inclusion (splicing in) of this splicing event. No increased exclusion (splicing out) of A3SS by *SF3B1* hotspot variants has been validated to date.

Taken together, majority of previously reported splicing events related to *SRSF2*^{P95} that were expressed in both BeatAML and TCGA here, specifically 83% (5/6), were successfully reproduced in our analysis, and therefore we may be confident in using our BeatAML and TCGA analysis for validation of novel splicing related to *SRSF2* hotspot variants identified from our own studies.


A
Splicing events assessed (yellow) in SRSF2P95 BeatAML patients
SF3B1K700
SRSF2P95
U2AF1S34
18
0
8
0
0
2
13
B
Splicing events assessed (yellow) in SRSF2P95 TCGA patients
SRSF2P95
U2AF1S34
6
2
13
C
BeatAML
PRMT2 (SE)
chr21:46636439-46636547
TCGA
D
BeatAML
EZH2 (SE)
chr7:148817222-148817391
TCGA
E
BeatAML
HNRNPA2B1 (SE)
chr7:26193575-26193694
TCGA
F
BeatAML
IDH3G (SE)
chrX:153786801-153786947
TCGA
PSI
WT Normal (n=32)
WT Patient (n=290)
WT Patient (n=287)
WT Patient (n=291)
WT Patient (n=125)
WT Patient (n=121)
SF3B1 K700 (n=4)
SF3B1 K666 (n=9)
SF3B1 K666 (n=8)
SF3B1 K700 (n=2)
SF3B1 K666 (n=1)
SRSF2 P95 (n=29)
SRSF2 P95 (n=12)
SRSF2 P95 (n=11)
U2AF1 S34 (n=10)
U2AF1 S34 (n=6)
U2AF1 Q157 (n=6)
U2AF1 Q157 (n=5)
U2AF1 Q157 (n=1)
Others Mut (n=11)
Others Mut (n=3)

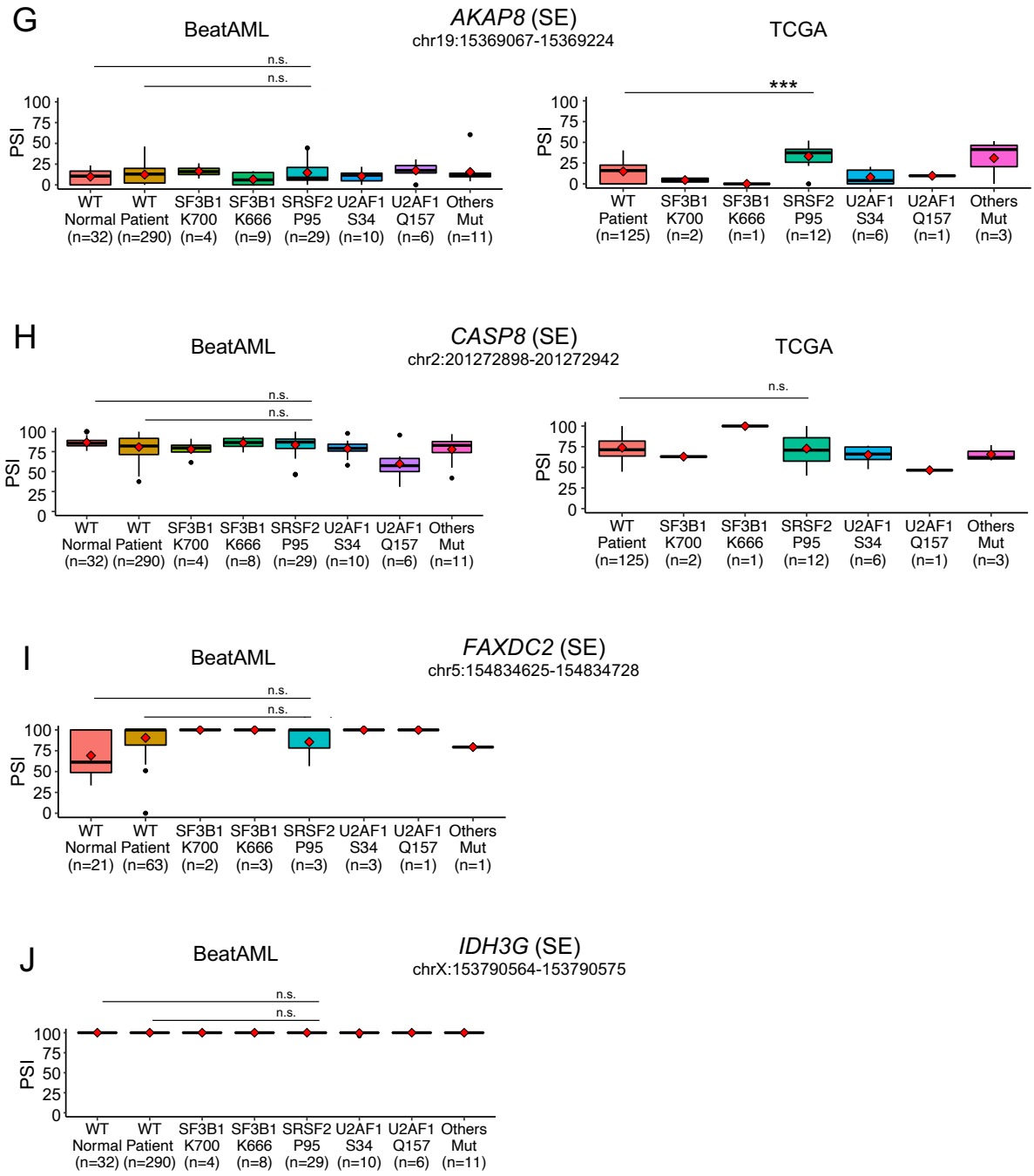


Figure 5.3: Splicing events previously reported to be aberrantly spliced by *SRSF2*^{P95}. (A-B) Number of previously reported *SRSF2*^{P95}-related splicing events that were expressed and therefore assessed in our BeatAML and TCGA analysis highlighted in yellow. **(C-F)** Previously reported *SRSF2*^{P95}-related splicing events that were confirmed to be differentially spliced in our BeatAML and TCGA analysis. **(G-J)** Previously reported *SRSF2*^{P95}-related splicing events that were confirmed to be differentially spliced **(G)** only in TCGA AML analysis or **(H-J)** in neither cohort.

Wilcoxon rank-sum test used here. WT: Wildtype. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

There was sufficient number of *U2AF1*^{S34} patients identified in BeatAML and TCGA, and therefore we were able to assess if previously reported *U2AF1*^{S34}-related splicing events can be recapitulated here in both cohorts. Specifically, 10 and 6 *U2AF1*^{S34} patients were identified in BeatAML and TCGA, respectively.

Of the 18 previously reported *U2AF1*^{S34}-specific splicing events, 13 were found to be expressed in both BeatAML and TCGA (Figures 5.4A and B). Nine of the previously reported *U2AF1*^{S34}-related splicing events were differentially spliced in *U2AF1*^{S34} patients from both BeatAML and TCGA, namely *CTNNB1* (A3SS), *H2AFY* (MXE), *IRAK4* (SE), *MED24* (SE), *PICALM* (A3SS), *DEK* (SE), *KDM6A* (SE), *SMARCA5* (SE), and *STRAP* (SE) (Figures 5.4C-K).

Two previously reported *U2AF1*^{S34}-related splicing events was differentially spliced in in BeatAML, but not TCGA, namely *RHBDD2* (SE) and *SERPINB8* (A3SS) (Figure 5.4L and M). On the other hand, one previously reported *U2AF1*^{S34}-related splicing events was differentially spliced in *U2AF1*^{S34} patients from TCGA, but not BeatAML, namely *BCOR* (A3SS) (Figure 5.4N). Finally, only one previously reported *U2AF1*^{S34}-related splicing events was not differentially spliced in *U2AF1*^{S34} patients from either BeatAML or TCGA, namely *ITGB3BP* (SE) (Figure 5.4O).

It is noteworthy that there were two types of splicing events that were preferentially targeted by *U2AF1*^{S34}, namely A3SS and SE. This is in contrast to *SF3B1*^{K700} and *SRSF2*^{P95} that have preference for only one splicing event type, namely A3SS or SE, respectively. Not surprisingly, there may be convergence of target genes by *U2AF1*^{S34} and *SF3B1*^{K700} or *U2AF1*^{S34} and *SRSF2*^{P95}. Indeed, SE of *EZH2* and *GNAS* were reported to be differentially spliced by both *U2AF1*^{S34} and *SRSF2*^{P95} (Shiozawa et al., 2018; Wheeler et al., 2022).

Previous studies have focused on investigating *U2AF1*^{S34}-related splicing events, but not splicing events related to other *U2AF1* hotspot variants. Our BeatAML cohort analysis included sufficient samples with *U2AF1*^{Q157} hotspot variant (n=6) and therefore enabled us to compare the splicing profile of *U2AF1*^{Q157} to that of *U2AF1*^{S34}.

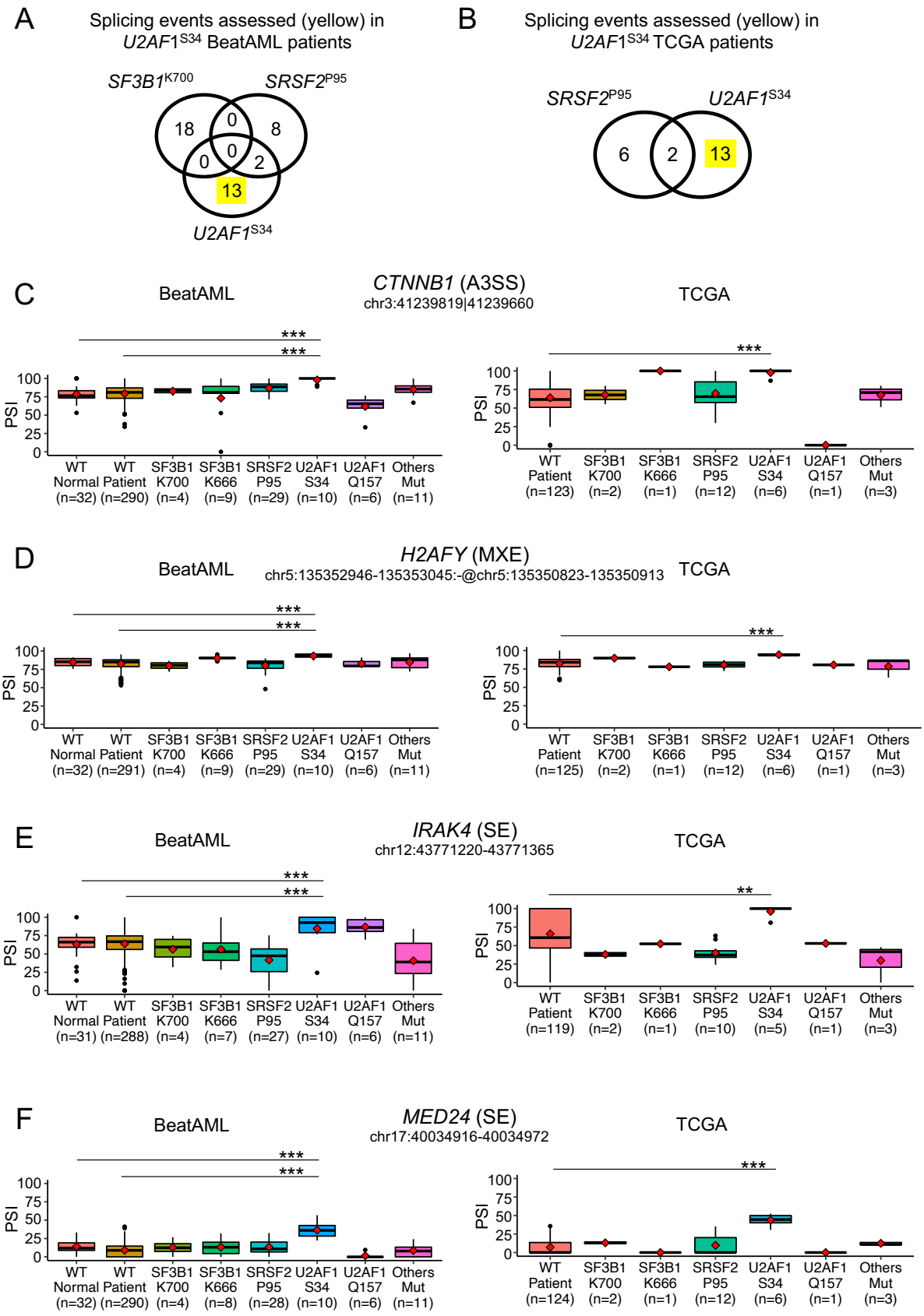
Certain alternative splicing events were differentially spliced to the similar degree for both *U2AF1*^{S34} and *U2AF1*^{Q157}. For example, *IRAK4* (SE) (Figure 5.4E) and

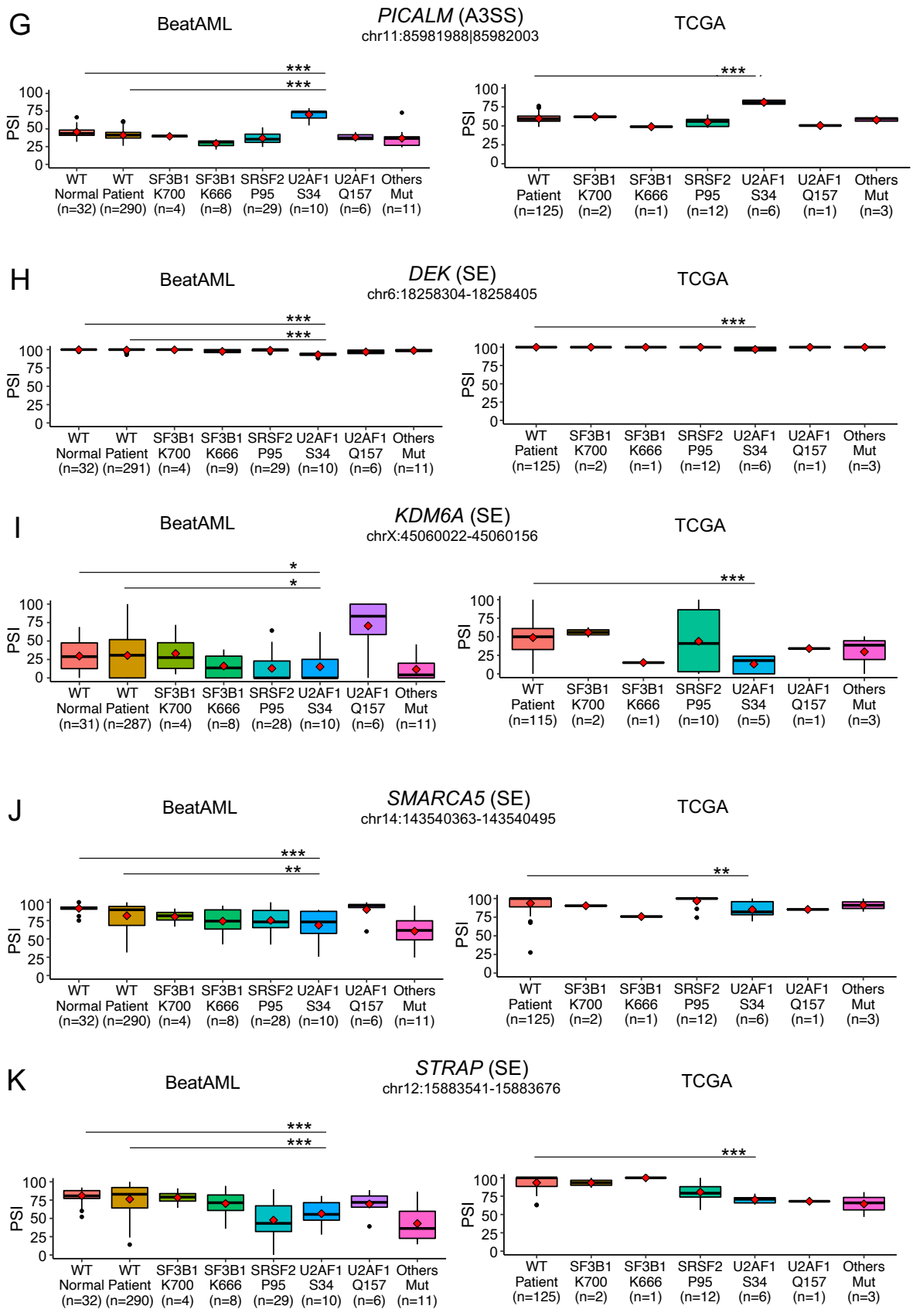
BCOR (A3SS) (Figure 5.4N). This suggests that *U2AF1* hotspot variants may converge on overlapping target genes.

However, certain splicing events were differentially spliced in the opposite direction for *U2AF1*^{S34} compared to *U2AF1*^{Q157}. For example, *CTNNB1* (SE) (Figure 5.4C), *H2AFY* (MXE) (Figure 5.4D), and *MED24* (SE) (Figure 5.4F) had increased PSI values in *U2AF1*^{S34} patients relative to wildtype patients and healthy donors whereas the same splicing events had decreased PSI values in *U2AF1*^{Q157} patients relative to wildtype patients and healthy donors.

On the other hand, *KDM6A* (SE) (Figure 5.4I) and *SMARCA5* (SE) (Figure 5.4J) had decreased PSI values in *U2AF1*^{S34} patients relative to wildtype patients and healthy donors whereas the same splicing events had increased PSI values in *U2AF1*^{Q157} patients relative to wildtype patients and healthy donors. This suggests opposite effects of *U2AF1*^{S34} and *U2AF1*^{Q157} on overlapping target genes. This is in contrast to *SF3B1*^{K700} and *SF3B1*^{K666} whereby both hotspot variants were consistently associated with increased PSI values of A3SS of their overlapping target genes.

Taken together, majority of previously reported splicing events related to *U2AF1*^{S34}, specifically 93% (12/13), were successfully reproduced in our analysis, and therefore we may be confident in using our BeatAML and TCGA analysis for validation of novel splicing events related to *U2AF1* hotspot variants identified from our own studies. Furthermore, the inclusion of patients with *U2AF1*^{Q157} hotspot variant will enable us to investigate aberrant splicing events related to this understudied genotype.


A
Splicing events assessed (yellow) in U2AF1S34 BeatAML patients
SF3B1K700
SRSF2P95
U2AF1S34
18
0
8
0
0
2
13
B
Splicing events assessed (yellow) in U2AF1S34 TCGA patients
SRSF2P95
U2AF1S34
6
2
13
C
BeatAML
CTNNB1 (A3SS)
chr3:41239819|41239660
TCGA
***
***
***
PSI
100
75
50
25
0
WT Normal (n=32)
WT Patient (n=290)
SF3B1 K700 (n=4)
SF3B1 K666 (n=9)
SRSF2 P95 (n=29)
U2AF1 S34 (n=10)
U2AF1 Q157 (n=6)
Others Mut (n=11)
WT Patient (n=123)
SF3B1 K700 (n=2)
SF3B1 K666 (n=1)
SRSF2 P95 (n=12)
U2AF1 S34 (n=6)
U2AF1 Q157 (n=1)
Others Mut (n=3)
D
BeatAML
H2AFY (MXE)
chr5:135352946-135353045:-@chr5:135350823-135350913
TCGA
***
***
***
WT Normal (n=32)
WT Patient (n=291)
SF3B1 K700 (n=4)
SF3B1 K666 (n=9)
SRSF2 P95 (n=29)
U2AF1 S34 (n=10)
U2AF1 Q157 (n=6)
Others Mut (n=11)
WT Patient (n=125)
SF3B1 K700 (n=2)
SF3B1 K666 (n=1)
SRSF2 P95 (n=12)
U2AF1 S34 (n=6)
U2AF1 Q157 (n=1)
Others Mut (n=3)
E
BeatAML
IRAK4 (SE)
chr12:43771220-43771365
TCGA
***
***
**
WT Normal (n=31)
WT Patient (n=288)
SF3B1 K700 (n=4)
SF3B1 K666 (n=7)
SRSF2 P95 (n=27)
U2AF1 S34 (n=10)
U2AF1 Q157 (n=6)
Others Mut (n=11)
WT Patient (n=119)
SF3B1 K700 (n=2)
SF3B1 K666 (n=1)
SRSF2 P95 (n=10)
U2AF1 S34 (n=5)
U2AF1 Q157 (n=1)
Others Mut (n=3)
F
BeatAML
MED24 (SE)
chr17:40034916-40034972
TCGA
***
***
***
WT Normal (n=32)
WT Patient (n=290)
SF3B1 K700 (n=4)
SF3B1 K666 (n=8)
SRSF2 P95 (n=28)
U2AF1 S34 (n=10)
U2AF1 Q157 (n=6)
Others Mut (n=11)
WT Patient (n=124)
SF3B1 K700 (n=2)
SF3B1 K666 (n=1)
SRSF2 P95 (n=12)
U2AF1 S34 (n=6)
U2AF1 Q157 (n=1)
Others Mut (n=3)


G
BeatAML
PICALM (A3SS)
chr11:85981988|85982003
TCGA
H
DEK (SE)
chr6:18258304-18258405
I
KDM6A (SE)
chrX:45060022-45060156
J
SMARCA5 (SE)
chr14:143540363-143540495
K
STRAP (SE)
chr12:15883541-15883676
PSI
WT Normal (n=32)
WT Patient (n=290)
WT Patient (n=125)
SF3B1 K700 (n=4)
SF3B1 K666 (n=8)
SF3B1 K666 (n=1)
SRSF2 P95 (n=29)
SRSF2 P95 (n=12)
U2AF1 S34 (n=10)
U2AF1 S34 (n=6)
U2AF1 Q157 (n=6)
U2AF1 Q157 (n=1)
Others Mut (n=11)
Others Mut (n=3)

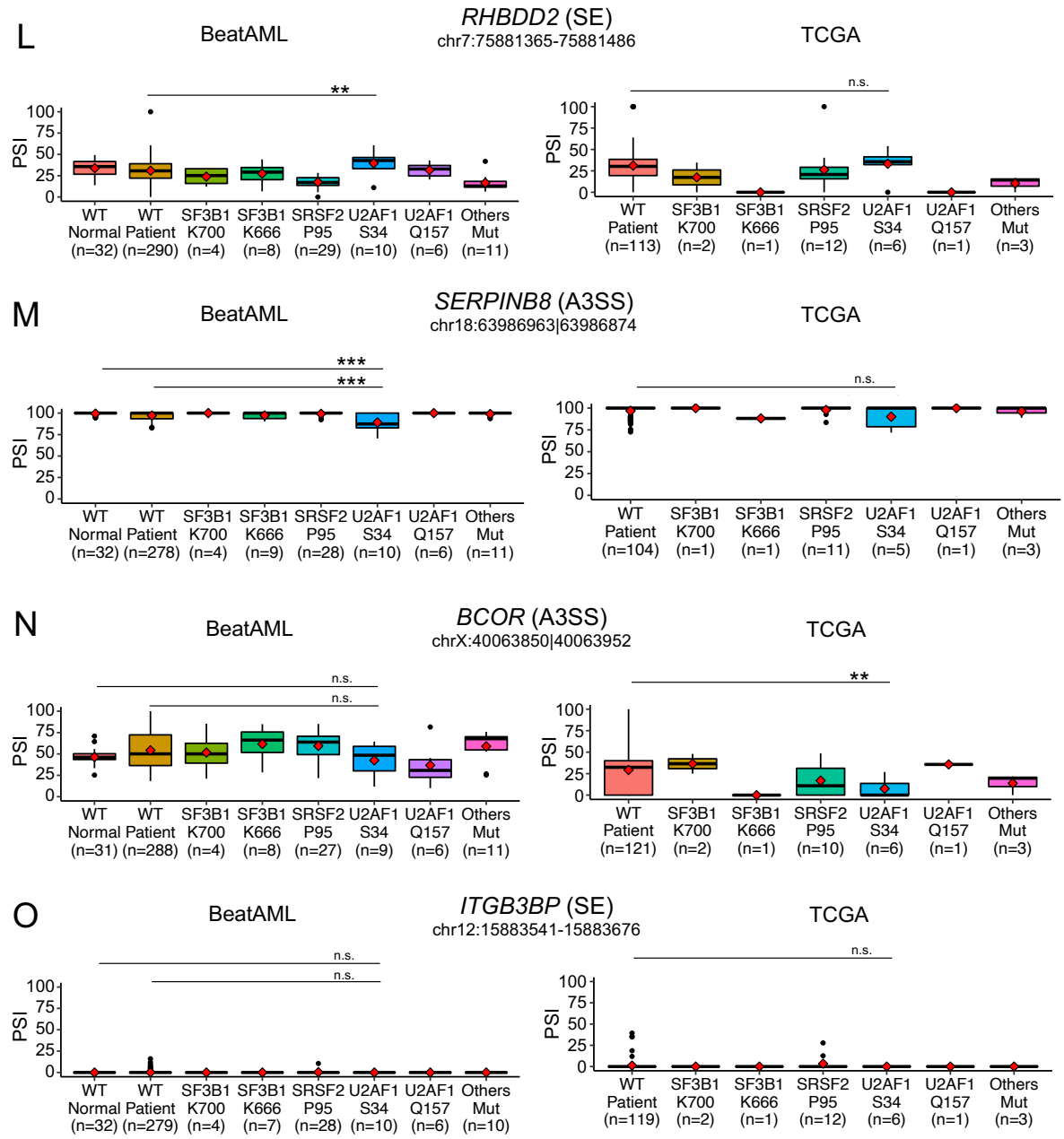


Figure 5.4: Splicing events previously reported to be aberrantly spliced by *U2AF1*^{S34}. (A-B) Number of previously reported *U2AF1*^{S34}-related splicing events that were expressed and assessed in our BeatAML and TCGA analysis highlighted in yellow. **(C-K)** Previously reported *U2AF1*^{S34}-related splicing events that were confirmed to be differentially spliced in our BeatAML and TCGA analysis. **(L-N)** Previously reported *U2AF1*^{S34}-related splicing events that were confirmed to be differentially spliced in either BeatAML or TCGA. **(O)** Previously reported *U2AF1*^{S34}-related splicing event that was not confirmed in either BeatAML or TCGA. Wilcoxon

rank sum-test used here. WT: Wildtype. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

There were only two splicing events that were reported to be aberrantly spliced by more than one splicing factor, and both events were expressed in both *SRSF2*^{P95} and *U2AF1*^{S34} genotypes in our BeatAML and TCGA analysis (Figures 5.5A and B). Both alternative exons were more spliced-in in both *SRSF2*^{P95} and *U2AF1*^{S34} patients in BeatAML and TCGA (Figures 5.5C and D) as previously reported (Shiozawa et al., 2018; Wheeler et al., 2022). The inclusion (splicing in) of an alternative exon in *EZH2* that interrupts the open reading frame and ultimately leads to the nonsense-mediated decay (NMD) of the transcript (Shiozawa et al., 2018). On the other hand, the inclusion (splicing in) of an alternative exon in *GNAS* does not disrupt the open reading frame, rather this alternative exon encodes for a disordered region within the final protein structure, which in turn increases *GNAS* activation and adenylyl cyclase activity, and downstream ERK/MAPK pathway (Wheeler et al., 2022).

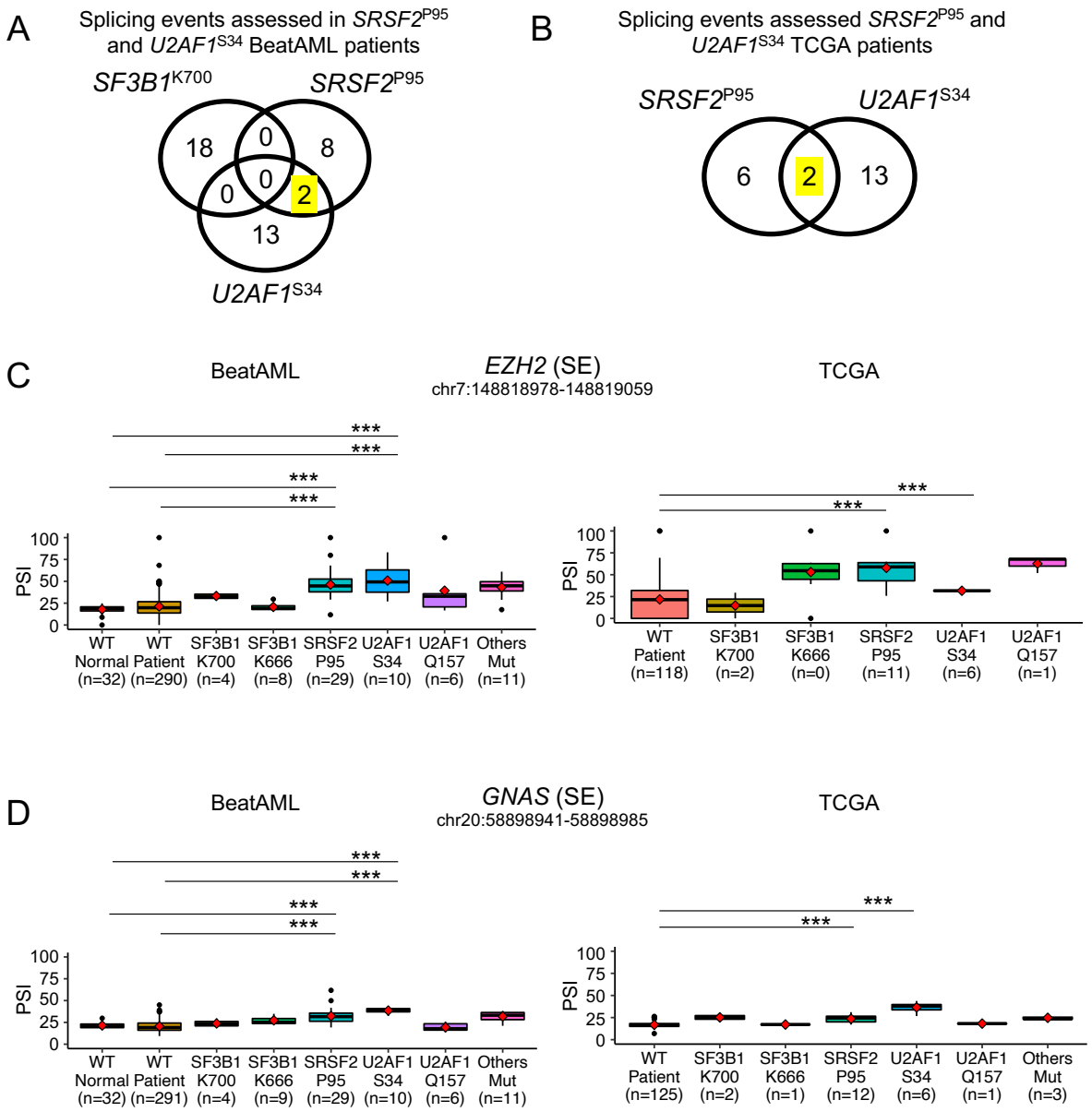


Figure 5.5: Splicing events previously reported to be aberrantly spliced by both *SRSF2*^{P95} and *U2AF1*^{S34}. (A-B) These were the only overlapping alternative splicing events associated with more than one splicing factor, namely *SRSF2*^{P95} and *U2AF1*^{S34} and were assessed in both BeatAML and TCGA highlighted in yellow. **(C-D)** Increased inclusion (splicing in) of alternative exon in both *SRSF2*^{P95} and *U2AF1*^{S34} patients in both BeatAML and TCGA. Wilcoxon rank-sum test used here. WT: Wildtype. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

Taken together, majority of previously reported splicing events related to *SF3B1*^{K700}, *SRSF2*^{P95}, and *U2AF1*^{S34} were successfully reproduced in our analysis.

Moreover, there was general agreement between BeatAML and TCGA cohorts. Therefore, we may be confident in using our BeatAML and TCGA analysis for validation of novel splicing related these hotspot variants identified from our own studies, in particular myeloid neoplasm. Furthermore, the inclusion of hitherto understudied genotypes, namely *SF3B1*^{K666} and *U2AF1*^{Q157}, will enable us to validate novel splicing events related these genotypes from our own studies.

It is noteworthy that previously reported splicing events used in our benchmarking analysis here were identified in a variety of leukaemia aside from AML, including myelodysplastic syndrome (MDS), myeloproliferative neoplasms (MPN), and chronic lymphocytic leukaemia (CLL). Therefore, our BeatAML and TCGA analysis may also be relevant for validating novel splicing events identified not only in AML patients but also in other types of leukaemia.

5.2 Validation of previously reported clinically relevant splicing events

One approach for prioritising candidate splicing events identified from high-throughput RNA-seq in myeloid neoplasm is to select for events associated with clinical features, including prognosis (survival), and neutrophil and platelet counts (Pellagatti et al., 2018; Smith et al., 2019). One such splicing event that has been reported to be aberrantly spliced in *U2AF1*^{S34} patients and was associated with poor prognosis in overall acute myeloid leukaemia (AML) patients was *IRAK4* exon 4 (Smith et al., 2019). Increased inclusion of this alternative exon in *U2AF1*^{S34} patients was validated using polymerase chain reaction (PCR) in both myelodysplastic syndrome (MDS) and AML patients. Indeed, our earlier BeatAML and The Cancer Genome Atlas (TCGA) analysis recapitulated this increased in *IRAK4* exon 4 inclusion in *U2AF1*^{S34} patients (Figure 5.4E). Here, we further investigated if our BeatAML and TCGA analysis could recapitulate the previously reported association between *IRAK4* exon 4 inclusion and worse prognosis.

We observed no association between increased *IRAK4* exon 4 inclusion with overall survival in our BeatAML analysis (Figure 5.6A). There was also no association between increased *IRAK4* exon 4 inclusion with overall survival when we only included *de novo* AML patients from BeatAML (Figures 5.6B). There was trend towards worse survival for patients with increased *IRAK4* exon 4 inclusion among transformed AML patients from the BeatAML, but the association was not statistically significant,

possibly due to small sample size. Lastly, we similarly did not observe any association between increased *IRAK4* exon 4 inclusion with overall survival in our TCGA analysis (Figure 5.6D).

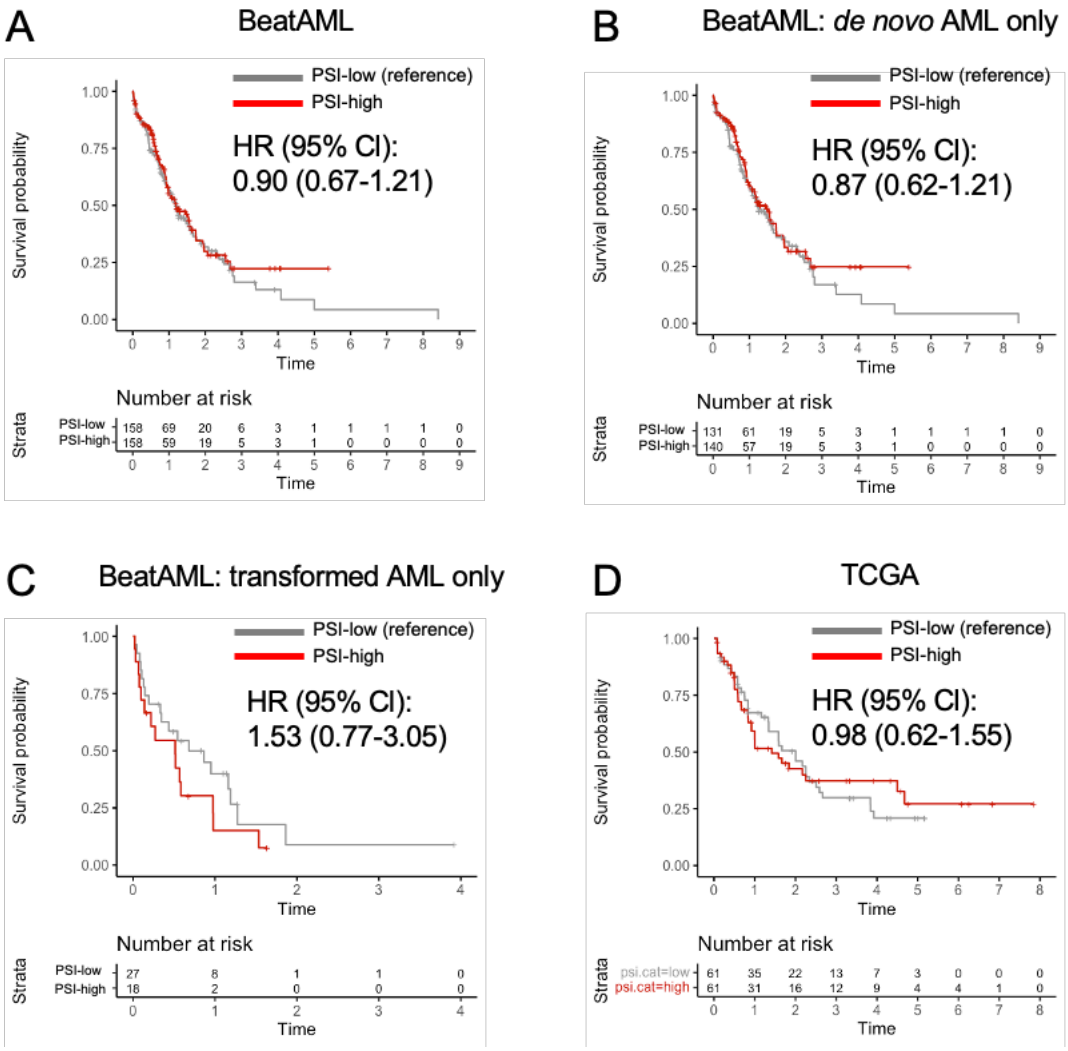


Figure 5.6: Association between *IRAK4* exon 4 inclusion (PSI high) with overall survival. (A-C) Survival analysis in BeatAML cohort for **(A)** all AML patients, **(B)** *de novo* AML patients only, and **(C)** transformed AML patients only. **(D)** Survival analysis in TCGA cohort. 95% CI: 95% confidence interval; HR: Hazard ratio.

One possible reason for the discrepancy between our survival analysis and that previously reported by the original study (Smith et al., 2019) may be the differences in quantifying the exon 4 inclusion rate of *IRAK4*. Specifically, we used a splice junction-based approach to compute the percent spliced-in (PSI) rate of *IRAK4* exon 4. This

splice junction-based approach for PSI quantification is a widely adopted approach for quantifying alternative exon inclusion rate (Kahles et al., 2018; Schischlik et al., 2019; Song et al., 2017). On the other hand, the original study used an unconventional quantification method for quantifying alternative exon inclusion rate whereby the *IRAK4* exon 4 expression is first quantified in reads per kilobase million (RPKM) followed by dividing this value with the total RPKM across all exons of this gene. Taken together, differences in quantifying alternative exon inclusion rate may lead to differences in correlation results.

5.3 Validation of previously reported drug-sensitive splicing events

Aside from validating novel splicing events and prioritising clinically relevant splicing events, IMPACT also aims to prioritise splicing events amenable to targeted therapy. To this end, we included *in vivo* and *ex vivo* drug sensitivity data from Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012; Ghandi et al., 2019) and BeatAML (Tyner et al., 2018), respectively.

To benchmark IMPACT, we first investigated if we were able to recapitulate previously reported correlations between splicing events and drug sensitivity. This is to ensure that our data processing including tabulation of drug sensitivity measurements and sample metadata, and splicing quantification, was performed correctly. One such reported correlation was between *MDM4* exon 6 and Nutlin-3 sensitivity (Ghandi et al., 2019).

MDM4 is a negative regulator of p53. The exclusion (splicing out) of *MDM4* exon 6 disrupts the open reading frame by introducing a premature stop codon and consequently leads to nonsense-mediate decay (NMD) of its mRNAs (Rallapalli, Strachan, Cho, Mercer, & Hall, 1999). Nutlin-3 has been shown to be confer sensitivity to cancer cell lines with *MDM4* exon 6 inclusion (splicing in) (Ghandi et al., 2019).

We observed *MDM4* exon 6 inclusion rate to be negatively correlated with Nutlin-3 AUC values (Pearson correlation = -0.65; *P* value < 0.01) in CCLE whereby *MDM4* exon 6 inclusion rate was associated with higher sensitivity to Nultin-3 (Figure 5.7A). Note that lower the area under curve (AUC) indicates higher sensitivity. The association between *MDM4* exon 6 inclusion rate and Nutlin-3 sensitivity may be partly explained by somatic variants in *TP53* whereby *TP53*^{WT} cell lines were more sensitivity to Nultin-3. *TP53*^{WT} as a prerequisite for Nutlin-3 sensitivity is well established in

multiple cancer types including ovarian cancer (Crane et al., 2015) and brain cancer (Kunkele et al., 2012). These suggest that a functional *MDM4* protein, i.e., the inclusion of *MDM4* exon 6, may also be a prerequisite of Nutlin-3 sensitivity.

However, we did not observe any correlation between *MDM4* exon 6 inclusion rate and Nutlin-3 sensitivity in BeatAML (Pearson correlation = -0.05; *P* value = 0.60) (Figure 5.7B). Next, we investigated if we could recapitulate the almost-mutually exclusive relationship between *MDM4* exon 6 inclusion and *TP53*^{MUT} previously observed in our CCLE analysis (Figure 5.7A). Similar to our observation in CCLE, we observed *TP53*^{WT} patients to have increased inclusion of *MDM6* exon 6 in BeatAML (Figure 5.7C). This observation was also successfully reproduced in TCGA (Figure 5.7D). *TP53*^{WT} patients in BeatAML demonstrated higher sensitivity to Nutlin-3 compared to *TP53*^{MUT} patients, as expected (Figure 5.7E).

Taken together, these suggest that factors in addition to somatic variants in *TP53* may affect the relationship between *MDM4* exon 6 inclusion and Nutlin-3 sensitivity in patient samples.

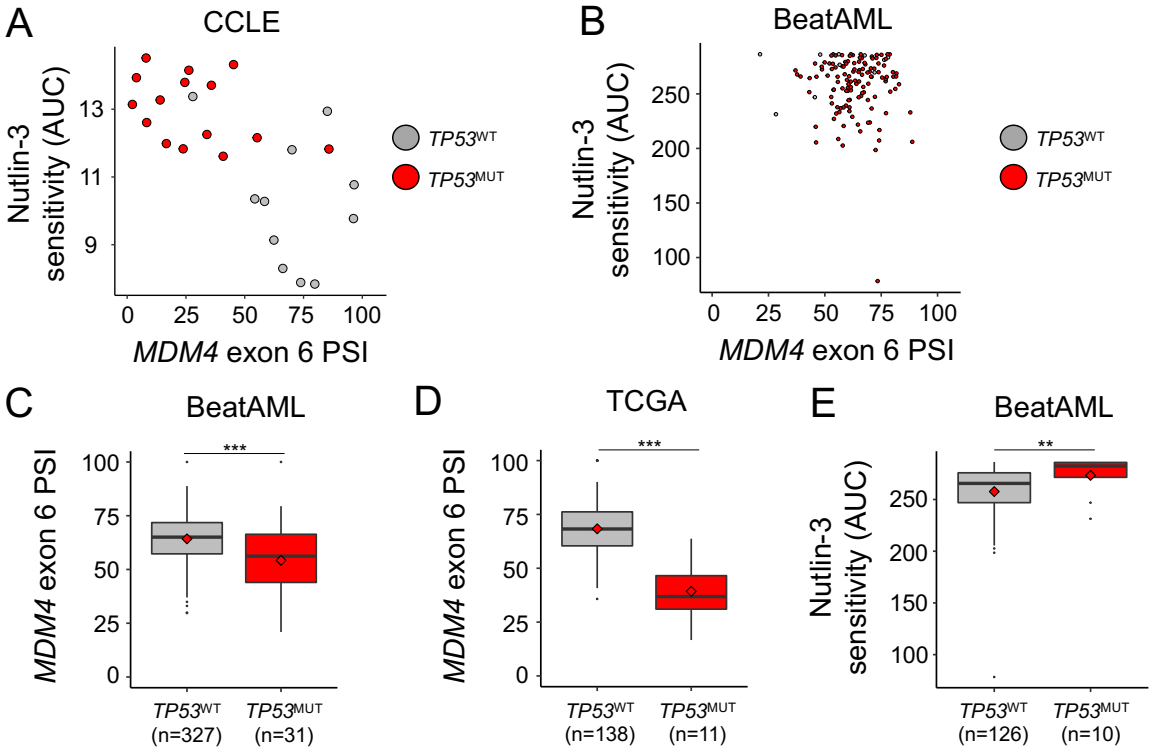


Figure 5.7: Correlation between *MDM4* exon 6 inclusion rate and sensitivity to Nutlin-3 treatment measured in AUC. (A-B) Correlation between *MDM4* exon 6

inclusion rate and Nultin-3 AUC for **(A)** haematopoietic cancer cell lines from CCLE and **(B)** isolated mononuclear cells from BeatAML patients. **(C-D)** *MDM4* exon 6 inclusion stratified by somatic variants in *TP53* among **(C)** BeatAML and **(D)** TCGA patients. **(E)** Nultin-3 AUC stratified by somatic variants in *TP53* among BeatAML patients. The discrepancy in the sample size between (C) and (E) was due to not every patient having drug treatment record also had sufficient coverage (>10x) at *MDM4* exon 6 for PSI quantification. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

Our previous analysis was performed using the AUC as the measurement of drug sensitivity. An alternative measurement of drug sensitivity is IC50 (half-maximal inhibitory concentration). AUC has been shown to be a more reproducible measurement of drug sensitivity compared to IC50 when assessing drug sensitivity across multiple datasets (Haibe-Kains et al., 2013). Moreover, AUC has been shown to be more predictive of drug sensitivity compared to IC50 (Kurilov, Haibe-Kains, & Brors, 2020). Indeed, re-analysis of *MDM4* exon 6 inclusion rate and Nutlin-3 sensitivity revealed no association between the two variables in CCLE or BeatAML (Figures 5.8A and B). Moreover, there was no association between somatic variants in *TP53* and Nutlin-3 sensitivity in our BeatAML analysis (Figure 5.8C).

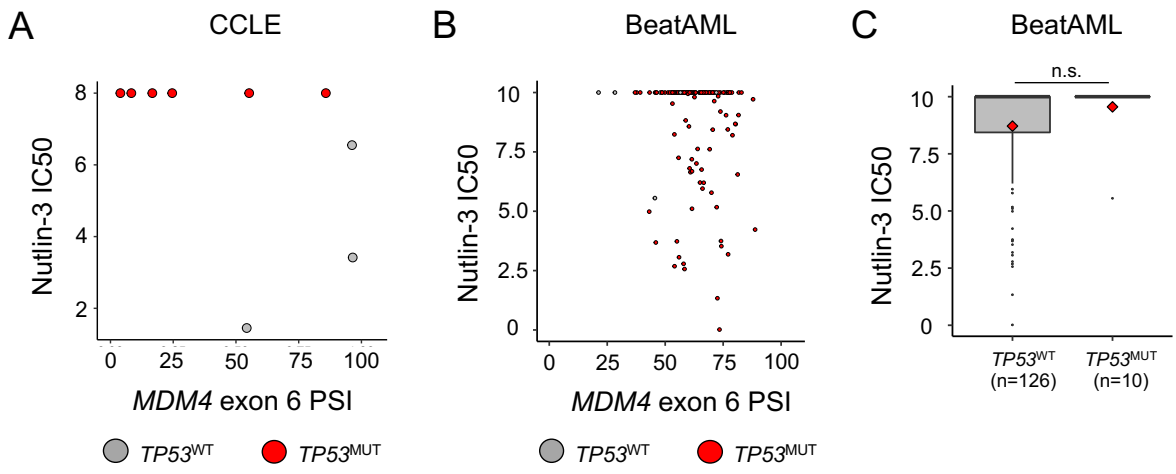


Figure 5.8: Correlation between *MDM4* exon 6 inclusion rate and sensitivity to Nutlin-3 treatment measured in IC50. (A-B) Correlation between *MDM4* exon 6 inclusion rate and Nultin-3 IC50 for **(A)** haematopoietic cancer cell lines from CCLE

and **(B)** isolated mononuclear cells from BeatAML patients. **(C)** Nultin-3 IC50 stratified by somatic variants in *TP53* among BeatAML patients. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

Taken together, we successfully recapitulated the previously reported association between *MDM4* exon 6 inclusion rate and Nutlin-3 sensitivity in CCLE, but not in BeatAML. This suggests homogeneous cell lines may be a better model system compared to heterogeneous patient samples for identifying correlations between splicing events and drug sensitivity. We also demonstrated that AUC may be more an accurate measurement of drug sensitivity compared to IC50 for identifying correlations between splicing events and drug sensitivity. Lastly, other factor aside of exon inclusion rate, for example somatic variants in cancer-related genes, may need to be taken into account to explain any correlations observed between splicing events and drug sensitivity.

6 Application of developed computational pipelines on myeloid neoplasm patients

6.1 Single-cell analysis of a *SF3B1*-mutant MDS patient

SF3B1 is involved in 3' splice site recognition during RNA splicing. Specifically, the interaction between SF3B1, p14, and intron sequence enables the binding of U2 snRNP complex to the intron through base-pairing interaction between the branchpoint sequence and U2 snRNA. The proper recognition of branchpoint sequence by U2 snRNP complex is essential for the subsequent recognition of canonical 3' splice site and ultimately intron removal. Genetic variants in *SF3B1* gene have been shown to lead to recognition of alternative branchpoint sequences and consequently recognition of alternative 3' splice sites (Alsafadi et al., 2016).

Genetic variants in *SF3B1* have been found in ~20-30% of myelodysplastic neoplasm (MDS) patients (Papaemmanuil et al., 2011; Pellagatti et al., 2018; Shiozawa et al., 2018). The variants typically occur on the HEAT (Huntingtin, Elongation factor 3, protein phosphatase 2A, Targets of rapamycin 1) domain. *SF3B1*^{K700} variants are the most common variants identified in MDS, and as a consequence, the impact of these variants on the transcriptome is the most well characterised to date. Nevertheless, there exists other less common *SF3B1* variants, such as R625, N626, and K666, that have been reported in MDS patients. Nevertheless, the impact of these variants on the transcriptome has not been well characterised. It is conceivable that different hotspot variants of *SF3B1* may give rise to distinct pattern of mis-splicing, and by extension, clinical phenotype. Indeed, it has been recently reported that *SF3B1*^{K666N} was associated with increased progression of MDS (Dalton et al., 2020). This is in contrast with the low-risk MDS with ring sideroblasts (MDS-RS) associated with the more common *SF3B1*^{K700E} variant (Kanagal-Shamanna et al., 2021).

We have earlier demonstrated the application and proof-of-principle of our single-cell splicing pipeline on plate-based RNA-seq data generated from homogeneous cell lines derived from induced pluripotent stem cells (iPSCs) and endoderm cells (see section 3.4: "Demonstration on plate-based RNA-seq dataset" section). Here, we will demonstrate the application and proof-of-principle of our single-cell splicing pipeline on plate-based RNA-seq data (Rodriguez-Meira et al., 2019) generated from heterogeneous haematopoietic stem and progenitor cell populations.

This dataset consists of a single MDS patient with both *SF3B1*^{K666N} and *SF3B1*^{N626D} variants detected. Specifically, the *SF3B1*^{N626D} clone was the major clone at diagnosis, but the *SF3B1*^{K666N} clone became the major clone as the disease progressed (Figures 6.1A and B). Additional three healthy donors aged 39, 49, and 75 were also included in this analysis to serve as a reference group relative to *SFB31*^{MUT} cells.

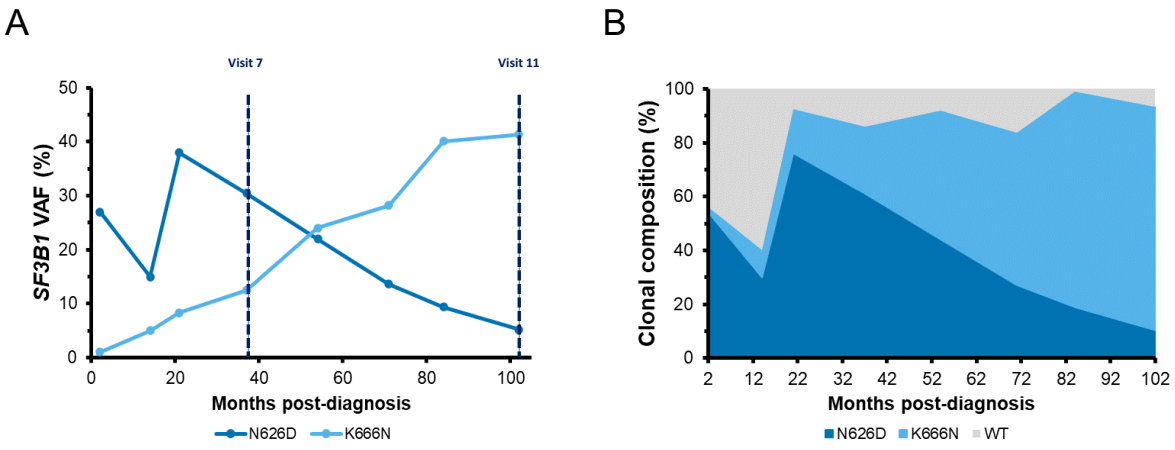


Figure 6.1: *SF3B1* variant analysis for a MDS patient. *(A) SF3B1* variant analysis using DNA-sequencing. *(B) SF3B1* variant analysis using droplet digital PCR. Figures courtesy of Affaf Aliouat.

We first assessed if our computational pipeline could recapitulate previously reported *SF3B1*^{MUT}-associated mis-spliced events. Of the 28 A3SS mis-spliced A3SS events tabulated from the literature (Table 1.1), two were expressed in our cell populations included in our study, namely haematopoietic stem cells (HSCs) and megakaryocyte-erythroid progenitors (MEPs; Figure 6.2A). The percent spliced-in (PSI) of *MAP3K7* A3SS event was significantly increased in *SF3B1*^{K666N} cells relative to *SF3B1*^{WT} cells among the HSCs (Figure 6.2B). The PSI of *SEPTIN6* A3SS event was significantly increased in *SF3B1*^{K666N} and *SF3B1*^{N626D} cells relative to *SF3B1*^{WT} cells among both HSCs and MEPs (Figure 6.2C). It is noteworthy that *SF3B1*^{MUT} -associated *MAP3K7* A3SS mis-splicing has been reported and validated by multiple independent studies (Lee et al., 2018; Lieu et al., 2022; Z. Liu et al., 2020).

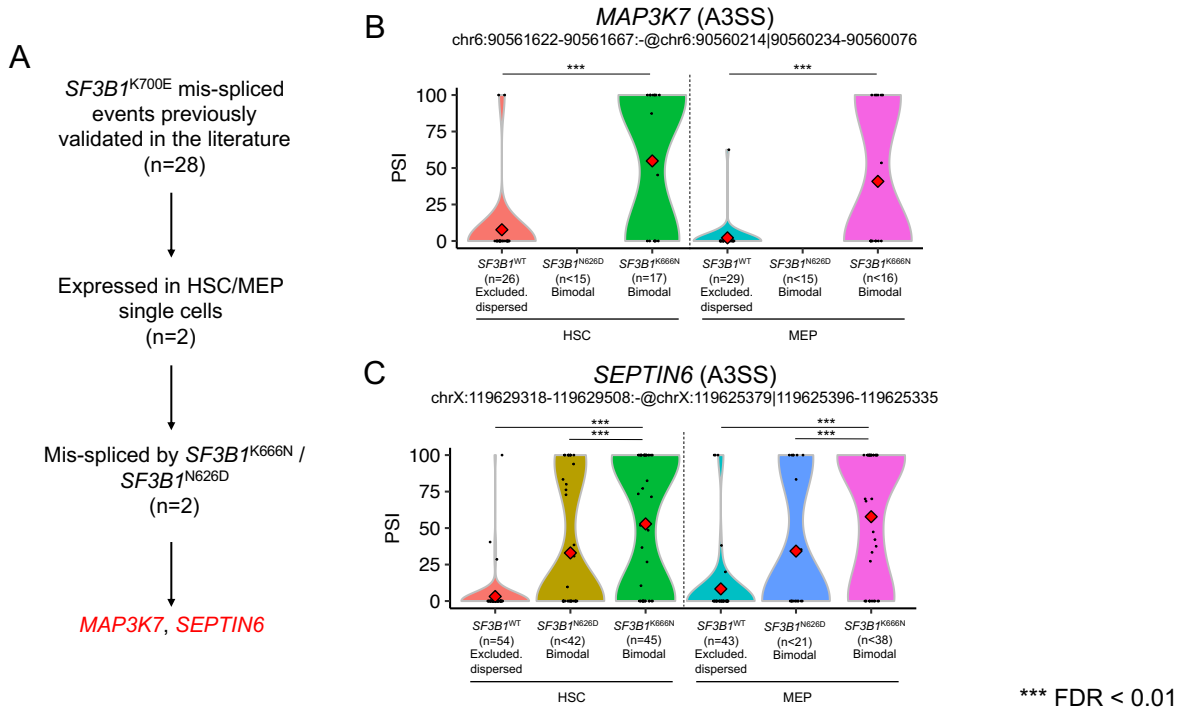


Figure 6.2: Recapitulating previously reported *SF3B1*^{MUT}-associated mis-spliced events in our single-cell dataset. (A) Two of 28 previously reported mis-spliced events were expressed in our dataset, and therefore included in our assessment here. **(B-C)** PSI distributions of **(B)** *MAP3K7* and **(C)** *SEPTIN3* A3SS splicing events across the different *SF3B1* genotypes among HSCs and MEPs. Single-cell dataset generated by Affaf Aliouat under the supervision of Sten Eirik Jacobsen, Eva Hellström-Lingberg, and Seshi Ogawa.

Aften recapitulating previously reported *SF3B1*^{MUT}-associated mis-spliced events, we proceeded with characterising the global transcriptomic landscape of *SF3B1*^{K666N} and *SF3B1*^{N626D} cells. To this end, we performed differential splicing analysis of *SF3B1*^{K666N} and *SF3B1*^{N626D} cells relative to *SF3B1*^{WT} cells (Figure 6.3A). In total, we identified 142, 97, 125, and 90 splicing events with increased PSI (ΔPSI > 10, FDR < 0.10) in HSC *SF3B1*^{K666N}, HSC *SF3B1*^{N626D}, MEP *SF3B1*^{K666N}, and MEP *SF3B1*^{N626D} cells, respectively, relative to *SF3B1*^{WT} cells (Figures 6.3B-E). On the other hand, 151, 129, 132, and 100 splicing events had decreased PSI (ΔPSI < -10, FDR < 0.10) in HSC *SF3B1*^{K666N}, HSC *SF3B1*^{N626D}, MEP *SF3B1*^{K666N}, and MEP *SF3B1*^{N626D} cells, respectively, relative to *SF3B1*^{WT} cells.

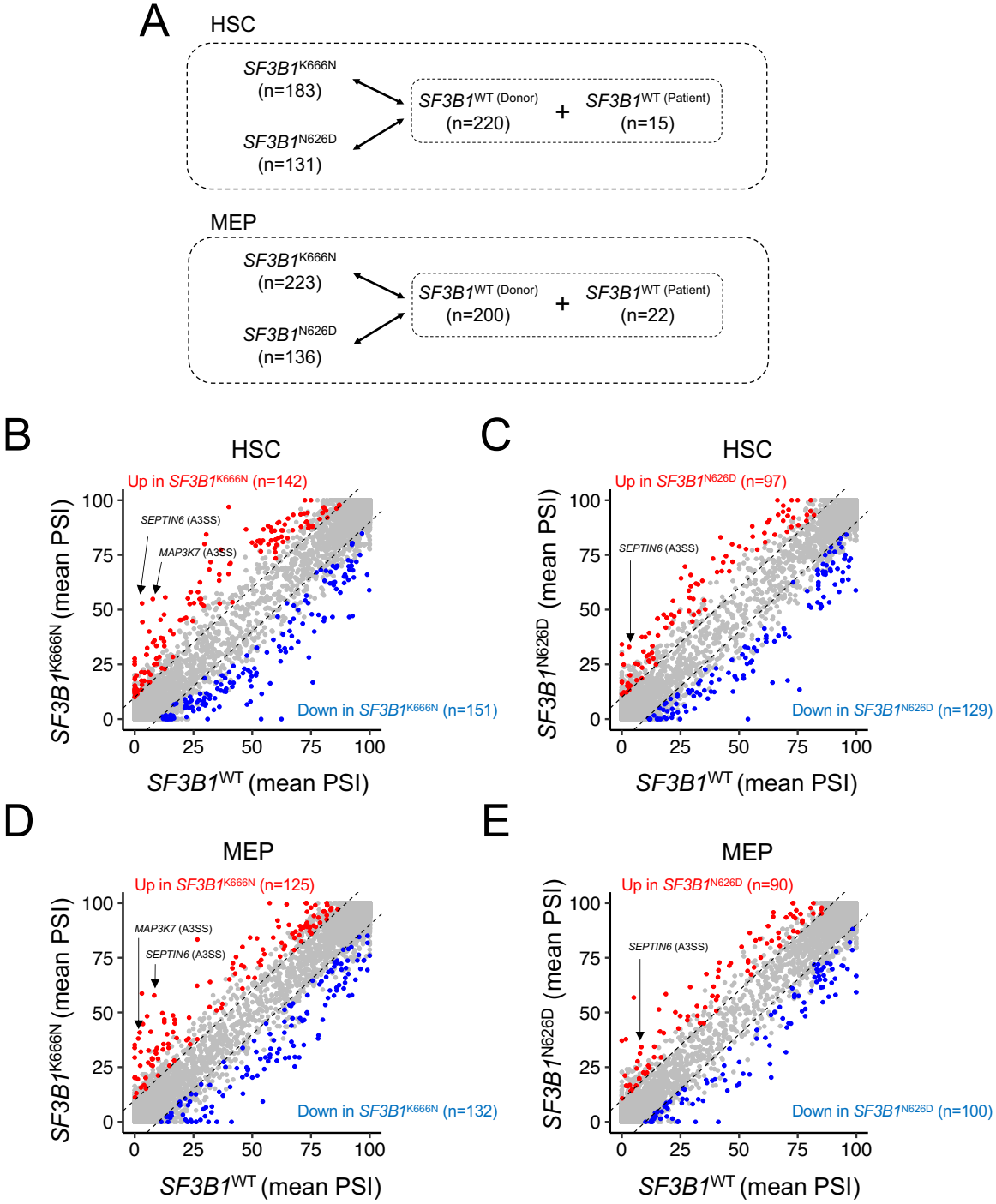


Figure 6.3: Differential splicing analysis between *SF3B1*^{MUT} vs *SF3B1*^{WT} cells. **(A)** Schematic diagram demonstrating the different pair-wise comparison performed for differential splicing analysis. **(B-C)** Differential splicing analysis between HSC **(B)** *SF3B1*^{K666N} and **(C)** *SF3B1*^{N626D} vs *SF3B1*^{WT} cells. **(D-E)** Differential splicing analysis between MEP **(D)** *SF3B1*^{K666N} and **(E)** *SF3B1*^{N626D} vs *SF3B1*^{WT} cells.

Pathway enrichment analysis of differentially spliced genes identified RNA splicing gene sets to be the most significantly enriched for HSC and MEP *SF3B1*^{K666N} and *SF3B1*^{N626D} vs *SF3B1*^{WT} comparisons (Figure 6.4). This re-affirms the role of *SF3B1*^{MUT} in dysregulated RNA splicing (Alsafadi et al., 2016). Additional pathways identified to be enriched among differentially spliced genes included pathways related to ubiquitination, cellular division, transcription, and apoptosis.

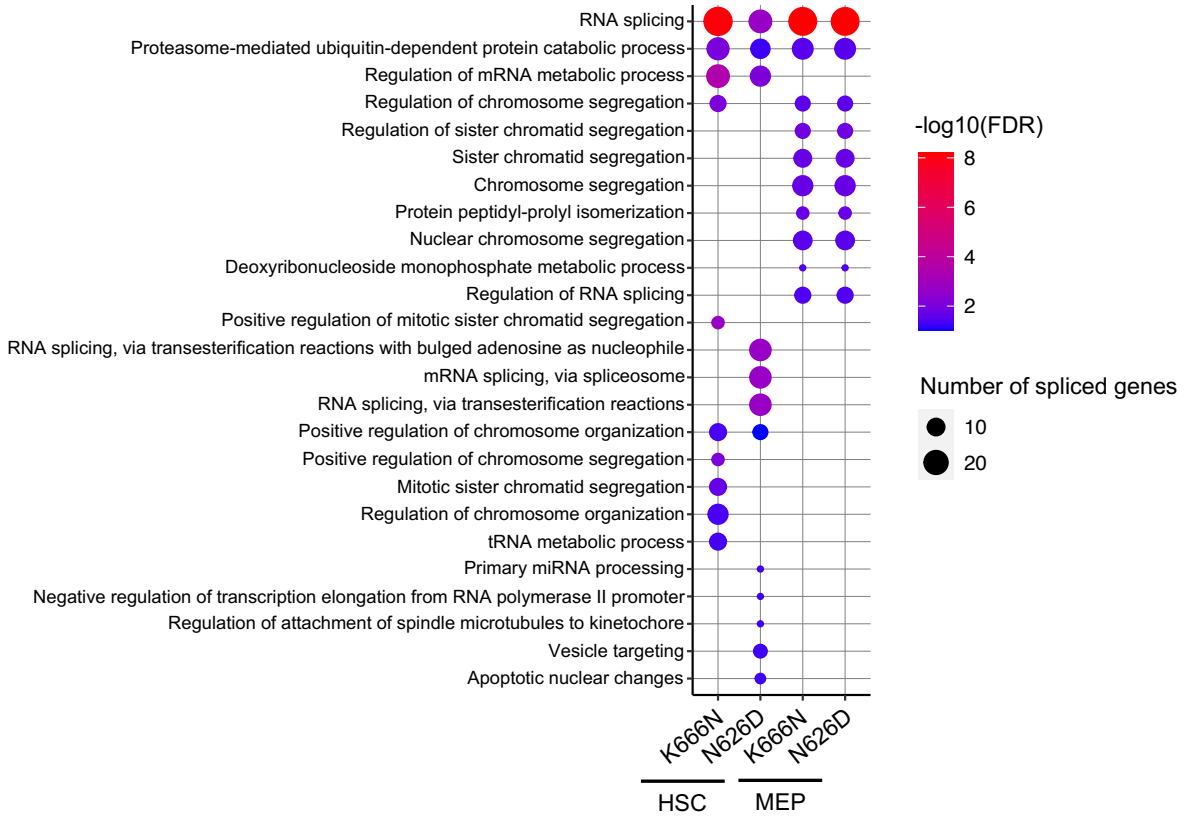


Figure 6.4: Pathway enrichment analysis of differentially spliced genes identified from HSC and MEP *SF3B1*^{K666N} and *SF3B1*^{N626D} vs *SF3B1*^{WT} comparisons (related to Figures 6.3B-D). RNA splicing and ubiquitination pathways enriched across all comparisons.

Using our list of differentially spliced events, we investigated if splicing represents an additional layer of complexity underlying gene expression profile. Principal component analysis (PCA) using differentially spliced events (|ΔPSI| > 10 and FDR < 0.10) expressed in HSC and MEP *SF3B1*^{K666N}, *SF3B1*^{N626D}, and *SF3B1*^{WT}

cell groups successfully delineated *SF3B1*^{MUT} from *SF3B1*^{WT} cells among both HSC and MEP populations (Figure 6.5A). On the other hand, PCA using differentially expressed genes (|log2fc| < 0.5 & FDR > 0.10) delineated HSC from MEP, but not *SF3B1*^{MUT} from *SF3B1*^{WT} cells (Figure 6.5B). Similarly, PCA using highly variable genes delineated HSC from MEP, but not *SF3B1*^{MUT} from *SF3B1*^{WT} cells (Figure 6.5C-D).

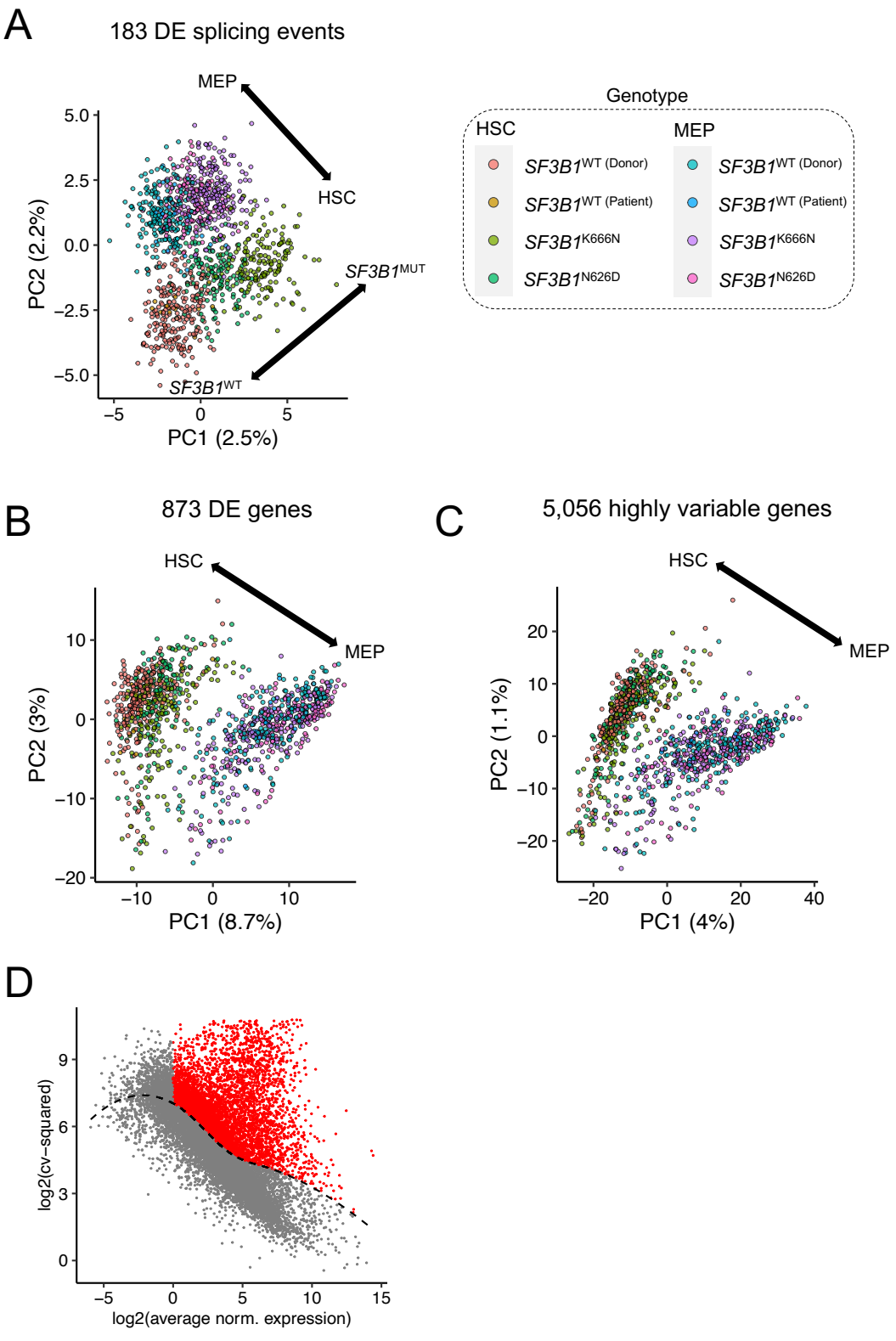


Figure 6.5: Principal component analysis (PCA) to assess the ability of splicing and gene expression values in delineating *SF3B1*^{MUT} from *SF3B1*^{WT} cells. (A-C) PCA using **(A)** differentially spliced events, **(B)** differentially expressed genes, and **(C)**

highly variable genes. **(D)** Highly variable genes (red) identified by fitting a loess curve across the square of coefficient of variance and average normalised gene expression values.

We next sought to validate the novel differentially spliced events identified from our single-cell dataset in a publicly available acute myeloid leukaemia (AML) dataset, namely the BeatAML cohort (Tyner et al., 2018). Because *SF3B1*^{K666N} clone overtook the *SF3B1*^{N626D} as the major clone in our MDS patient, we focused on *SF3B1*^{K666N}-associated mis-spliced events identified from HSCs and MEPs for validation. Furthermore, given the role of *SF3B1*^{WT} and *SF3B1*^{MUT} in recognising canonical and alternative 3' splice site, respectively (Alsafadi et al., 2016), we further focused *SF3B1*^{K666N}-associated A3SS events for validation.

In total, 49 and 30 splicing events with increased and decreased PSI, respectively, in either HSC or MEP *SF3B1*^{K666N} cells were included for validation (Figure 6.6). One splicing event which had decreased PSI in HSC *SF3B1*^{K666N} but increased PSI in MEP *SF3B1*^{K666N} was excluded from validation. Fifteen of 49 (31%) splicing events with increased PSI in *SF3B1*^{K666N} cells were successfully validated in BeatAML whereas five of 30 (17%) splicing events with decreased PSI in *SF3B1*^{K666N} cells were successfully validated in BeatAML.

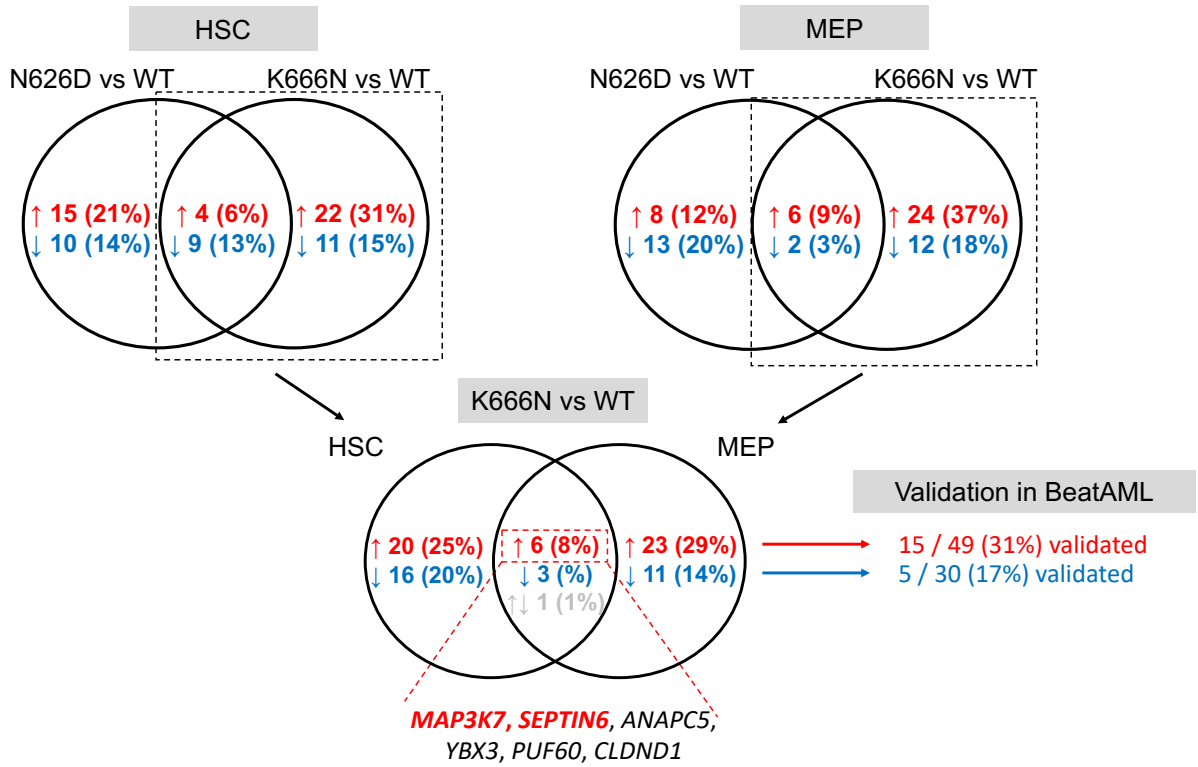


Figure 6.6: Retrieval of *SF3B1*^{K666N}-associated mis-spliced events identified from our single cells for validation using BeatAML. Among the splicing events that were concordantly mis-spliced in both HSC and MEP *SF3B1*^{K666N} cells were *MAP3K7* and *SEPTIN6*, both of which were previously reported in the literature and validated in BeatAML.

Next, we performed additional validation of the *SF3B1*^{K666N}-associated mis-spliced events by investigating the distance between the alternative and canonical 3' splice site for each mis-spliced event. Multiple independent studies have shown that *SF3B1*^{MUT} is associated with A3SS located within ~10-30bp upstream of the canonical splice site (Alsafadi et al., 2016; Z. Liu et al., 2020). Therefore, *bona fide SF3B1*^{K666N}-associated mis-spliced events should be enriched with A3SSs located within this range.

Indeed, we observed *SF3B1*^{K666N}-associated A3SSs that were increased in our single-cell dataset and also validated in BeatAML to be enriched with A3SSs located within 10-30bp upstream of their corresponding canonical splice sites (Figure 6.7). On the other hand, *SF3B1*^{K666N}-associated A3SSs that were decreased in our single-cell dataset but not validated in BeatAML demonstrated a bimodal distribution.

Nevertheless, there were only five splicing events from this category and therefore more down-regulated A3SSs may be required to draw more robust conclusions.

Interestingly, *SF3B1*^{K666N}-associated A3SSs identified from our single cells that were not validated in BeatAML were not enriched for A3SSs located within 10-30bp upstream of their corresponding canonical splice sites. Instead, these unvalidated A3SSs had a uniform distribution for the distance between the A3SSs and their corresponding canonical splice sites. This suggests that unvalidated A3SSs may not be mediated by *SF3B1*^{K666N}. This analysis also highlights the importance of orthogonal validation of novel mis-spliced events to increase the likelihood of identifying true mis-splicing events for downstream experimental studies.

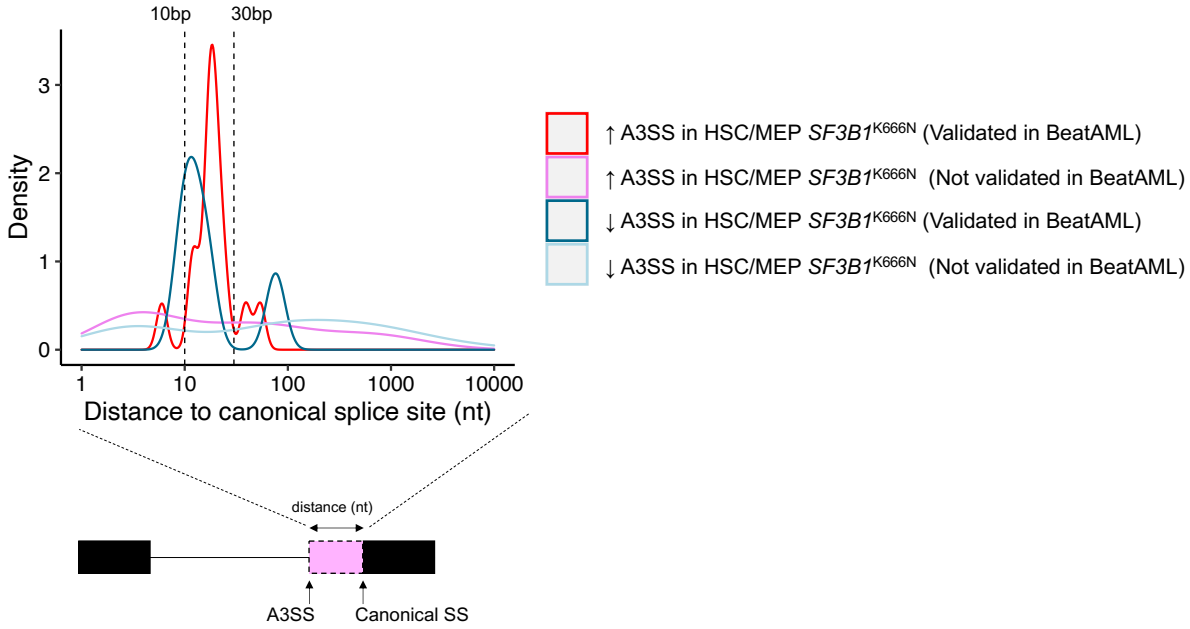


Figure 6.7: The relative distance between A3SSs and their corresponding canonical splice sites. A3SSs stratified by whether the A3SSs had increased or decreased PSI in HSC/MEP *SF3B1*^{K666N} cells and whether the A3SSs were successfully validated in BeatAML.

Among the genes that were mis-spliced by both HSC and MEP *SF3B1*^{K666N} cells and were also successfully validated by BeatAML included *SEPTIN6*, *MAP3K7*, and *ANAPC5* (Figure 6.7). Both *SEPTIN6* and *MAP3K7* mis-splicing have been reported to be associated with *SF3B1*^{MUT} (Dolatshad et al., 2015; Lee et al., 2018).

ANAPC5 has been reported to be mis-spliced in a combined analysis of *SF3B1*^{K700E} chronic lymphocytic leukaemia, breast cancer, and MDS, but has yet to be experimentally validated or characterised.

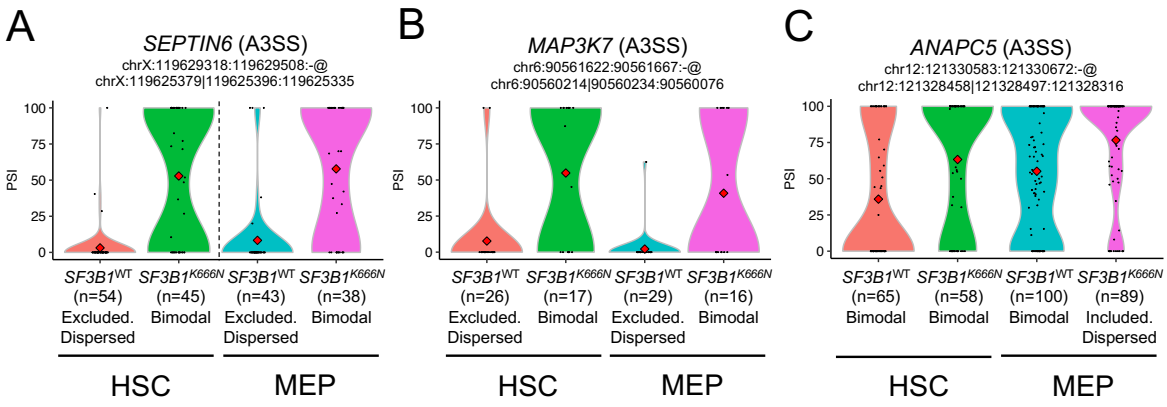


Figure 6.7: PSI profile of *SF3B1*^{K666N}-associated mis-spliced events identified in both HSCs and MEPs.

Among the genes that were mis-spliced by HSC, but not MEP, *SF3B1*^{K666N} cells but were successfully validated by BeatAML included *TCEA2*, *ABLIM1*, *UXS1*, *CELF2*, *EDEM2*, *SLTM*, and *STX4* (Figure 6.8). Notably, *CELF2* is an RNA-binding protein involved in mediating intron retention at 3'-untranslated regions (3' UTRs) (Chatrikhi et al., 2019). *CELF2* was expressed in HSC, but not MEP (Figure 6.8D), and therefore may hint at an HSC-specific role for this gene.

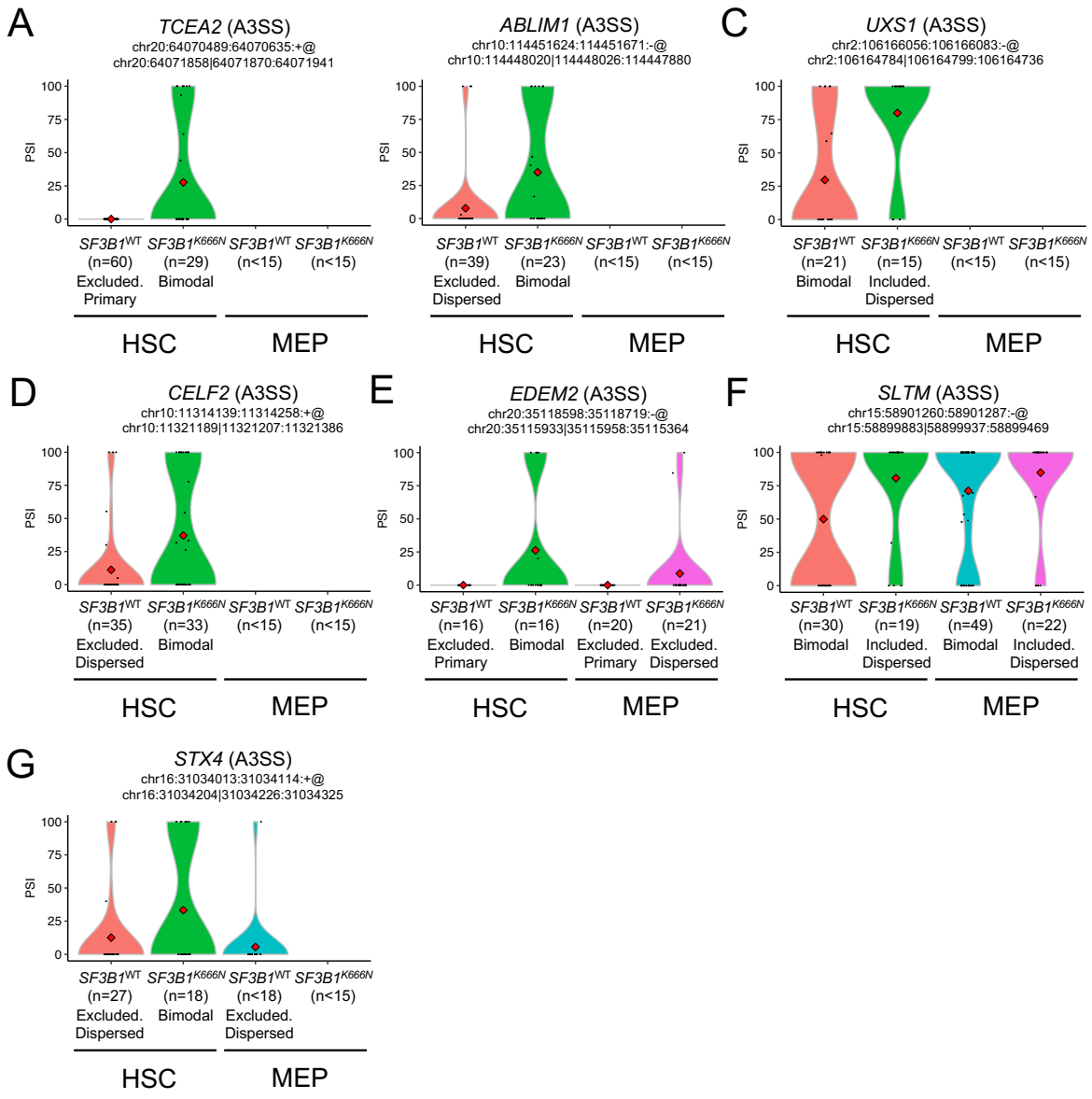


Figure 6.8: PSI profile of HSC-specific *SF3B1*^{K666N}-associated mis-spliced events.

Among the genes that were mis-spliced by MEP, but not HSC, *SF3B1*^{K666N} cells but were successfully validated by BeatAML included *ERGIC3*, *EI24*, *SERBP1*, *ERCC3*, and *ZDHHC16* (Figure 6.9). Notably, *EI24* is regulated by p53 and has been shown to contribute to pancreatic cancer cell proliferation (Hwang et al., 2019; Zhao et al., 2012). Furthermore, *SERBP1* has been shown to contribute to glioblastoma development (Kosti et al., 2020).

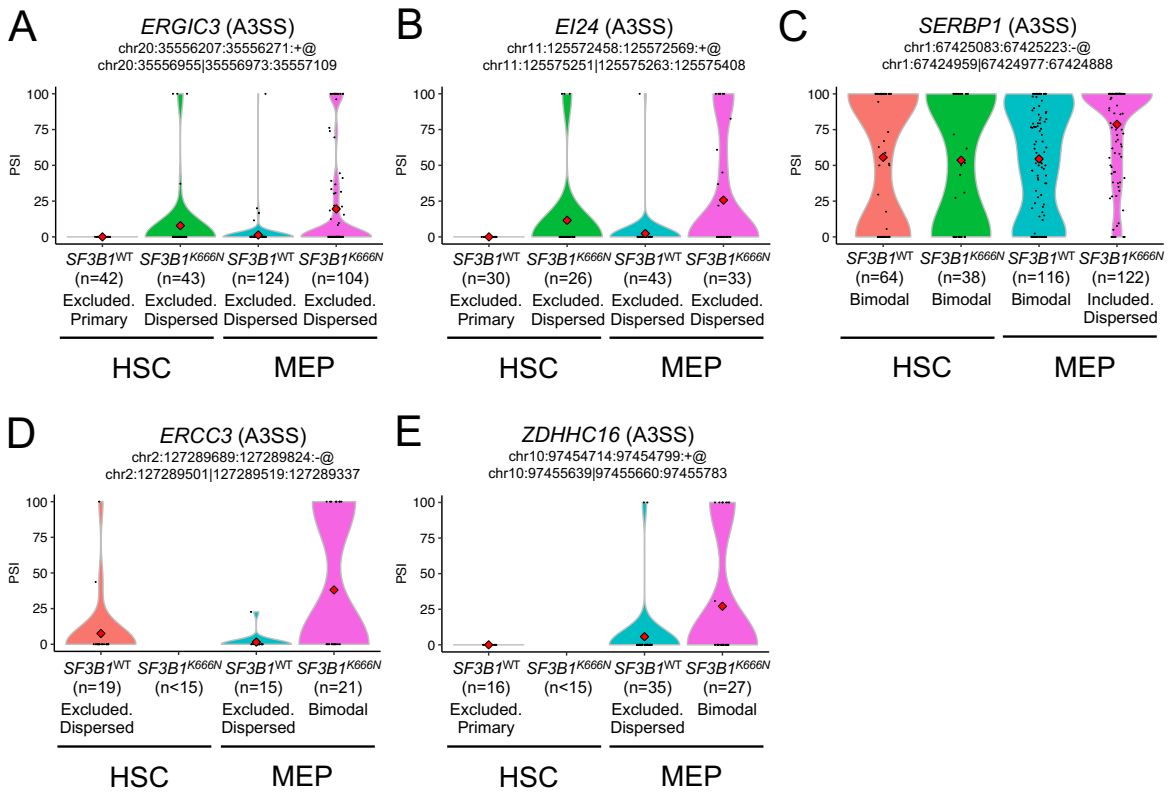


Figure 6.9: PSI profile of MEP-specific *SF3B1*^{K666N}-associated mis-spliced events.

SF3B1^{K666N}-associated A3SS usage has been shown to potentially disrupt the open reading frame (ORF) of the corresponding gene. Specifically, the insertion (splicing in) of A3SS may introduce a premature stop codon (PTC) into the ORF (Figure 6.10A). This may consequently lead to nonsense-mediated decay (NMD) of the mRNA transcript, and this may ultimately be reflected as a decreased in the corresponding gene and protein expression levels (Shiozawa et al., 2018). Therefore, NMD analysis may aid in identifying potential candidate mis-spliced genes for downstream experimental studies.

Here, we performed NMD prediction on all 15 *SF3B1*^{K666N}-associated A3SSs identified in our single cells that were also validated in the BeatAML cohort. Two of which were predicted to be subjected to NMD, namely *MAP3K7* and *STX4* (Figure 6.10B). *MAP3K7* A3SS was specifically mis-spliced in *SF3B1*^{K700} and *SF3B1*^{K666} patients, but not in patients with variants in splicing factors *SRSF2* or *U2AF1* (Figure

6.10C). Furthermore, *MAP3K7* gene expression was significantly down-regulated in *SF3B1*^{K666} patients (Figure 6.10D).

On the other hand, *STX4* A3SS mis-splicing was not specific to *SF3B1*^{K700} and *SF3B1*^{K666} patients. It was also mis-spliced in *SRSF2*^{P95} and *U2AF1*^{S34} patients (Figure 6.10E). Furthermore, *STX4* gene expression was significantly up-regulated in *SF3B1*^{K700} patients but significantly down-regulated in *SF3B1*^{K666} patients (Figure 6.10F). Therefore, compared to *MAP3K7*, it is more difficult to attribute *STX4* mis-splicing directly to *SF3B1*^{K666} and to attribute *STX4* reduced gene expression in *SF3B1*^{K666} patients directly to NMD.

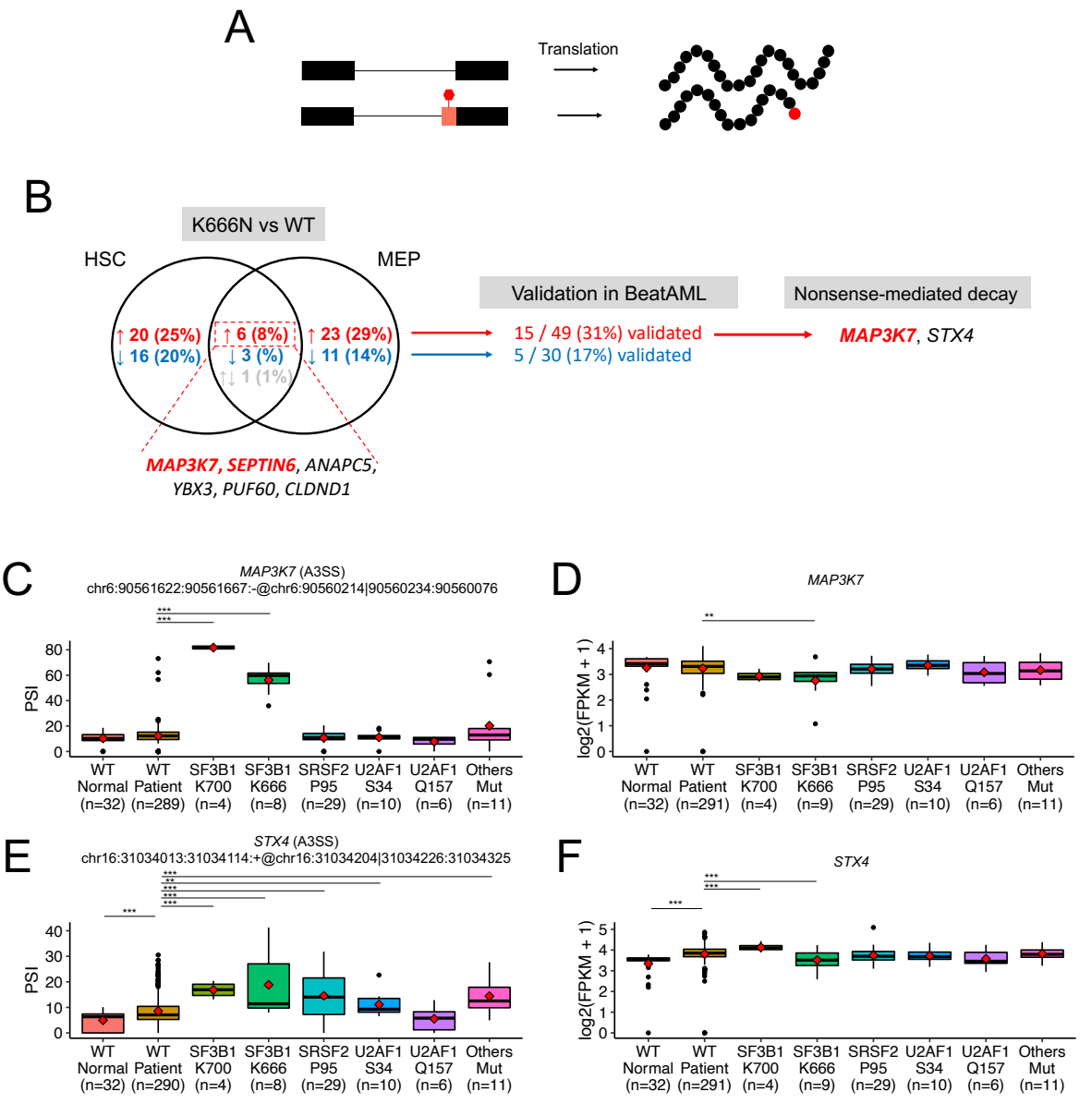


Figure 6.10: Nonsense-mediated decay (NMD) prediction on *SF3B1*^{K666N}-associated mis-spliced events identified from HSC or MEP and validated in BeatAML. (A) Schematic diagram illustrating the creation of a premature stop codon (PTC) by an A3SS. **(B)** *MAP3K7* and *STX4* predicted to undergo splicing-mediated NMD. **(C-D)** PSI distribution and gene expression profile of *MAP3K7* across BeatAML patients stratified by splicing factor hotspot variants. **(E-F)** PSI distribution and gene expression profile of *STX4* across BeatAML patients stratified by splicing factor hotspot variants. FDR *** < 0.01, ** < 0.05, * < 0.10, n.s. non-statistically significant.

Taken together, *MAP3K7* is the most robust candidate mis-spliced genes identified from our single cell analysis and it is noteworthy that the functional consequence of splicing-mediated NMD of *MAP3K7* has been extensive characterised Specifically, *SF3B1*^{MUT}-associated mis-splicing of *MAP3K7* has been shown to hyperactivate the NF-kB signalling pathway and also lead to higher rates of erythrocyte apoptosis (Lee et al., 2018; Lieu et al., 2022). The former observation suggests a role of *MAP3K7* mis-splicing in driving MDS while the latter observation suggests a role of *MAP3K7* mis-splicing in anaemia development in MDS patients.

6.2 Single-cell analysis of *SRSF2*-mutant MDS patients

SRSF2 is a member of the serine/arginine-rich (SR) protein family. It consists of an RNA recognition motif (RRM) and a domain rich in arginine and serine residues (RS domain). SRSF2 mediates the interaction between U1 snRNP with 5' and 3' splice sites during splicing. SRSF2 also mediates the interaction between U2 snRNP and branchpoint-point sequence during splicing (Fu & Maniatis, 1992).

Genetic variants in *SRSF2* are found in ~10-20% of myelodysplastic syndrome (MDS) patients (Pellagatti et al., 2018; Shiozawa et al., 2018). Majority of genetic variants identified in *SRSF2* is P95 whereby a point mutation changes the proline (P) residue into histidine (H), leucine (L), threonine (T), or arginine (R) residue (Liang et al., 2018; Pellagatti et al., 2018; Shiozawa et al., 2018). The P95 amino acid is located on the hinge region of SRSF2 which is sandwiched between the RRM domain towards the N-terminal and RS domain towards the C-terminal. Paired whole-genome sequencing and scRNA-seq of bone marrow sample from an AML patient identified *SRSF2*^{P95} as the major clone and in the haematopoietic stem cell (HSC) and common

myeloid progenitor (CMP) compartments (Petti et al., 2019). This suggests that *SRSF2*^{P95} is an early event in blood myeloid neoplasm development.

Nevertheless, a comprehensive characterisation of SRSF2 P95-mediated aberrant splicing across the different cellular compartments is lacking. Identifying *SRSF2*^{P95}-mediated aberrant splicing in haematopoietic stem and progenitor cells (HSPCs) may aid in identifying biomarkers for targeted therapy and to understand the mechanism by which *SRSF2*^{P95}-mediated aberrant splicing leads to disease development and progression. To this end, we have performed scRNA-seq on CD34+ cell population from six individuals consisting of five *SRSF2*^{P95} MDS patients and one healthy donor (Table 6.1).

Patient ID	Genetic variants									
	SRSF2	ASXL1	TET2	RUNX1	NRAS	SETBP1	ETV6	PTPN11	STAG2	BCOR
PV1506	p.P95H	p.Y591_Q592delinsX		p.384_384del	p.G12R	p.S869R	p.L205fs	p.D61G		
PV1553	p.P95H		p.A1341P | p.F1429fs							
MAN973	p.P95L	p.G646Wfs*12	p.H682Ifs*18							
PV1488	p.P95L		p.Q1834*							
MAN543	p.P95L	p.A735Lfs*9		p.R204*					p.R259*	p.S226Vfs*75
NOC161 (Healthy)										

Table 6.1: Genetic variant profile of myeloid neoplasm genes in the six individuals included in this study. Data generated by Juseong Lee under the supervision of Andrea Pellagatti and Jacqueline Boultwood.

Sequencing was performed in two batches. In each batch, three samples were pooled together. Sequencing was performed using Chromium Single Cell 3' Reagent Kit (v3.1 Chemistry) with Feature Barcoding technology for Cell Surface Protein. Therefore, each sample was tag with their respective hashtag oligo (HTO). To assign each cell to their donor of origin (demultiplexing), we developed a hybrid approach consisting of using both HTO and transcriptome-wide genetic variants to assign as many cells as possible to their donor of origin. Specifically, we used Cell Ranger *count* module to quantify the HTO expression for each cell and we also used Souporcell to cluster cells based on their transcriptome-wide genetic variants (Heaton et al., 2020). Our approach for demultiplexing for one pool of three samples is illustrated in Figure 6.11. Here, HTO-1, HTO-2, and HTO-3 correspond to patients PV1506, PV1553, and MAN973, respectively.

Using Seurat, each cell was classified into four broad categories, namely singlet, doublet, negative, or unassigned. Upon scrutiny of each HTO expression

distribution, we further categorise singlets into five categories (Figure 6.11A and B). Therefore, each cell may be classified into one of eight categories.

- Singlet (HTO-1), singlet (HTO-2), and singlet (HTO-3) were cells with high expression of HTO-1, HTO-2, and HTO-3, respectively.
- Singlet (low HTO signal) were cells with low expression of all three HTOs.
- Singlet (mixed HTO signal) were cells with high expression of more than two HTOs.
- Doublet, negative, and unassigned as per the original classification by Seurat.

Re-analysis of HTO expression for each category confirmed that singlet (HTO-1), singlet (HTO-2), and singlet (HTO-3) had high expression of HTO-1, HTO-2, and HTO-3, respectively (Figure 6.11C). However, ~20% of cells were classified as negative. This represents a substantial number of cells in which the donor of origin could not be identified. Therefore, we proceeded with Souporcell to assess if we were able to rescue these cells with missing donor identity.

Souporcell identified three clusters of cells. Cluster 0, 1, and 2 were enriched with cells with HTO-2, HTO-3, and HTO-1, respectively (Figure 6.11D). All cells within the same cluster were assigned to the more prevalent HTO identified within the cluster. For example, in Cluster 0, ~10% with cells with previously undetermined donor identify was assigned to HTO-2. Based on Souporcell, each cell was classified into three categories, namely singlet, doublet, and unassigned (Figure 6.11E). ~95% of cells were successfully assigned to their donor of origin.

Cross tabulation of HTO- and Souporcell-based classifications identified 1,379 cells that were previously assigned as negative were now assigned as belonging to donor with HTO-1 (Figure 6.11F). Combining the classifications of HTO and Souporcell, ~95% of cells were successfully and confidently assigned to their donor of origin (Figure 6.11G). Taken together, HTO may be first used to identify the donor of origin for majority of the cells, and Souporcell may be subsequently used to rescue cells with low or missing HTO information.

Next, cells were further filtered based on unique molecular identifier (UMI) counts and number of genes detected per cell (Figure 6.11H-J). Specifically, cells with <1,000 UMIs or <500 detected genes were excluded for all samples. On the other hand, the upper limit, i.e., the number of UMIs and detected genes above which the

cells were excluded, was determined for each sample individually. Additionally, cells with >5% mitochondrial reads were excluded for all samples. Quality control metrices are summarised in Table 6.2.

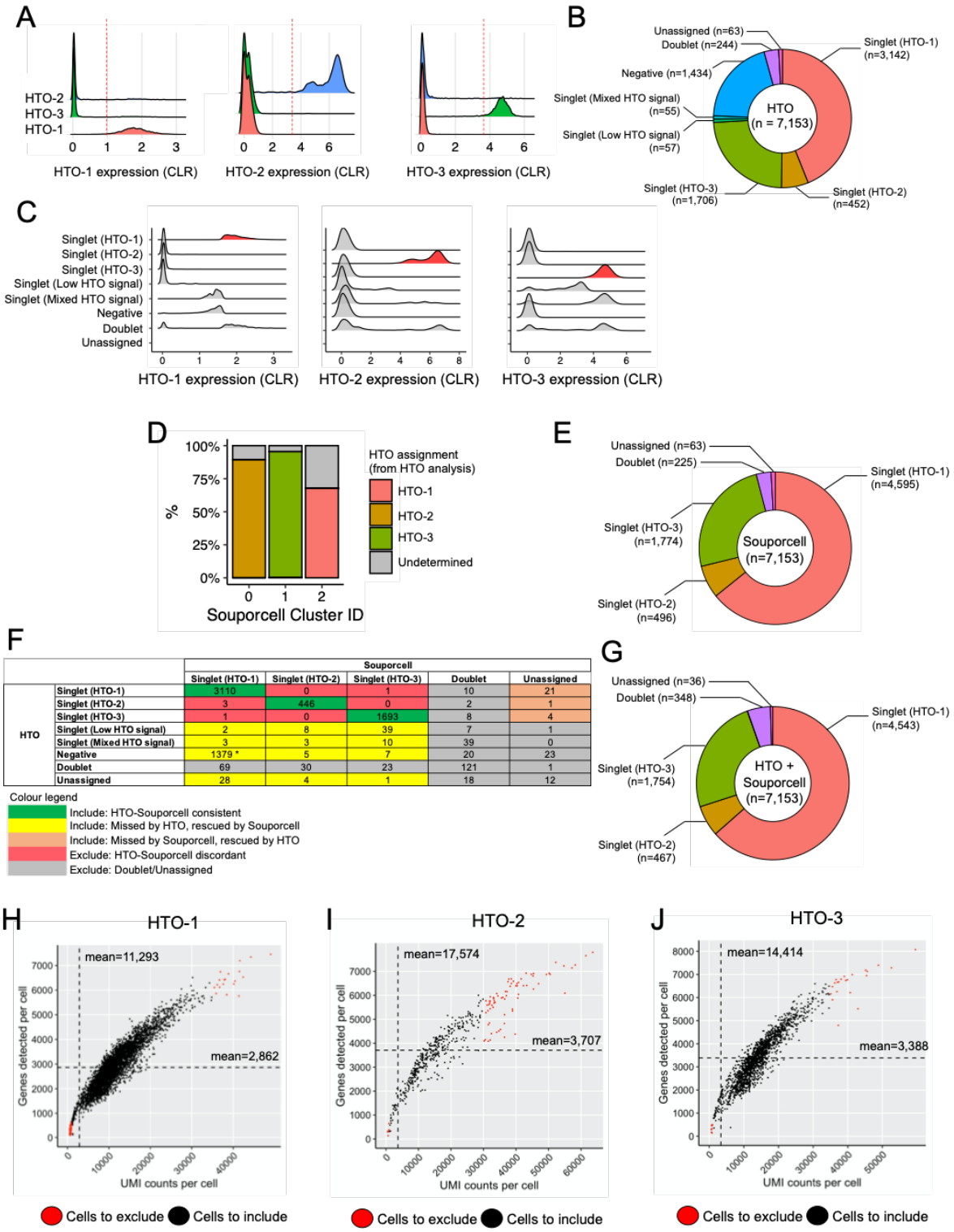


Figure 6.11: Hybrid approach for sample demultiplexing using HTO and Souporcell. Results for one pool of sample consisting of PV1506 (HTO-1), PV1553 (HTO-2), and MAN973 (HTO-3) illustrated here. (A) HTO expression distribution for each HTO. Vertical red dashed line indicates the expression threshold to assign cells to their donor of origin. Cells to the right of the threshold were considered to have high expression of the corresponding HTO. **(B)** Classification of cells based on HTO expression. **(C)** HTO expression based on classification of (B). **(D)** Cell clusters identified by Souporcell and annotated with % cells based on previous HTO classification. **(E)** Classification of cells based on Souporcell. **(F)** Cross tabulation of HTO and Souporcell classifications. **(G)** Final cell classification by combining results from both HTO and Souporcell. **(H-J)** Filtering of cells based on UMI counts and number of detected genes.

Table 6.2: Quality control metrices and the number of cells passing each quality metric. While the capture rates were variable and modest, the demultiplexing rate and cells with sufficient UMI counts and genes detected were high (>90%).

Sample ID	Donor ID	HTO ID	No. of cells pooled	No. of cells sequenced (cellranger)	No. of cells successfully demultiplexed (HTO + Souporcell)	No. of cells passed no. of genes, no. of UMIs, % MT QC (SingCellaR)
1	**PV1506**	**HTO1**	17,500	7,153 (30%)	4,543	4,372
	PV1553	**HTO2**	1,400		467	371
	MAN973	**HTO3**	5,250		1,754	1,689
	Total		24,150		6,764 (95%)	6,432 (95%)
2	**PV1488**	**HTO1**	3,500	2,757 (8%)	300	239
	NOC161	**HTO2**	12,250		827	750
	MAN543	**HTO3**	17,500		1,433	1,364
	Total		33,250		2,560 (93%)	2,353 (92%)

After successful assigning of majority of cells to their donor of origin and filtering for cells with sufficient UMIs and genes, we proceeded with integration of all cells for downstream analysis. Dimension reduction analysis revealed cells to cluster by sequencing batch and donor ID (Figures 6.12A and B). Therefore, we corrected for these confounding factors using Harmony and re-performed our dimension reduction analysis using the new embeddings from Harmony (Korsunsky et al., 2019). We observed cells no longer cluster by sequencing batch or donor ID (Figures 6.12C and D). Moreover, cells were observed to cluster by lineage identity when assessed using main cell lineage markers, namely erythroid, megakaryocyte, and myeloid gene expression markers (Figure 6.12E). This suggests that our integration approach was

successful based on mitigating sequencing batch and donor ID as confounding factors, while enabling cells with similar identity to cluster in close proximity.

Louvain clustering identified 22 cell clusters. We perform gene set enrichment analysis (GSEA) using 75 HSPC gene sets and subsequently assigned each cluster to their respective cell type (G. Wang et al., 2022) (Figure 6.12F). Based on our cell type annotation, we collapsed the original 22 cell clusters into 20 cell clusters (Figure 6.12G). Annotation of cell types on our UMAP indicated UMAP-1 distinguished myeloid from MEP cell populations whereas UMAP-2 distinguished HSC/MPP from lymphoid cell populations.

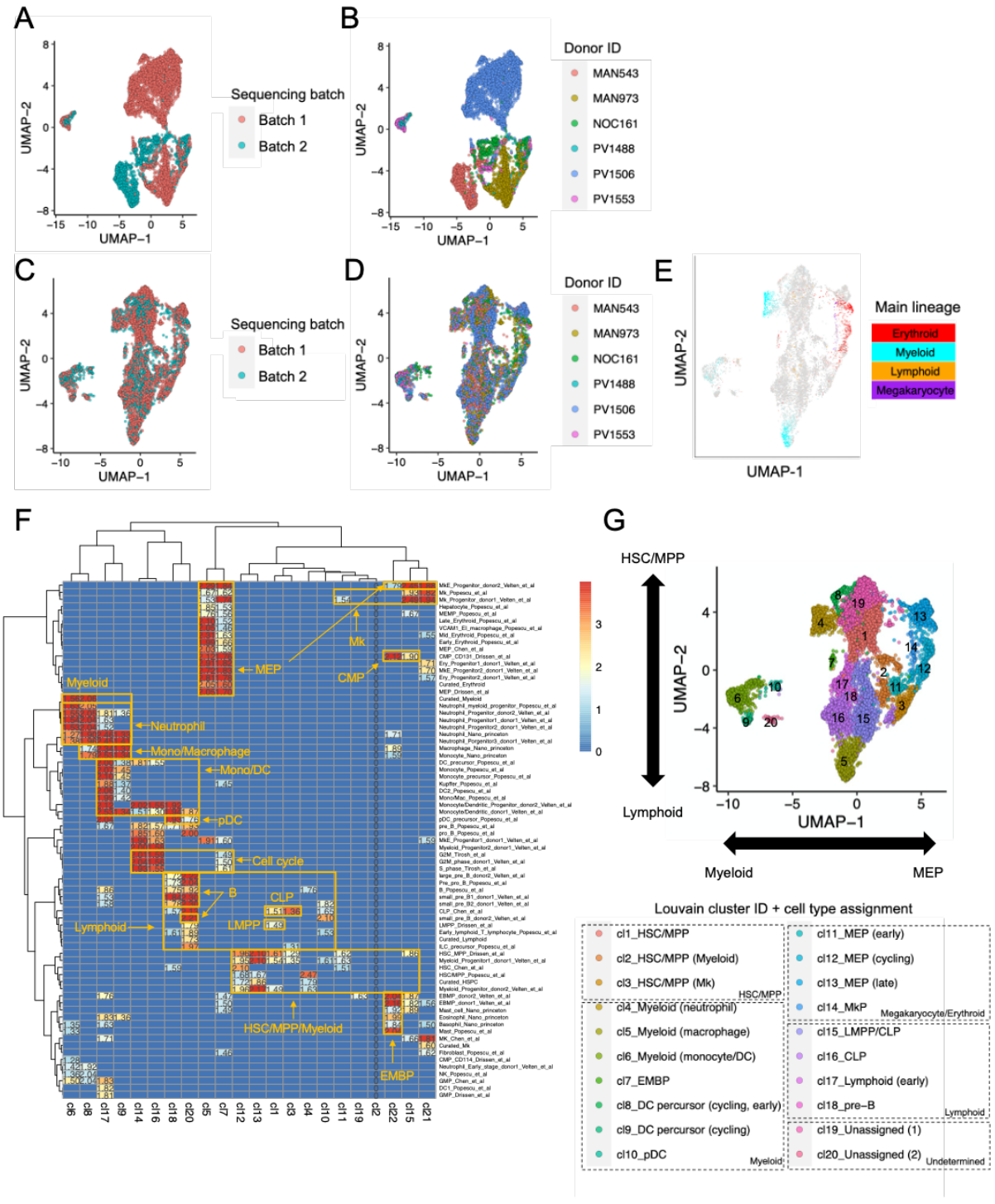


Figure 6.12: Sample integration and cell type annotation. (A-B) Cells clustered by sequencing batch and donor ID. **(C-E)** Cells no longer clustered by sequencing batch and donor ID after correction, but rather, cells clustered by their cell identity based on main cell lineage gene expression markers. **(F)** GSEA on each cluster using 75 HSPC gene sets. **(G)** UMAP annotated with cell type annotation determined from (F).

To identify the cell types with *SRSF2*^{P95} variant for differential splicing analysis between patients and healthy donor, we performed single-cell genotyping using VarTrix (Petti et al., 2019). We observed *SRSF2*^{P95} variant in all cell types, including the early HSC/MPP populations and downstream HSPC populations including myeloid, megakaryocyte-erythroid, and lymphoid cell populations (Figure 6.13). This is consistent with previous single-cell *SRSF2*^{P95} genotyping of an AML patient that demonstrated the presence of *SRSF2*^{P95} variant in HSCs and in more differentiated cell types (Petti et al., 2019).

It is noteworthy that the genotyping rate here is potentially low due to high-dropout rate inherent to scRNA-seq generated droplet-based platforms and also due to stochasticity of RNA expression across the single cells (Petti et al., 2019). Recent technological advances, such as GoT-Splice improved genotyping rate in scRNA-seq generated from droplet-based platforms. GoT-Splice leverages on the full-length cDNA generated during droplet-based library preparation. Specifically, targeted amplification of variant sites enabled confident assignment of mutant and wildtype cells whereas long-read sequencing of full-length cDNAs enabled full-length isoform analysis.(Gaiti et al., 2022)

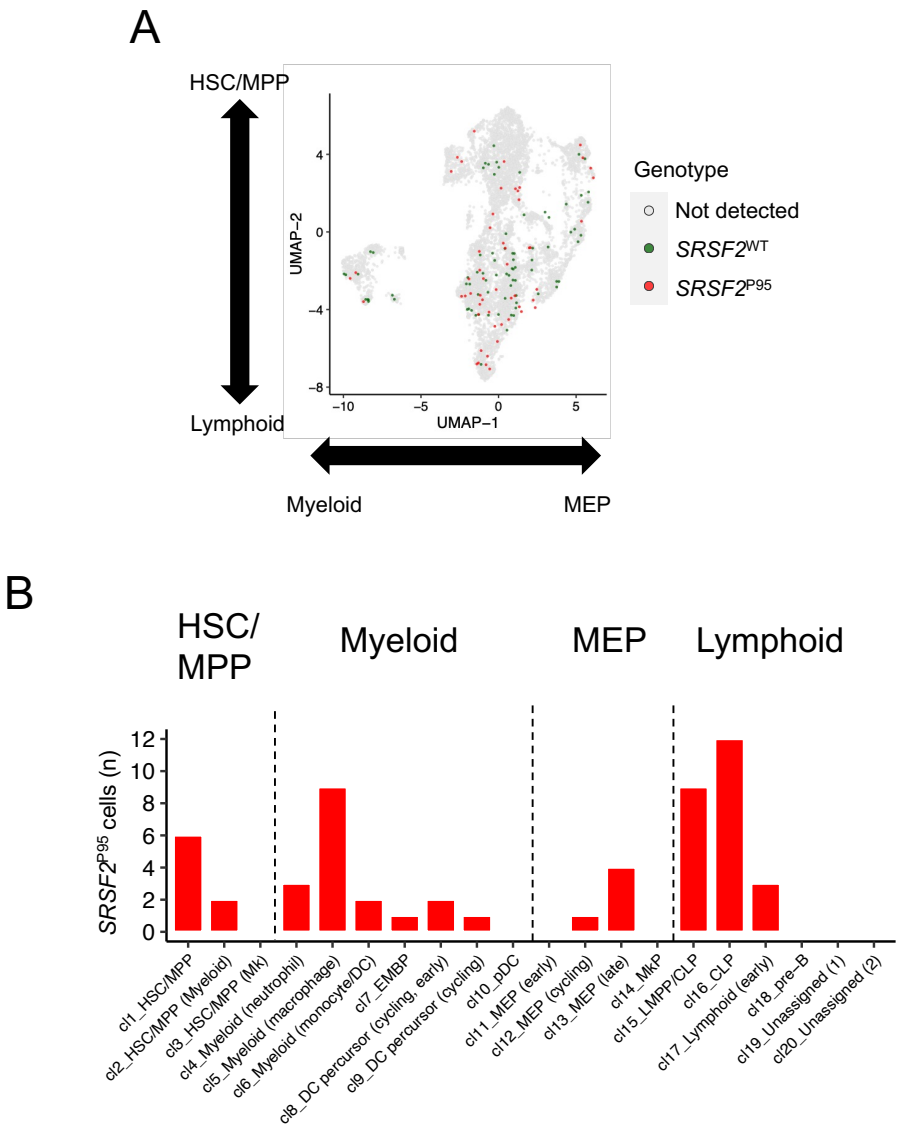


Figure 6.13: Single-cell *SRSF2*^{P95} genotyping. **(A)** UMAP annotated with *SRSF2*^{P95} genotype. **(B)** Number of *SRSF2*^{P95} cells in each cell population.

Moving forward, we focused our differential splicing analysis with MARVEL on the HSC/MPP and myeloid cell populations because (1) HSC/MPP populations may contain relevant *SRSF2*^{P95}-mediated spliced genes that may serve as therapeutic biomarkers, (2) candidate *SRSF2*^{P95}-mediated spliced genes identified in myeloid cell populations may explain clinical phenotype of MDS patients such as myeloid expansion and dysplasia, and (3) combining both HSC/MPP and myeloid cell populations may increase our statistical power to identify *SRSF2*^{P95}-mediated spliced genes.

We first assessed the ability of our single-cell splicing computational framework to identify previously reported *SRSF2*^{P95}-associated splicing events. Specifically, we assess if *EZH2* exon 11 was differentially spliced between *SRSF2*^{P95} patients and healthy donor (Figure 6.14A). We chose this splicing event for assessment because this is the most reproducible splicing event reported to be differentially spliced by *SRSF2*^{P95} in the literature (Rahman et al., 2020; Shiozawa et al., 2018; Tyner et al., 2018; Wheeler et al., 2022). *SRSF2*^{P95} mediates the inclusion (splicing in) of this exon and consequently disrupt *EZH2* open reading frame by introducing a premature stop codon (PTC).

Consistent with previous reports, we observed the splice junctions that support the inclusion of *EZH2* exon 11 to be more highly expressed in *SRSF2*^{P95} patients among HSC/MPP and myeloid cell populations (Figures 6.14B and C). Conversely, the splice junction that support the exclusion (splicing out) of *EZH2* exon 11 was down-regulated in *SRSF2*^{P95} patients among HSC/MPP and myeloid cell populations (Figure 6.14D). These observations were generalisable to majority of the cell populations identified in this study (Figures 6.14E-G).

Interestingly, there was no differences in *EZH2* gene expression levels between *SRSF2*^{P95} patients and healthy donor (Figure 6.14H). Therefore, differential *EZH2* exon 11 usage by *SRSF2*^{P95} would have been missed based on gene expression analysis alone. Moreover, although a PTC was introduced by this novel exon, and nonsense-mediated decay (NMD) was presumed to have taken place, this did not affect the corresponding gene expression levels. This suggests that NMD may not always be reflected in down-regulation of gene expression and/or that there are other factors that may contribute to gene expression regulation aside from splicing-mediated NMD.

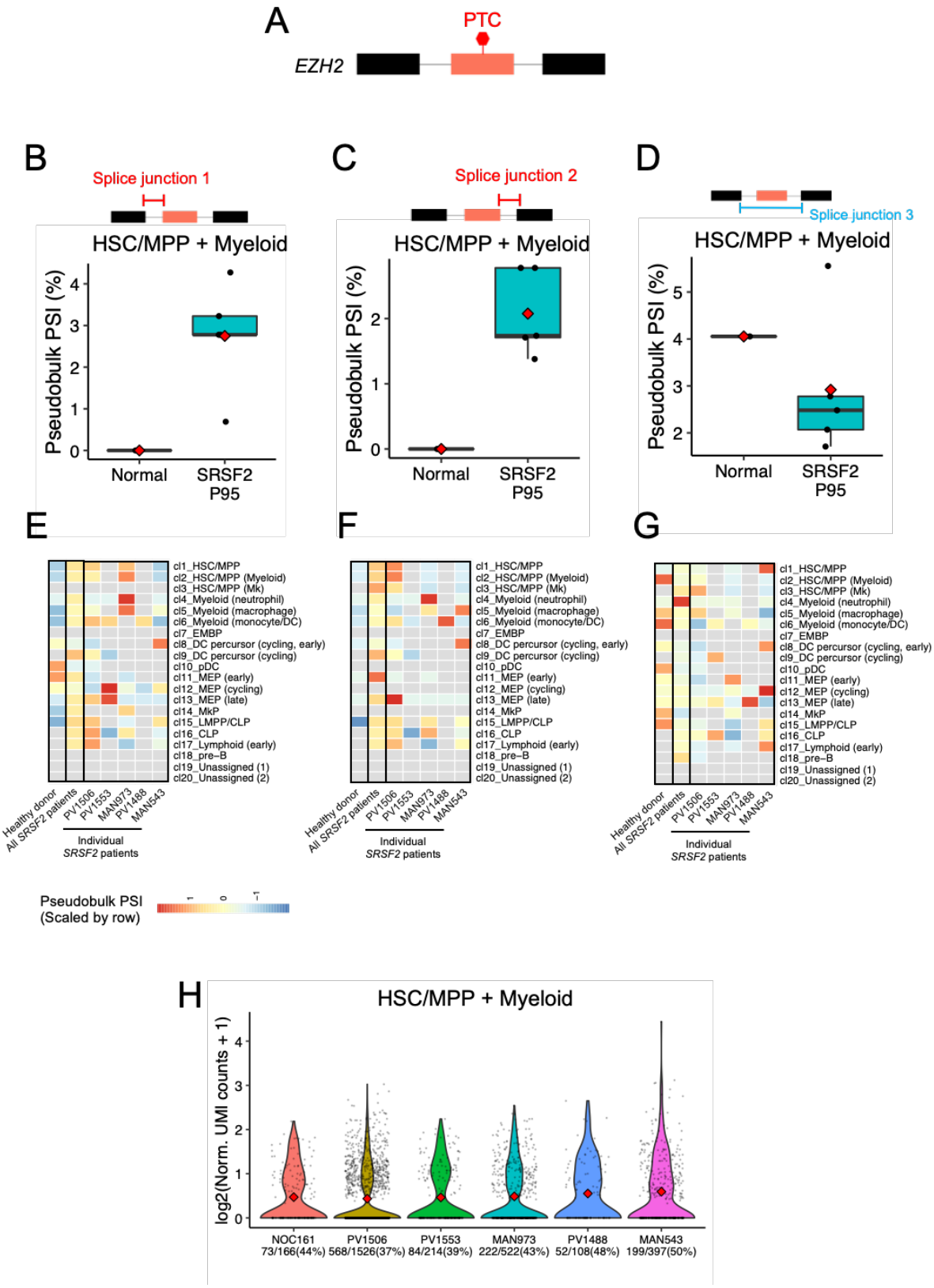


Figure 6.14: Differential *EZH2* exon 11 usage by *SRSF2*^{P95} patients. **(A)** Exon 11 inclusion (splicing in) introduces a PTC into the open reading frame of *EZH2*. **(B-D)** Differential exon 11 usage by HSC/MPP and myeloid cell populations of *SRSF2*^{P95}

patients. **(E-G)** Splicing rate of exon 11 across all cell populations and across all *SRSF2*^{P95} individuals. **(H)** *EZH2* gene expression profile in HSC/MPP and myeloid cell populations across all individuals.

Transcriptome-wide differential splicing analysis identified 77 splice junctions that were up-regulated in *SRSF2*^{P95} patients and 314 splice junctions that were down-regulated in *SRSF2*^{P95} patients (Figure 6.15A). Differentially spliced junctions were defined as |ΔPSI| > 5 and *P* value < 0.05 and average log2 gene expression > 0.5. Pathway enrichment analysis identified pathways associated with immune response, inflammation, cell cycle, and stem cell to be enriched among on differentially spliced genes (Figure 6.15B). Lastly, majority of differentially spliced genes (95%) had no concurrent differences in gene expression levels between *SRSF2*^{P95} patients and healthy donor, i.e., isoform-switching (Figure 6.15C). Therefore, majority of differentially spliced genes would have been missed based on differential gene expression analysis alone.

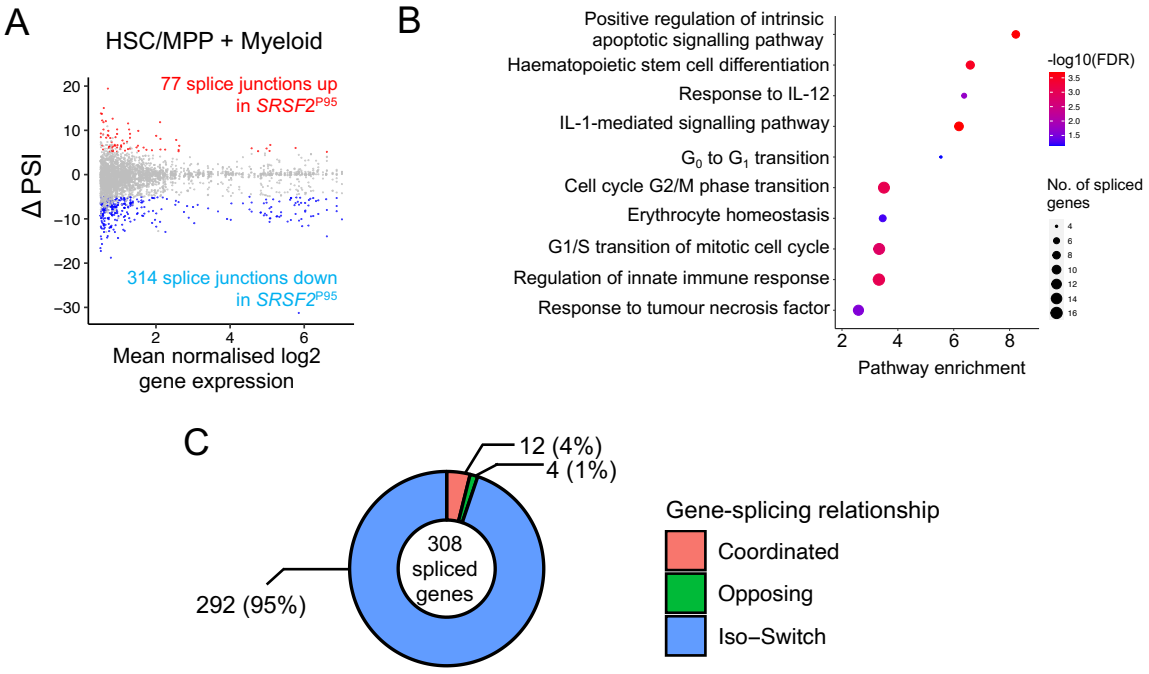


Figure 6.15: Transcriptome-wide differential splicing analysis between *SRSF2*^{P95} patients vs healthy donor. (A) Volcano plot presentation of differential splicing analysis results. **(B)** Pathway enrichment analysis of differentially spliced genes. **(C)**

Classification of changes in splice junction usage relative to changes in gene expression profile between *SRSF2*^{P95} patients and healthy donor.

Taken together, we have demonstrated the utility of our single-cell splicing analytical framework on RNA-seq data generated using a droplet-based library preparation method (10x Genomics) from MDS patient samples consisting of heterogeneous cell populations. We were successful in recapitulating a previously reported *SRSF2*^{P95}-associated splicing event, namely *EZH2* exon 11, in our analysis here. Furthermore, transcriptome-wide differential splicing analysis identified potential biologically relevant pathways enriched among differentially spliced genes that may explain MDS phenotype and may be amenable for therapeutic development.

6.3 Single-cell analysis of *U2AF1*-mutant MPN patients

We have demonstrated the application of our single-cell splicing analysis pipeline on RNA-seq data generated with plate- and droplet-based library preparation methods on HSPCs obtained from myeloid neoplasm patients with *SF3B1* and *SRSF2* genetic variants, respectively. In addition to *SF3B1* and *SRSF2*, another commonly mutated splicing factor gene is *U2AF1*. Together, individuals with genetic variants in *SF3B1*, *SRSF2*, and *U2AF1* constitute >50% of MDS and MPN patients (Grinfeld et al., 2018; Pellagatti et al., 2018; Schischlik et al., 2019; Shiozawa et al., 2018).

Two hotspot *U2AF1* variants have been reported, namely the more common S34 and less common Q157. To date, *U2AF1*^{S34}-associated mis-spliced events have been investigate in bulk samples (Shiozawa et al., 2018; Shirai et al., 2015). Nevertheless, the high variant allele frequency of *U2AF1*^{S34} suggests that this variant arises in the early stages in haematopoiesis, possibly in the HSC compartment (Graubert et al., 2011). Therefore, characterising the splicing landscape in the *U2AF1*^{S34} HSCs may be of particular interest for identifying biomarkers for targeted therapy.

To this end, we performed parallel single-cell genotyping and RNA-seq in phenotypically defined HSCs derived from MPN patients (Rodriguez-Meira et al., 2019). We focused our analysis on *U2AF1*^{S34} cells in addition to *U2AF1*^{WT} from healthy donors and patients as controls (Figure 6.16).

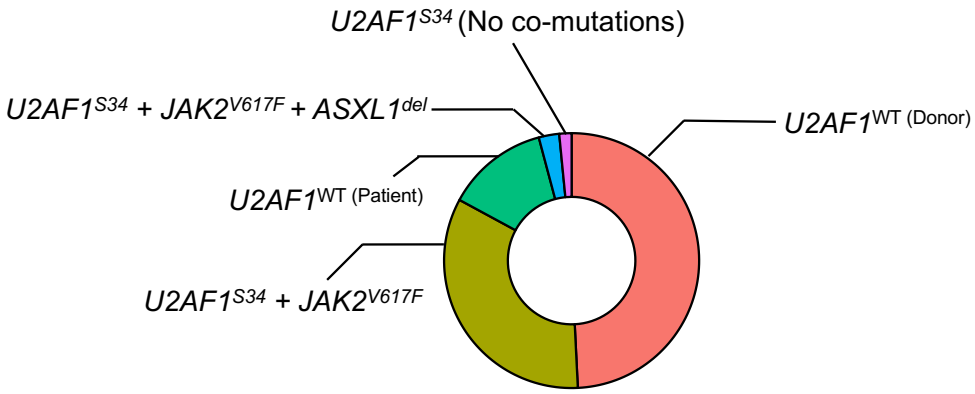


Figure 6.16: Proportion of single-cell genotypes of phenotypically defined HSCs derived from MPN patients. Single-cell dataset generated by Alba Rodriguez-Meira under the supervision of Adam Mead.

We first assessed if we were able to recapitulate previously reported *U2AF1*^{S34}-associated mis-spliced events. We identified 18 previously reported *U2AF1*^{S34}-associated mis-spliced events from the literature for assessment (Table 1.1). Of which, 4 splicing events were expressed in our dataset, namely *DEK*, *GNAS*, *H2AFY*, and *STRAP* (Figure 6.16A). Of these, 3 were differentially spliced by *U2AF1*^{S34} HSCs (Figure 6.16B-E). Both *GNAS* and *H2AFY* had increased percent spliced-in (PSI) in *U2AF1*^{S34} HSCs whereas *STRAP* had decreased PSI in *U2AF1*^{S34} HSCs. The changes in PSI values of these events in *U2AF1*^{S34} HSCs relative to *U2AF1*^{WT} HSCs were consistent with the literature (Wheeler et al., 2022; Yip et al., 2017a).

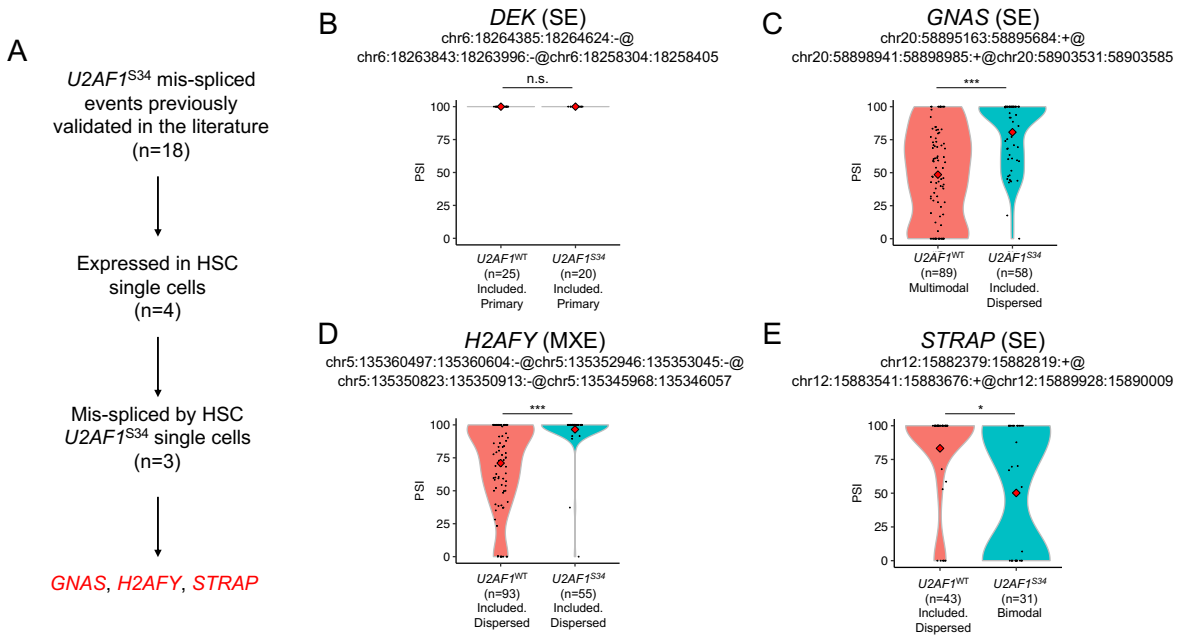


Figure 6.16: Recapitulating previously reported *U2AF1*^{S34}-associated mis-spliced events for assessment in our single-cell dataset. (A) Selection of previously reported *U2AF1*^{S34}-associated mis-spliced events. **(B)** *DEK* was not differentially spliced in *U2AF1*^{S34} HSCs. **(C-E)** *GNAS*, *H2AFY*, and *STRAP* were differentially spliced in *U2AF1*^{S34} HSCs.

After successfully recapitulating previously reported *U2AF1*^{S34}-associated mis-spliced events, we proceeded with transcriptome-wide splicing analysis of *U2AF1*^{S34} vs *U2AF1*^{WT} HSCs. In total, 99 splicing events had significantly increased PSI (ΔPSI > 10 and FDR < 0.1) in *U2AF1*^{S34} relative to *U2AF1*^{WT} HSCs (Figure 6.17A). On the other hand, 87 splicing events had significantly decreased PSI (ΔPSI < -10 and FDR < 0.1) in *U2AF1*^{S34} relative to *U2AF1*^{WT} HSCs. Stratification of the differentially spliced events by splicing event type revealed slight preference for exon skipping and intron retention by *U2AF1*^{S34} HSCs (Figure 6.17B). Pathway enrichment analysis of differentially spliced genes revealed RNA splicing-related gene sets to be enriched (Figure 6.17C). This reaffirms the role of *U2AF1*^{S34} in dysregulated splicing. Other gene sets enriched among differentially spliced genes include mRNA catabolism, translation, and cell cycle.

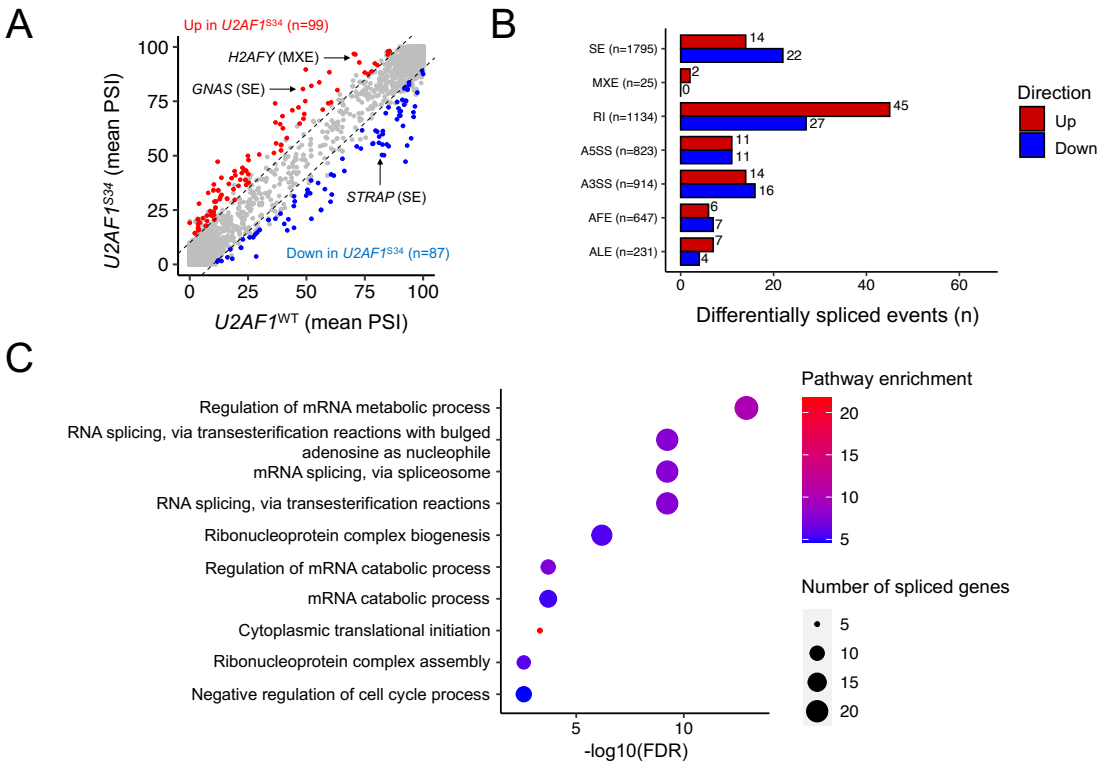


Figure 6.17: Differential splicing analysis between *U2AF1*^{S34} vs *U2AF1*^{WT} HSCs. **(A)** Mean PSI values of splicing events in *U2AF1*^{S34} vs *U2AF1*^{WT} HSCs. Significant splicing events (|ΔPSI| > 10 and FDR < 0.1) are colour coded. **(B)** Significant splicing events stratified by splicing event type. Number in parenthesis indicates total number of expressed events included for differential splicing analysis. **(C)** Top pathways enriched among differentially spliced genes.

Next, we compared the ability of splicing and gene expression profile in distinguishing *U2AF1*^{S34} from *U2AF1*^{WT} HSCs. Dimension reduction analysis using differentially spliced events successfully distinguished *U2AF1*^{S34} from *U2AF1*^{WT} HSCs on the principal component analysis (PCA) space (Figure 6.18A). Similarly, differentially expressed genes and highly variable genes were able to distinguish *U2AF1*^{S34} from *U2AF1*^{WT} HSCs (Figure 6.18B-D).

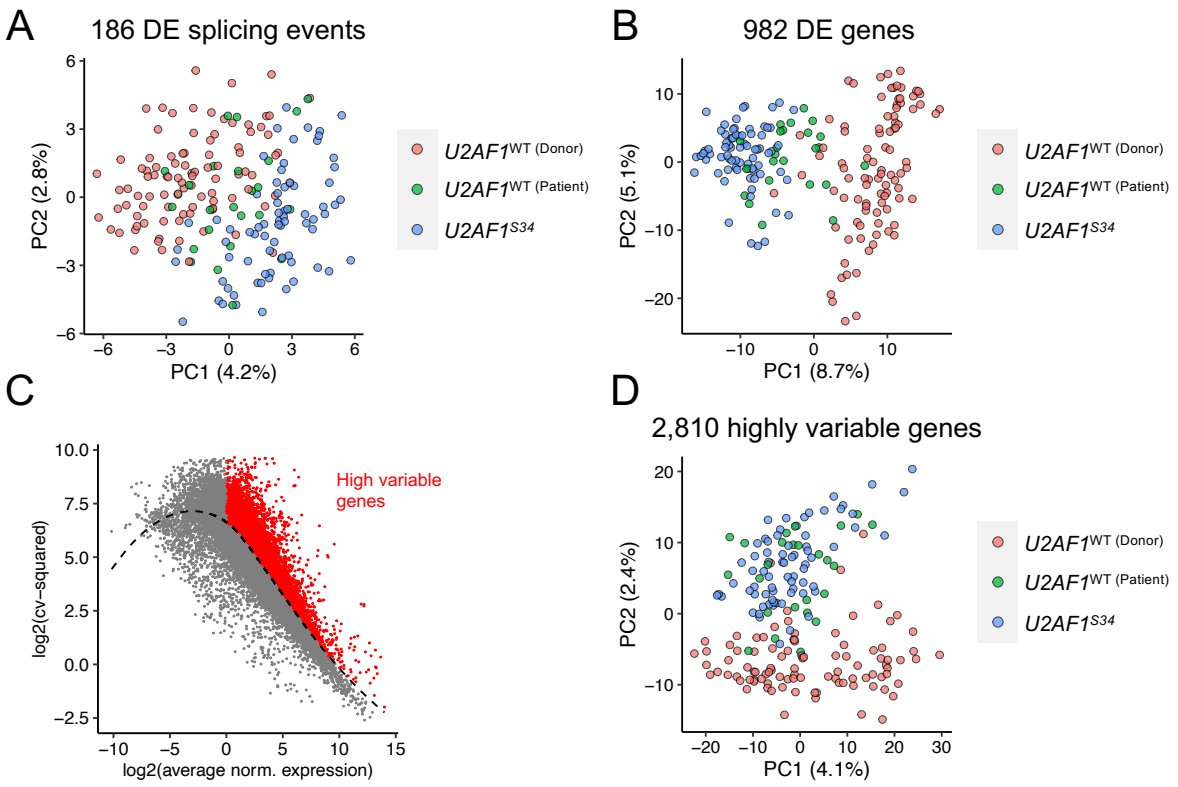


Figure 6.18: Principal component analysis (PCA) using splicing and gene expression values. (A) PCA using differentially spliced events. **(B)** PCA using differentially expressed genes. **(C)** Selection of highly variable genes. **(D)** PCA using highly variable genes.

To assess if the differentially spliced events identified from our analysis were reproducible, we proceeded with identifying these events in the BeatAML cohort (Tyner et al., 2018). Of the 114 non-RI splicing events identified from *U2AF1*^{S34} HSCs, 60 were differentially spliced in *U2AF1*^{S34} patients relative to *U2AF1*$^{WT(Donor)}$ (Figure 6.19). Of these, 45 mis-spliced events had the PSI changes in *U2AF1*^{S34} patients relative to *U2AF1*$^{WT(Donor)}$ or relative to *U2AF1*$^{WT(patient)}$ that were in the same direction as *U2AF1*^{S34} HSCs relative to *U2AF1*^{WT} HSCs.

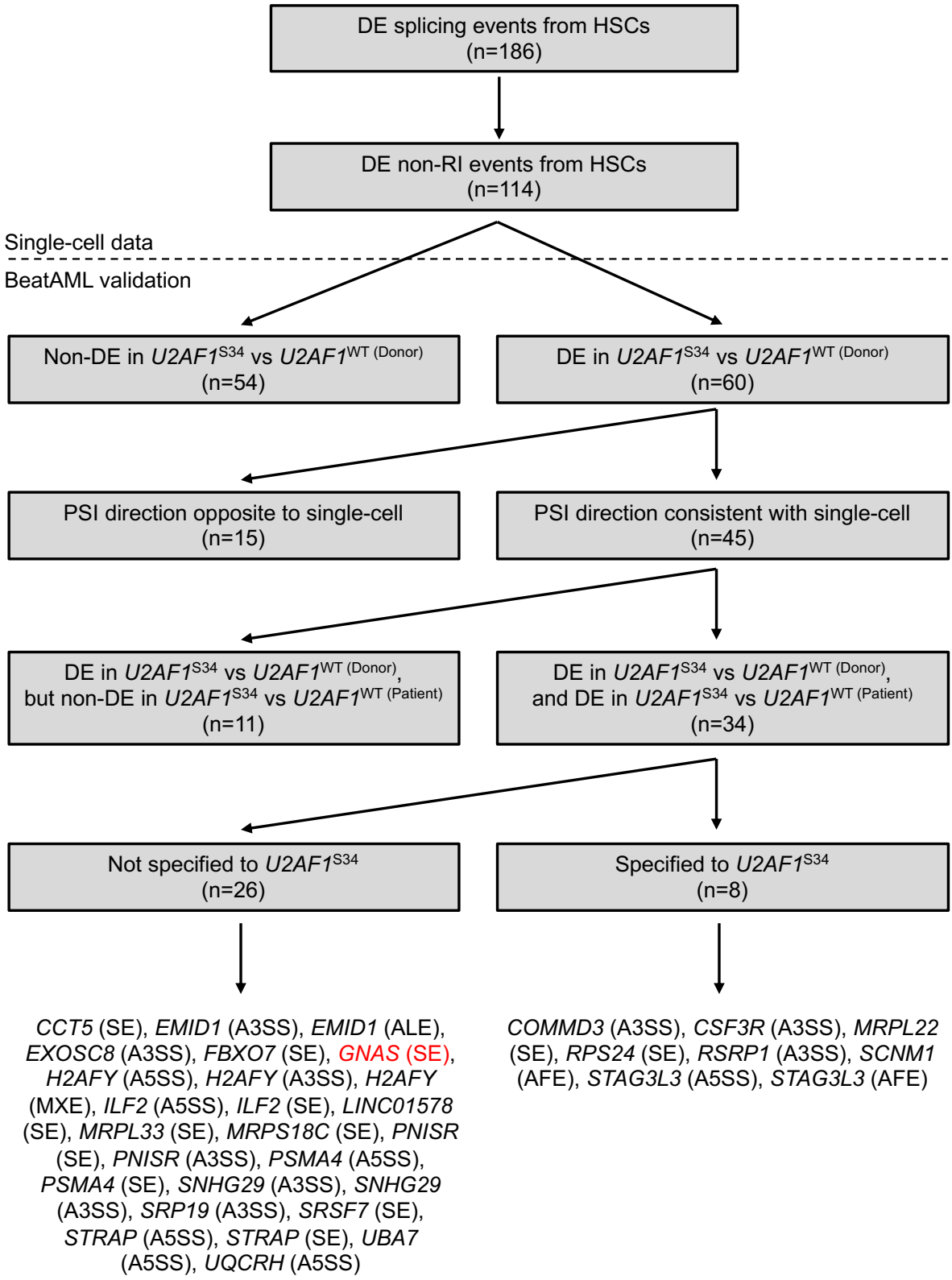


Figure 6.19: Flowchart demonstrating the steps in validating differentially spliced events identified in *U2AF1*^{S34} HSCs using the BeatAML cohort. In total,

34 of 114 *U2AF1*^{S34} HSC-associated mis-spliced events (29%) were successfully recapitulated in the BeatAML cohort.

Of these, 34 splicing events were differentially spliced in *U2AF1*^{S34} patients relative to *U2AF1*$^{WT(Donor)}$ and also to *U2AF1*$^{WT(patient)}$. An example of splicing event that was differentially spliced in *U2AF1*^{S34} relative to *U2AF1*$^{WT(Donor)}$, but not relative to *U2AF1*$^{WT(patient)}$ was *PCBP2* (Figure 6.20A). This suggests that the mis-splicing of this event may not be specific to *U2AF1*^{S34}, but rather, it is associated with myeloid neoplasm in general (with or without spliceosome mutations).

Of the events mis-spliced in *U2AF1*^{S34} relative to *U2AF1*$^{WT(Donor)}$ and also to *U2AF1*$^{WT(patient)}$, 26 were also mis-spliced in AML patients with genetic variants in splicing factors other than *U2AF1*^{S34} while 8 were exclusively mis-spliced in *U2AF1*^{S34}. An example of the former is *EXOSC8* which was mis-spliced in *SF3B1*^{K666}, *SRSF2*^{P95}, *U2AF1*^{S34}, and *U2AF1*^{Q157} patients (Figure 6.20B). An example of the latter is *SCNM1* which was mis-spliced in *U2AF1*^{S34}, but not in AML patients with genetic variants in other splicing factors (Figure 6.20C). *EXOSC8* is an rRNA metabolism-related gene that has been shown to play an oncogenic role in colorectal carcinoma (Cui, Liu, Li, Zhang, & Li, 2020). Increased inclusion of A3SS in this gene in *U2AF1*^{S34} HSCs was not predicted to lead to NMD and therefore may hint at A3SS-mediated activation of this gene. SCNM1 is a putative splicing factor and genetic variants have been reported in this gene in biliary tract cancer (Buchner, Trudeau, & Meisler, 2003).

In total, 34 of 114 (29%) *U2AF1*^{S34} HSC-associated mis-spliced events were successfully validated in BeatAML. This rate was similar to our validation of *SF3B1*^{K666} HSC/MEP-associated mis-spliced events in BeatAML (Figure 6.6).

It is noteworthy that RI was not assessed for validation in BeatAML due to logistic reasons. Only the splice junction information in tab-delimited files were required for validation of SE, A3SS, A5SS, AFE, and ALE splicing events. On the other hand, the large-sized BAM files are required for computing intronic coverage for RI PSI quantification. Finding sufficient storage space to host the BAM files of 360 BeatAML patients currently presents a logistic challenge to yours truly.

Notably, *GNAS*, which has been reported to be mutated in myeloid neoplasm (Cancer Genome Atlas Research et al., 2013), was identified to be mis-spliced in both *U2AF1*^{S34} HSCs and *U2AF1*^{S34} AML patients. *GNAS* encodes the α-subunit of the stimulatory G protein, which is involved in many signalling pathways such as cellular

growth and proliferation (Turan & Bastepe, 2013). *U2AF1*^{S34} HSC-associated inclusion (splicing in) of the alternative exon in *GNAS* was predicted by MARVEL to not lead to nonsense-mediated decay (NMD) of this gene. Therefore, it is conceivable that in lieu of NMD, the inclusion (splicing in) of this alternative exon may instead hyperactivate GNAS. Indeed, *in silico* modelling and experimental characterisation demonstrated that the insertion of this alternative exon in *GNAS* lead to increased G protein activation and adenylyl cyclase activity of the α-subunit of the stimulatory G protein (Wheeler et al., 2022).

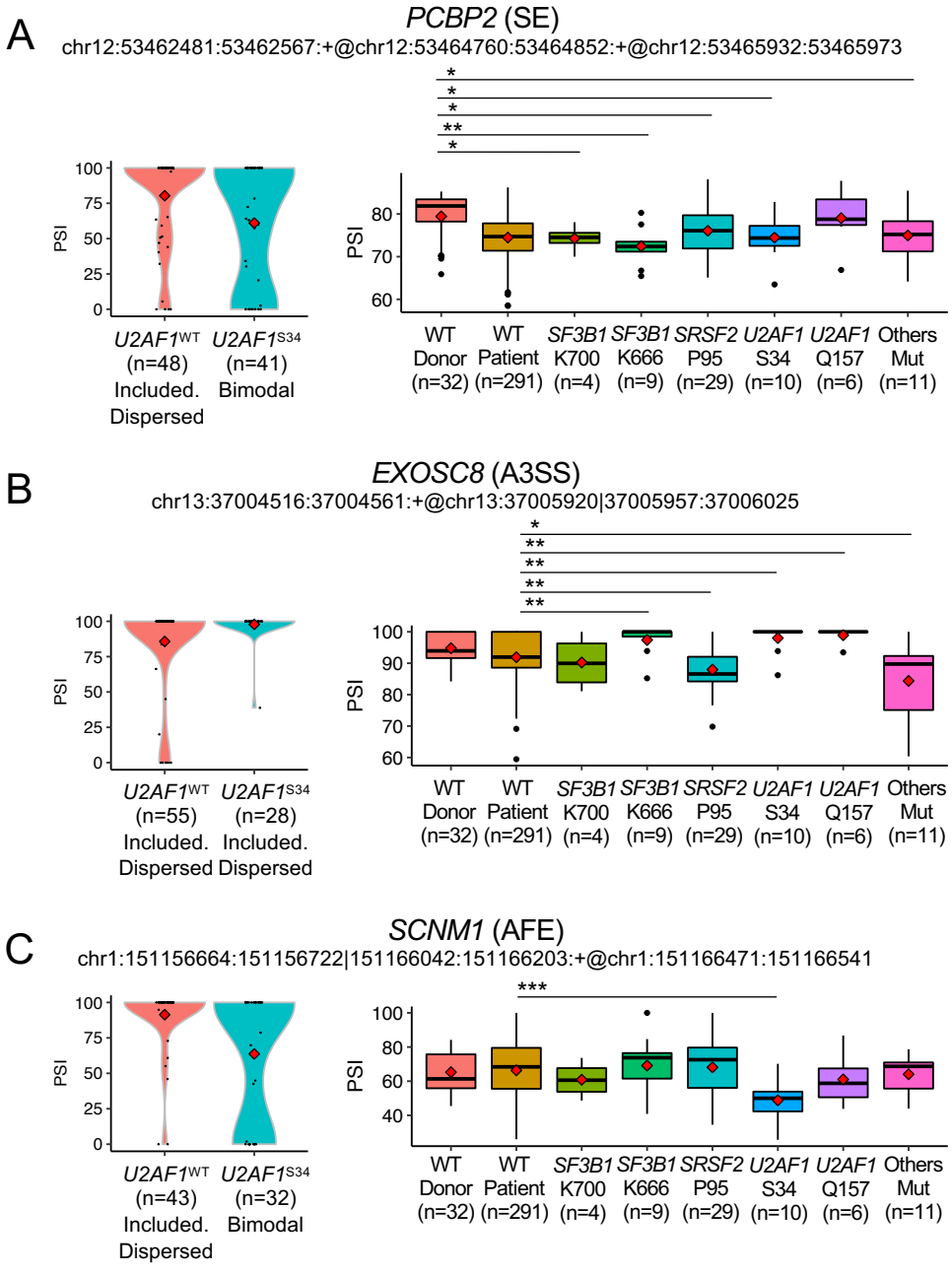


Figure 6.20: Representative examples of *U2AF1*^{S34} HSC-associated mis-spliced events that were validated in BeatAML. (A) Example of mis-spliced event that was associated with AML in general, but not with *U2AF1*^{S34} in particular. **(B)** Example of mis-spliced event that was associated with genetic variants in splicing factors in general, but not with *U2AF1*^{S34} in particular. **(C)** Example of mis-spliced event that was specifically associated with *U2AF1*^{S34}.

To further confirm the mis-splicing of *GNAS* in *U2AF1*^{S34} HSCs, we applied VALERIE to visually inspect the sequencing read alignment at the genomic locus corresponding to this splicing event. Indeed, sequencing read alignment confirmed the alternative exon to be more included (spliced in) in *U2AF1*^{S34} relative to *U2AF1*^{WT} HSCs (Figure 6.21A). Unexpectedly, VALERIE revealed a cryptic A3SS at this genomic locus. Further inspection of the sequencing read alignment corresponding to this cryptic A3SS demonstrated decreased expression of this cryptic A3SS in *U2AF1*^{S34} relative to *U2AF1*^{WT} HSCs (Figure 6.21B). The revelation of this cryptic A3SS suggests that there are at least 4 possible isoforms expressed from this genomic locus of *GNAS* (Figure 6.21C).

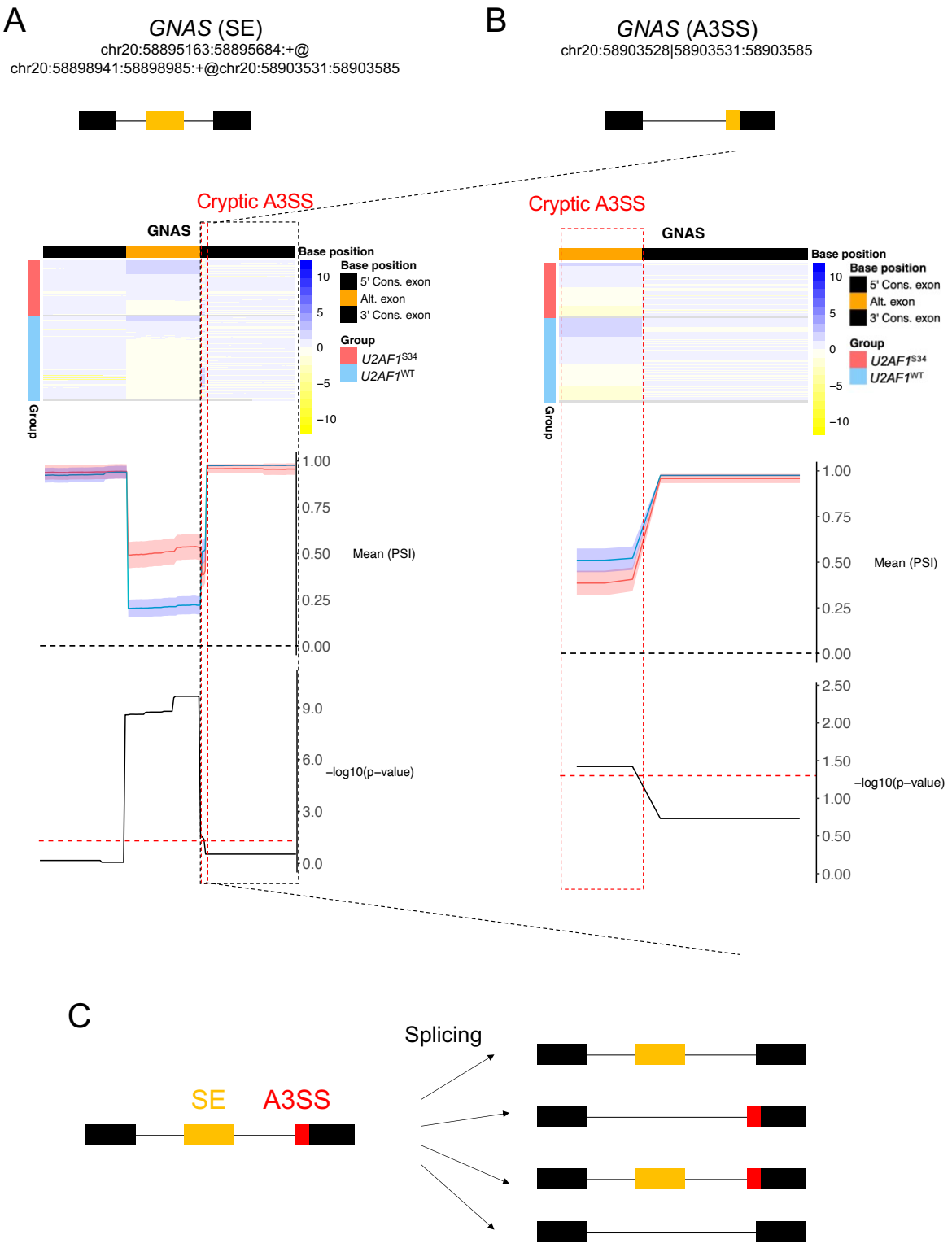


Figure 6.21: Visual-based validation of *GNAS* mis-spliced event using VALERIE. **(A-B)** Sequencing read alignment profile of *GNAS* **(A)** SE event and **(B)** cryptic A3SS event. **(C)** Possible *GNAS* isoforms based on different combination of SE and cryptic A3SS events at this genomic locus.

The deconvolution of complex splicing event, such as *GNAS* SE and cryptic A3SS, may not be straightforward using short-read RNA-seq. Therefore, to delineate the different *GNAS* isoforms, we performed single-cell long-read RNA-seq using Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio). We pooled 10 single cells from *U2AF1*^{WT} and 10 single cells from *U2AF1*^{S34} for sequencing on these two platforms (Figure 6.22).

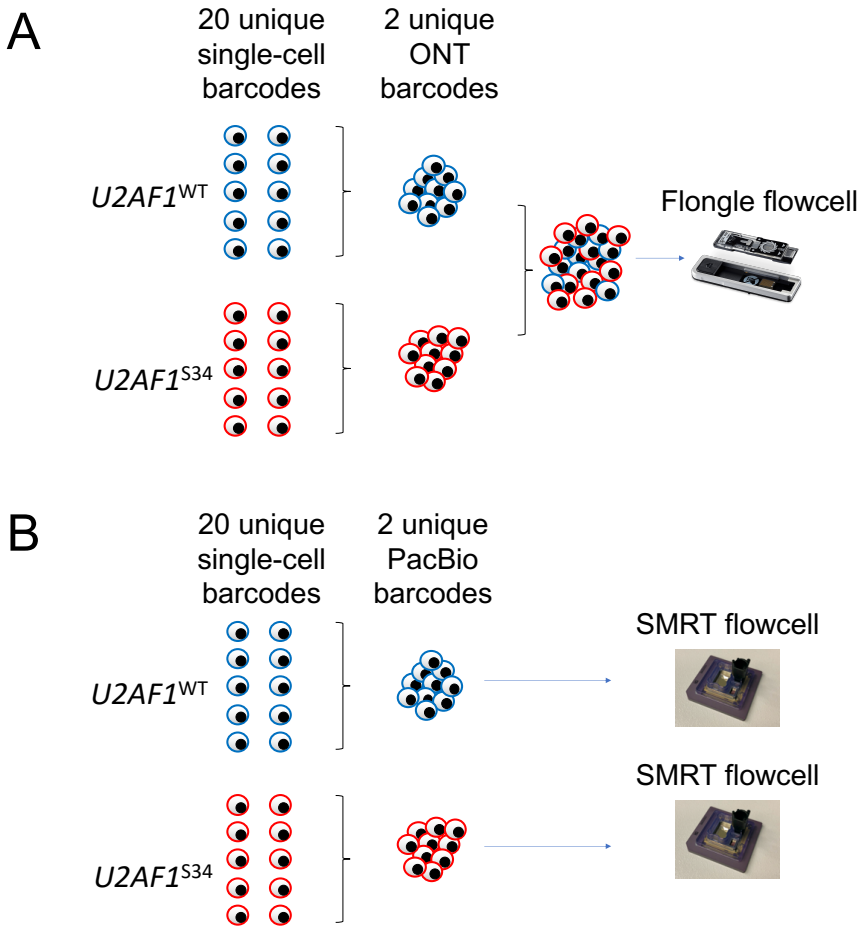
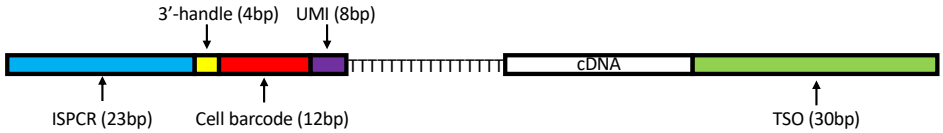


Figure 6.22: Pooling of *U2AF1*^{WT} and *U2AF1*^{S34} HSCs for sequencing on ONT and PacBio. (A-B) Single-cell multiplexing design for **(A)** ONT and **(B)** PacBio**.** ONT data generated by Carika Weldon under the supervision of Rory Bowden. PacBio data generated by Laura Mincarelli under the supervision of Iain Macaulay.

Prior to identifying the different *GNAS* isoforms from our single-cell long-read RNA-seq data, we had to first retrieve and demultiplex the sequencing reads. To this

end, we systematically demultiplexed the sequencing reads based on ONT/Pacbio sample barcode, polyA tail, cell barcode, template switching oligo and finally mappability of reads to the human reference genome (Figure 6.23). Overall, 2,810 (2.6%) and 5,480 (4.6%) of sequencing reads from *U2AF1*^{WT} and *U2AF1*^{S34} pool, respectively, were successfully demultiplexed for ONT. On the other hand, 72,375 (3.8%) and 143,602 (5.4%) of sequencing reads from *U2AF1*^{WT} and *U2AF1*^{S34} pool, respectively, were successfully demultiplexed for PacBio. Due to the higher number of demultiplexed reads obtained from PacBio, we focused our downstream analysis using the demultiplexed reads obtained from PacBio.

A



B

Demultiplexing step	No. of reads: ONT		No. of reads: PacBio	
	U2AF1^{WT}	*U2AF1*^{S34}	*U2AF1*^{WT}	*U2AF1*^{S34}
Sample bacode identified	107,390	119,746	2,203,775	2,661,860
PolyA tail identified (≥9 consecutive A's)	21,434 (20.0%)	28,214 (23.6%)	166,976 (7.6%)	279,790 (10.5%)
Cell barcode identified (Levenshtein distance=2)	15,882 (14.8%)	22,072 (18.4%)	163,695 (7.4%)	276,194 (10.4%)
TSO identified	7,846 (7.3%)	10,078 (8.4%)	103,003 (4.7%)	173,498 (6.5%)
Mapped to genome	2,810 (2.6%)	5,480 (4.6%)	72,375 (3.8%)	143,602 (5.4%)

Figure 6.23: Demultiplexing of sequencing reads. (A) Read construct subjected to long-read RNA-seq on ONT and PacBio. **(B)** Based on the known position of the certain features (e.g., cell barcode) on the read construct in (A), we may systematically demultiplex the sequencing reads as per the direction of the arrow.

Single-cell long-read RNA-seq revealed 5 expressed *GNAS* isoforms, one of which was novel, i.e., not previously reported in publicly available databases. Notably, while the SE was identified, the cryptic A3SS was not expressed or captured here (Figure 6.24). Comparison of the proportion of the different *GNAS* isoforms between *U2AF1*^{S34} and *U2AF1*^{WT} HSCs revealed ENST00000265620.11 transcripts to be increased in *U2AF1*^{S34} HSCs while ENST00000477931.5 and ENST00000371085.7-

1 transcripts to be decreased in *U2AF1*^{S34} HSCs (Figure 6.24). On the other hand, ENST00000371085.7 and novel isoform m64036_191029_065537/98175834/ccs were expressed at similar proportions in both *U2AF1*^{S34} and *U2AF1*^{WT} HSCs.

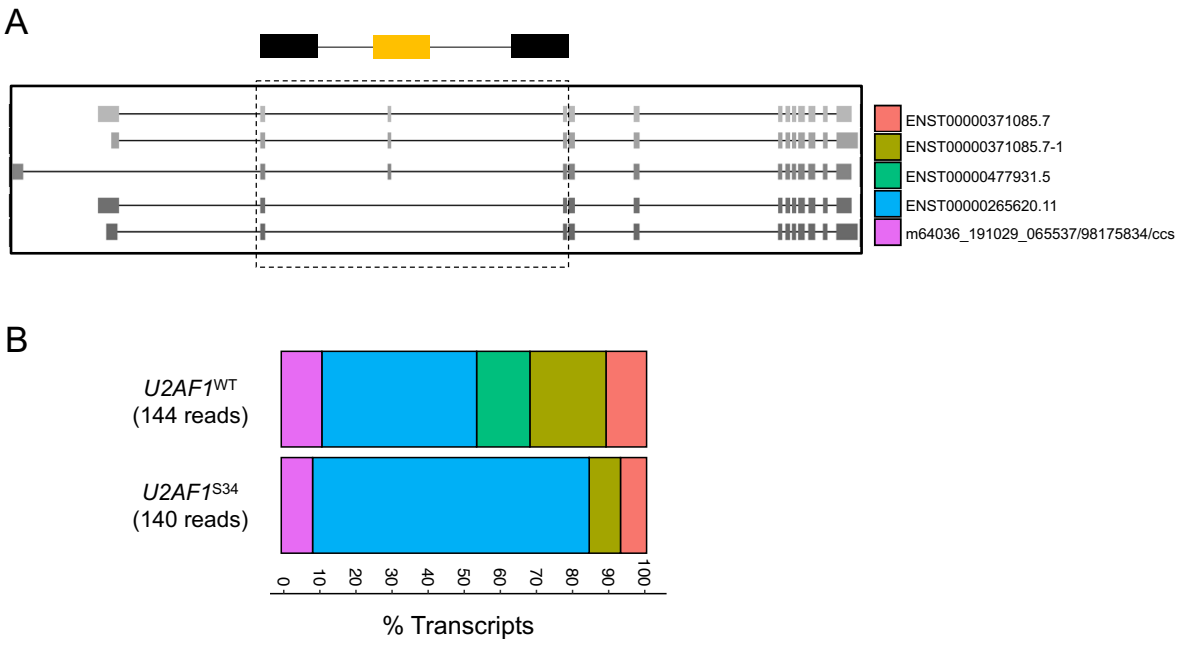


Figure 6.24: Single-cell long-read isoform analysis of *GNAS*. (A) *GNAS* isoforms identified. **(B)** Differences in the proportion of each isoform in *U2AF1*^{WT} and *U2AF1*^{S34} HSCs.

Taken together, we successfully recapitulated previously reported *U2AF1*^{S34}-associated mis-spliced events in our single-cell HSC population. We have also successfully validated novel *U2AF1*^{S34} HSC mis-spliced events in an external myeloid neoplasm cohort. Notably, *GNAS* appeared as a potential candidate gene due to previous reports of *GNAS* genetic variants in myeloid neoplasms (Tate et al., 2019), and GNAS role in G protein activation and signalling (Weinstein, Liu, Sakamoto, Xie, & Chen, 2004). Indeed, *GNAS* was very recently validated as an *U2AF1*^{S34} target gene, and the mis-spliced *GNAS* has been experimentally shown to hyperactivate the ERK/MAPK signalling pathway (Wheeler et al., 2022). Single-cell visual-based inspection of the sequencing read alignment at the *GNAS* genomic locus revealed complex alternative splicing consisting SE and cryptic A3SS events. Lastly, we attempted to deconvolute the different *GNAS* isoforms using single-cell long-read RNA-seq.

6.4 Bulk analysis of MBNL1-deficient DM1 patients

We demonstrated the application of our pipeline on single-cell splicing analysis of myeloid neoplasm patients. In principle, our pipeline is also applicable to bulk splicing analysis. Here, we demonstrate the application our pipeline on bulk RNA-seq data generated from myotonic dystrophy type 1 (DM1) patients with deficiency in splicing factor MBNL1.

MBNL1 (Muscleblind like splicing regulator 1) is a splicing factor that binds competitively with U2AF65 at the acceptor splice site. The binding of MBNL1 at the acceptor splice site, in lieu of U2AF65, leads to the exclusion (splicing out) of the adjacent exon from the final mRNA transcript (Figure 6.24) (E. T. Wang et al., 2012; Warf et al., 2009). Binding of MBNL1 along the exon body similarly leads to exclusion (splicing out) of the exon whereas binding of MBNL1 at the donor splice site leads to inclusion (splicing in) of the adjacent exon in the final mRNA transcript (Cheng et al., 2014; E. T. Wang et al., 2012).

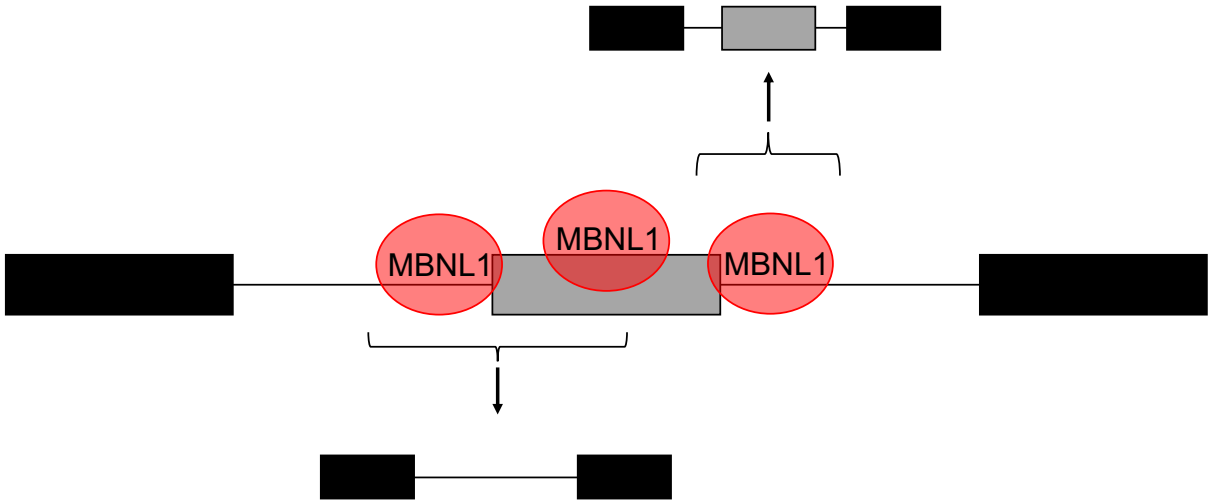


Figure 6.24: Binding of MBNL1 at different positions relative to the alternative exon. Binding at acceptor splice site and exon body leads to inclusion (splicing in) of the alternative exon whereas binding at the donor splice site leads to exclusion (splicing out) of the alternative exon. Grey colour represents the alternative exon and black colour represents the constitutive exons.

DM1 is an inherited disease characterised by progressive muscle weakness and wasting, cataract development, testicular atrophy, and cardiac conduction defects (Cho & Tapscott, 2007). Affected individuals have extended CTG repeats at the 3' untranslated region (3' UTR) of *DMPK* gene. MBNL1 proteins are sequestered within this CTG repeat region, leading to decreased MBNL1 levels, and consequently dysregulated splicing of MBNL1 target genes.

Several therapeutic approaches have been developed to reverse DM1 dysregulated splicing profile. One approach is to use antisense oligonucleotides (ASOs) with sequence complementary to the CTG repeat region of *DMPK* (N. Hu et al., 2021). The competitive binding of ASOs to this region precludes MBNL1 from binding to this region. Another approach is to use RNA-targeting Cas9 (RCas9) to target and eliminate the CTG repeat region of *DMPK* (Batra et al., 2017). Here, we introduced a novel approach whereby we utilised dCas13 to bind the CTG repeat region of *DMPK*, consequently precluding MBNL1 from being sequestered to this region.

To assess the effectiveness of our approach, we performed bulk RNA-seq on *DMPK* wildtype cells (WT), DM1 cells, WT cells treated with non-targeting (NT) ASO, DM1 cells treated with NT ASO, DM1 cells treated with low-dose ASO, DM1 cells treated with high-dose ASO, WT cells treated with NT dCas13, DM1 cells treated with NT dCas13, and DM1 cells treated with dCas13. Each sample group had three biological replicates. Five individual samples had two technical replicates while the remaining individual samples had only one technical replicate. Coverage assessment of the *DMPK* gene loci revealed increased coverage at the 3' UTR of DM1 samples, thus confirming the extended CTG repeats in DM1 patients included in this study (Figure 6.25).

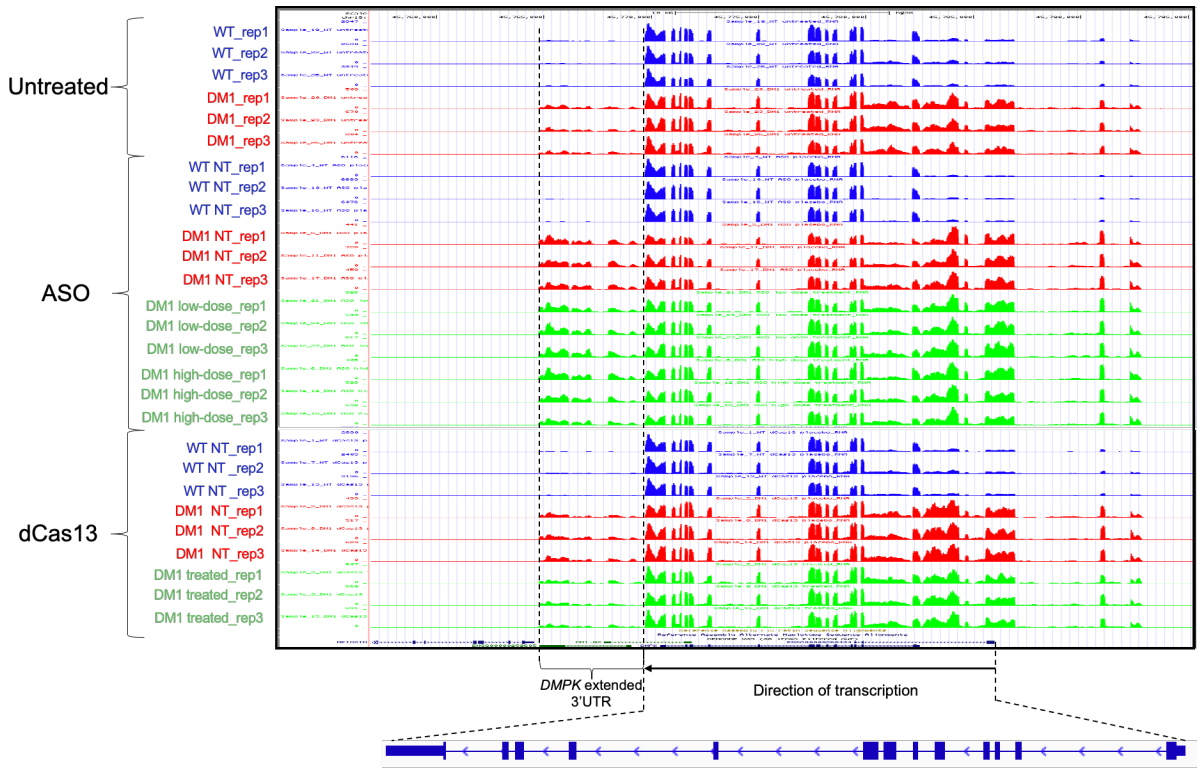


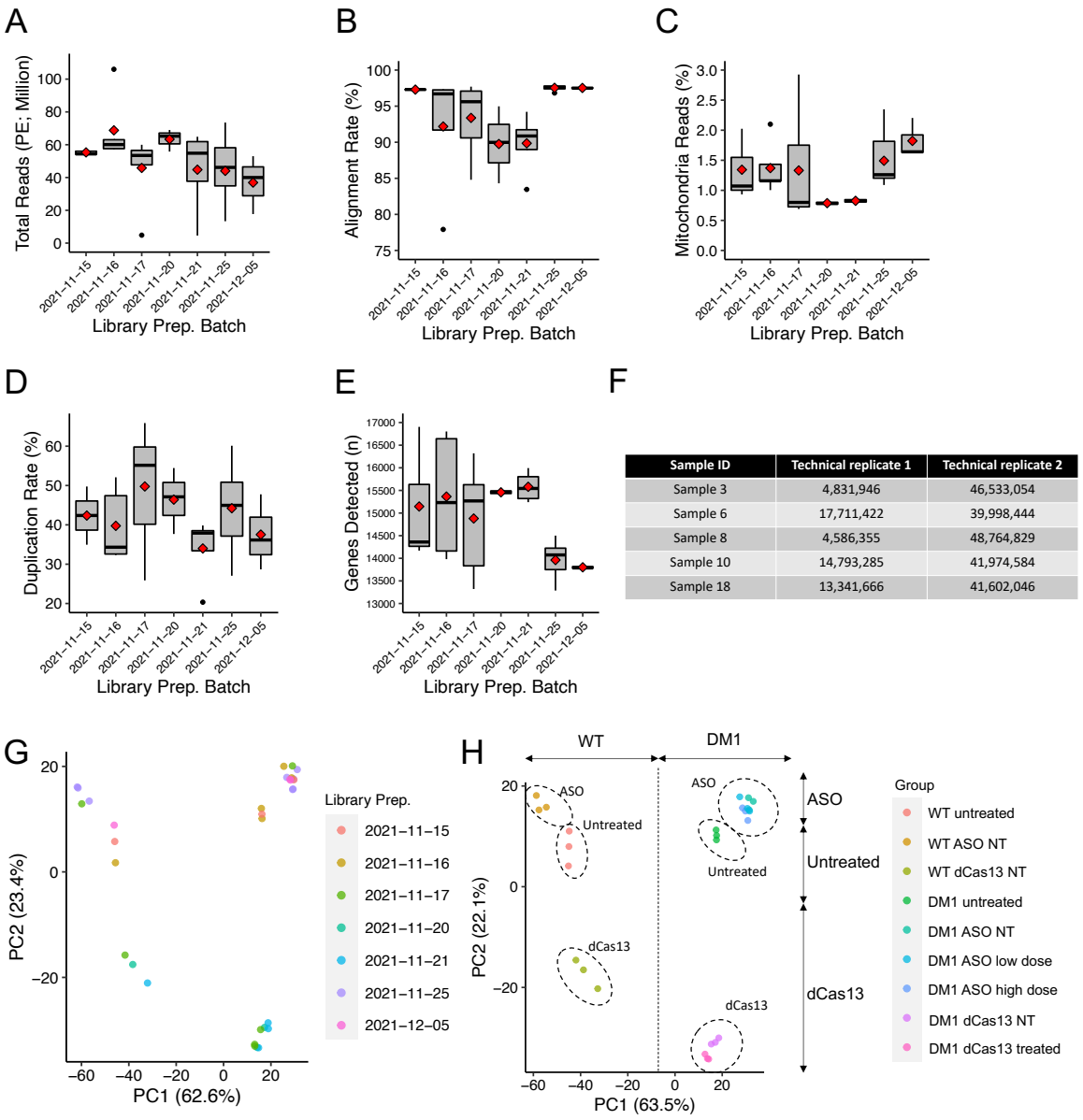
Figure 6.25: Coverage assessment of *DMPK* gene loci on the UCSC Genome Browser. Coverage tracks ordered by treatment type and colour coded based both disease status and treatment group. Blue represents *DMPK* WT samples, red represents DM1 samples treated with non-targeting ASO/dCas13, and green represents DM1 samples treated with ASO/dCas13. Coverage tracks revealed increased coverage at the 3' UTR of *DMPK* in DM1 samples compared to *DMPK* WT samples. Note that technical replicates were merged in this figure. Bulk RNA-seq dataset generated by Muhammad Hanifi under the supervision of Tatjana Sauka-Spengler.

Sequencing quality control (QC) assessment using total number of reads (sequencing depth), alignment rate, mitochondrial read contribution, duplication rate, and number of detected genes did not reveal any poor-quality samples (outliers) for removal prior to downstream analysis (Figure 6.25A-E). The five samples with technical replicates had higher number of sequencing reads for replicate no. 2 compared to replicate no. 1, indicating that re-sequencing of these samples successfully increased sequencing depth (Figure 6.25F).

As the samples were sequenced across multiple batches, we next check for the presence of batch effect contributed by the different library preparation dates. We first identified 18,360 expressed genes (defined as genes with counts per million (CPM) of at least 1 in at least three samples), of which, we selected the top 10% most variable genes based on variance across the samples. These 1,836 high variable genes were used for clustering the samples on the principal component analysis (PCA) space and the samples were annotated by their library preparation dates (Figure 6.25G). We observed no obvious samples clustering by library preparation dates. Moreover, the technical replicates of each sample clustered in close proximity. Therefore, we did not perform batch correction and proceeded with merging the technical replicates.

Unsupervised clustering using these highly variable genes revealed samples clustered by *DMPK* WT and DM1 status on the 1st principal component (PC1) (Figure 6.25H). This suggests that disease status is the strongest distinguishing factor in this study. Nevertheless, it is noteworthy that *DMPK* WT samples were derived from immortalized human myoblast cell lines (MRC CNMD Biobank London) whereas DM1 samples were derived from human patients (Arandel et al., 2017). Therefore, segregation of samples on the 1st PC may be partly due to differences in sample origin.

Furthermore, samples clustered by treatment group (untreated, ASO, and dCas13) on PC2. For example, samples within the WT group clustered by untreated and NT ASO and NT dCas13. This suggests that transfection alone, even in the absence of therapeutic ASO and dCas13, was sufficient to lead to changes in transcriptome profile. This needs to be considered when identifying samples groups for differential gene expression and splicing analysis. For example, to control for the effect of ASO/dCas13 transfection on the samples' transcriptome, we should compare samples within the same treatment group, such as DM1 dCas13 treated vs DM1 NT dCas13 (in lieu of vs DM1 untreated).



Sample ID	Technical replicate 1	Technical replicate 2
Sample 3	4,831,946	46,533,054
Sample 6	17,711,422	39,998,444
Sample 8	4,586,355	48,764,829
Sample 10	14,793,285	41,974,584
Sample 18	13,341,666	41,602,046

Figure 6.25: Sequencing QC. (A-E) Sequencing QC based on **(A)** sequencing depth, **(B)** alignment rate, **(C)** mitochondrial reads contribution, **(D)** duplication rate, and **(E)** number of genes detected. **(F)** Comparison of sequencing depth between technical replicate no. 1 and 2 for five samples with technical replicates. **(G-H)** Unsupervised clustering using highly variable genes, samples annotated by **(G)** library preparation batch and **(H)** disease status and treatment group.

Next, we assessed whether our RNA-seq data could recapitulate reversal of DM1-associated mis-spliced events (*MBNL1*, *MBNL2*, *DMD*, *SOS1*, *INSR*) by ASO and dCas13 treatment, as previously validated by polymerase chain reaction (PCR; data not shown). We also included DM1-associated mis-spliced events previously

reported in the literature (*ANK2*, *ATP2A1*, and *CACNA1*) for this analysis (Nakamori et al., 2017).

We observed *MBNL1* exon 7, *MBNL2* exon 5, and *DMD* exon 78 in ASO- and dCas13-treated DM1 samples to have splicing profiles similar to that of WT samples (Figures 6.26A-C). For *MBNL1* exon 7 and *MBNL2* exon 5, the percent spliced-in (PSI) values were increased in DM1 samples compared to WT samples (Figures 6.26A and B). ASO treatment successfully reduced the PSI values in a dose-dependent manner whereby high-dose ASO treatment led to more reduction in the PSI values compared to low-dose ASO treatment. dCas13 treatment similarly reduced the PSI values to levels comparable to that of high-dose ASO treatment. This indicates that high-dose ASO and dCa13 treatment led to complete reversal of aberrant splicing profile of these alternative exons.

On the other hand, for *DMD* exon 78, the PSI values were decreased in DM1 samples compared to WT samples (Figure 6.26C). High-dose ASO treatment, but not low-dose ASO treatment, led to slight increase in PSI values. dCas13 treatment similarly led to slight increase in PSI values. Nevertheless, both ASO and dCas13 treatment did not increase PSI values to levels comparable to that of WT samples. This indicates that ASO and dCas13 treatment only led to partial reversal of aberrant splicing profile of this alternative exon.

The effectiveness of ASO treatment for *SOS1* exon 21 was harder to discern. Specifically, PSI values of WT and DM1 NT ASO samples were similar (Figure 6.26D) Nevertheless, dCas13 treatment successfully increased PSI values to levels comparable to that of WT samples. For *INSR* exon 11 and *ANK2* exon 21, we did not observe any differences in splicing profile between WT and DM1 samples (Figures 6.26E and F). Our observation on *INSR* exon 11 here was consistent with our PCR data that demonstrated only minimal difference in *INSR* exon 11 splicing rates between DM1 and WT samples (data not shown). Lastly, there were insufficient coverage at *ATP2A1* exon 22 and *CACNA1S* exon 29 for us to assess their splicing profiles across the different sample groups (Figures 6.26G-H).

It is noteworthy that *ANK2* and *CACNA1S* are late-stage biomarkers in DM1 (Wojciechowska et al., 2018). Here, muscle cells were differentiated for 6 days prior to RNA-seq. Taken together, the overall splicing profile observed here seems to recapitulate early stage of the disease with mis-splicing of early-to-medium responder

exons (*MBNL1*, *MBNL2,* and *SOS1*), but not the medium-to-late responder exons (*CACNA1S* and *ANK2*).

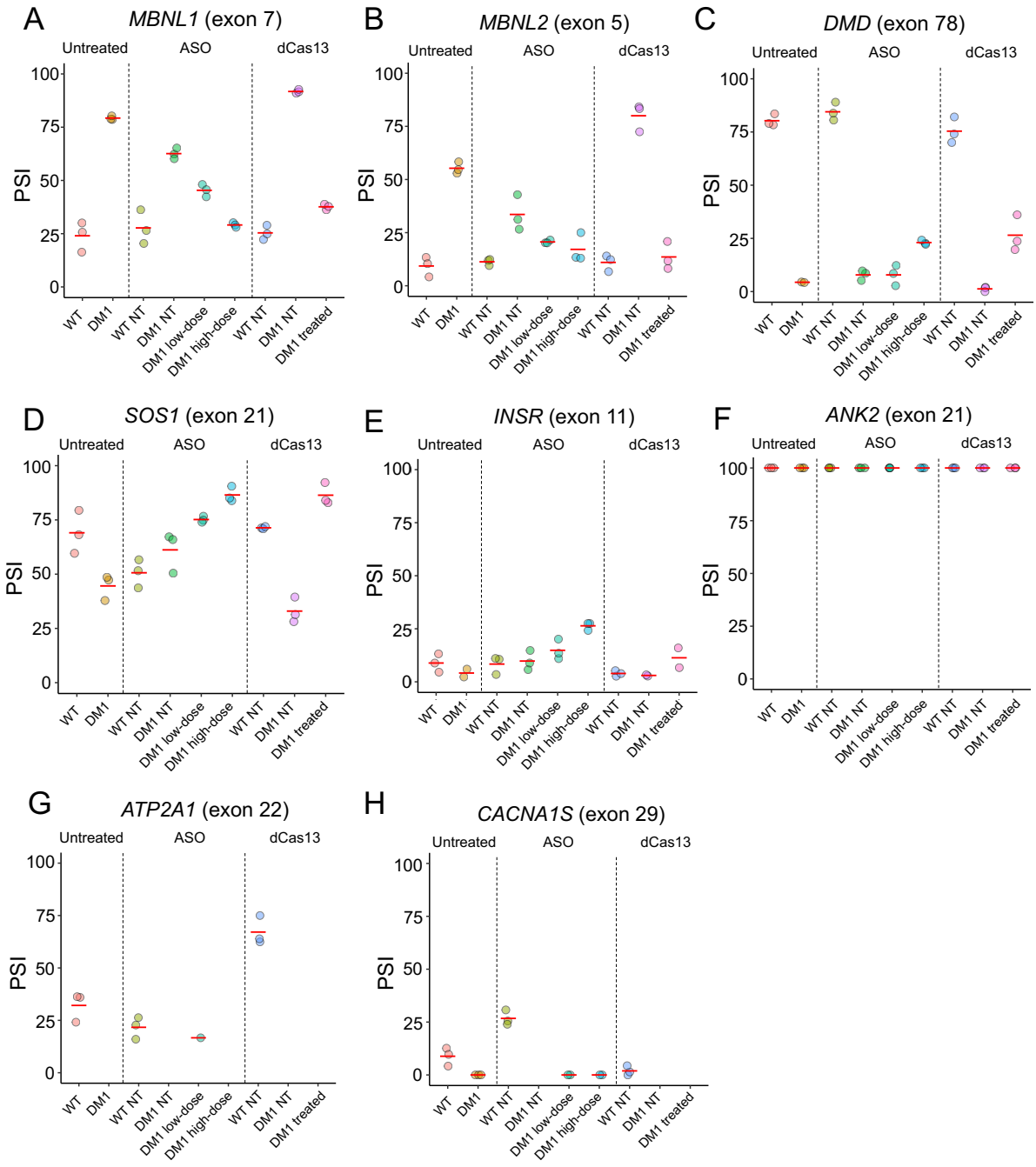


Figure 6.26: PSI values across the different sample groups. (A-B) *MBNL1* exon 7 and *MBNL2* exon 5 demonstrated complete reversal of splicing profile in ASO- and dCas13-treated DM1 samples. **(C)** *DMD* exon 78 demonstrated partial reversal of splicing profile in ASO- and dCas13-treated DM1 samples. **(D)** *SOS1* exon 21 demonstrated reversal of splicing profile in dCas13-treated samples. **(E-F)** *INSR* exon

11 and *ANK2* exon 21 did not demonstrate any differences in splicing profile between DM1 and WT samples. **(G-H)** Too few samples with sufficient coverage at *ATP2A1* and *CACNA1S* alternative exons for assessment.

Closer inspection of the gene expression levels of these genes suggests that alternative splicing events would require approximately a minimum gene expression of log2(CPM + 1) of 4 to have sufficient coverage for splicing analysis (Figures 6.27A-H).

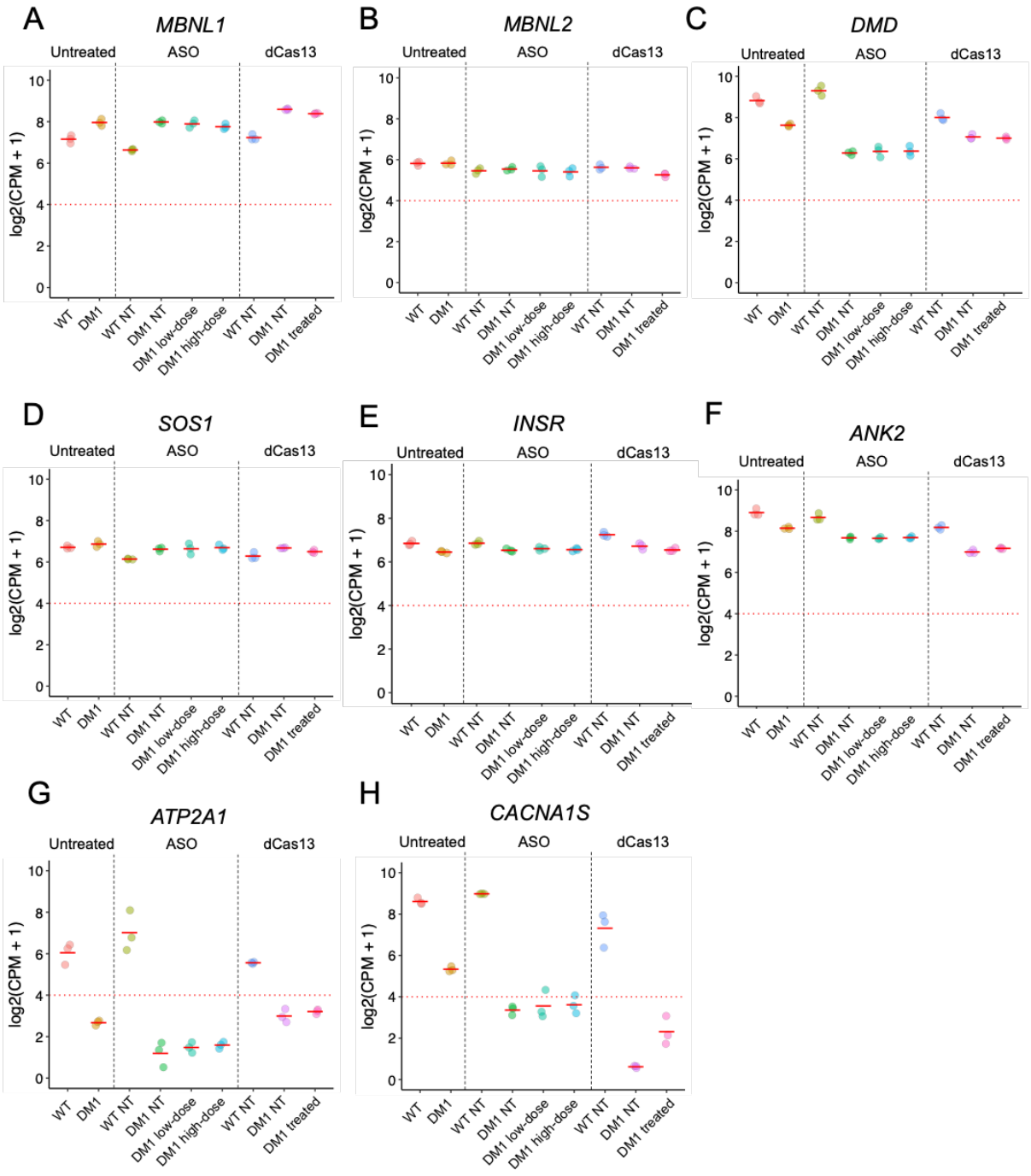


Figure 6.27: Normalised gene expression levels of genes included for splicing analysis in Figure 6.26. (A-F) Genes with moderate-to-high expression levels, and consequently all samples had sufficient coverage at alternative exons for splicing analysis (related to Figure 6.26A-F). **(G-H)** Genes with low expression levels, and consequently too few samples had sufficient coverage at alternative exons for splicing analysis (related to Figure 6.26G-H). CPM: Counts per million.

Now that we have demonstrated the ability of our RNA-seq data to recapitulate previously validated DM1-associated mis-spliced events, we proceeded to perform global differential splicing analysis between DM1 and WT samples to identify transcriptome-wide DM1-associated mis-spliced events. In total, 33,156 splicing events were expressed in both DM1 NT dCas13 and WT NT dCas13 samples and therefore were available for differential splicing analysis (Figure 6.28A). Of which 1,405 (4.2%) splicing events were differentially spliced. SE event type constituted majority of differential splicing events. This is consistent with the reported role of MBNL1 in regulating splicing via its binding to SEs (E. T. Wang et al., 2012). Similarly, 32,456 splicing events were expressed in both DM1 NT ASO and WT NT ASO samples and therefore were available for differential splicing analysis (Figure 6.28B). Of which, 1,948 (6.0%) splicing events were differentially spliced and SE event type constituted majority of differentially spliced events.

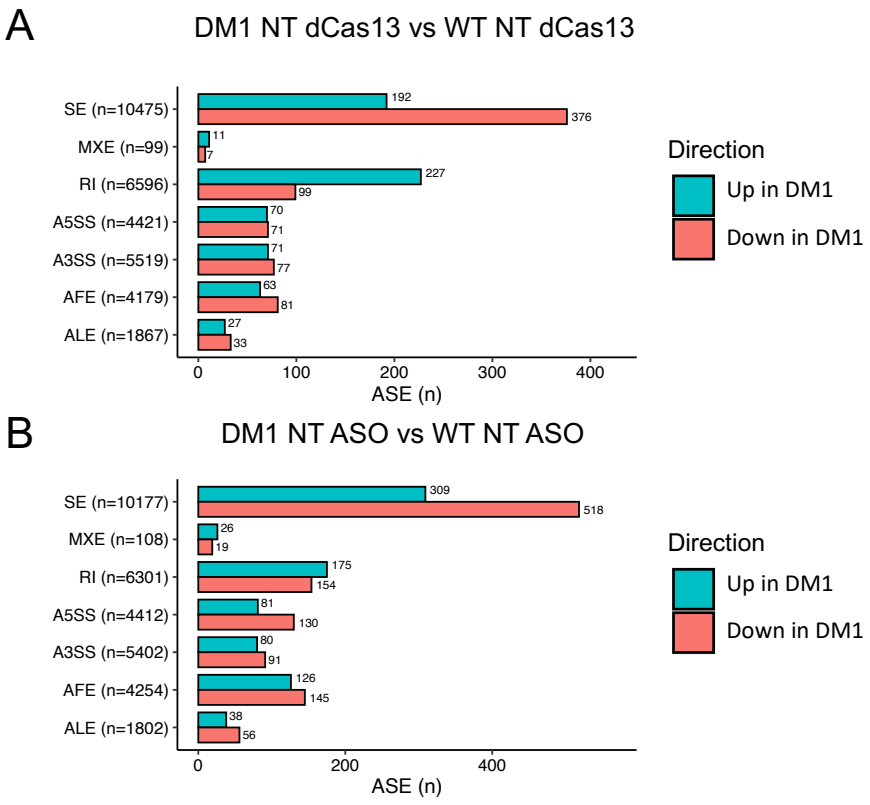


Figure 6.28: Number of differentially spliced events stratified by splicing event type. Number of differentially spliced events in **(A)** DM1 NT dCas13 vs WT NT dCas13 and **(B)** DM1 NT ASO vs WT NT ASO. SE event type is the predominant event type

that is differentially spliced. Number in parentheses indicate number of expressed splicing events included for differential splicing analysis.

Next, we performed gene ontology analysis to identify biological pathways that were aberrantly spliced in DM1 patients. We identified the greatest number of biological pathways enriched among differentially spliced genes of SE event type, and several biological pathways enriched among differentially spliced genes of MXE event type, but no biological pathways enriched among differentially spliced genes of other event types (Figure 6.29A). Top biological pathways enriched among differentially spliced genes of SE event type revealed pathways previously reported to be regulated by MBNL1, namely pathways related to muscle function and pathways related to localisation of mRNAs to specific cellular locations such as the Golgi apparatus and cell membrane (Figures 6.29B and C) (E. T. Wang et al., 2012).

Taken together, this suggests that DM1 main phenotypes, i.e., muscle weakness and wasting, may be potentially attributed to aberrant splicing of muscle-related genes, specifically that of SE event type. Therefore, we proceeded to focus our assessment of ASO and dCas13 treatment efficiency on SE mis-spliced events.

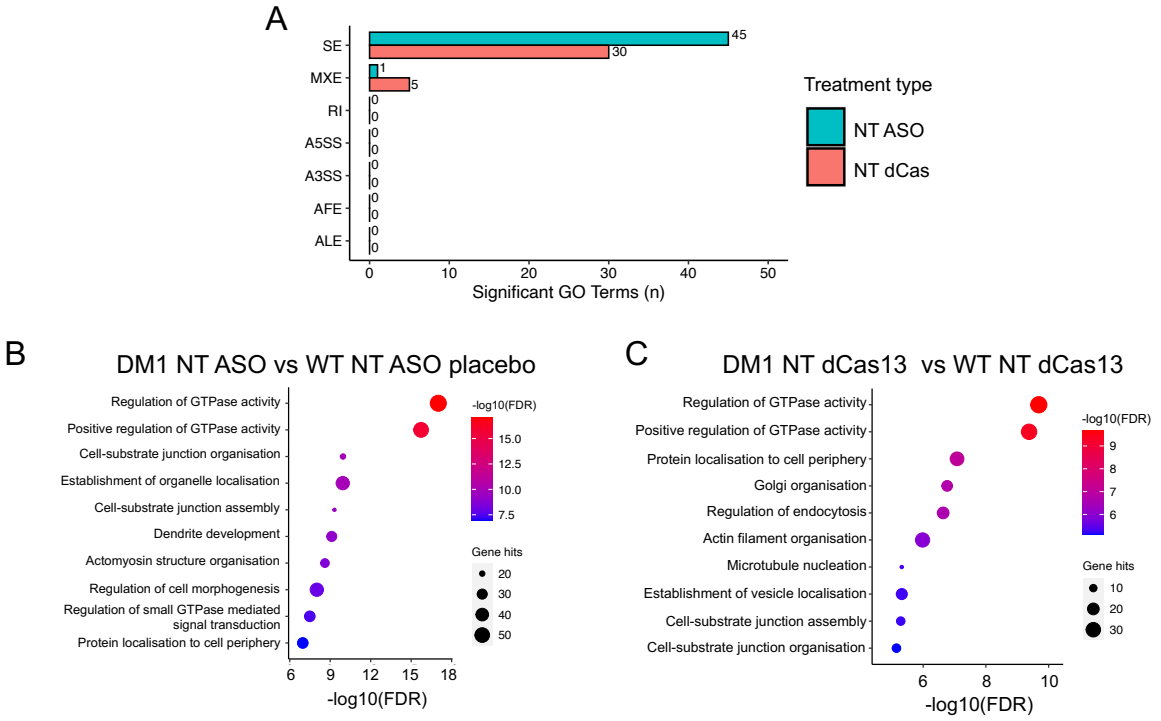


Figure 6.29: Pathway enrichment analysis of differentially spliced genes between DM1 and WT samples. (A) Number of biological pathways enriched among differentially spliced genes stratified by splicing event type. **(B-C)** Top biological pathways enriched among differentially spliced genes of SE event type from **(B)** DM1 NT dCas13 vs WT NT dCas13 and **(C)** DM1 NT ASO vs WT NT ASO comparisons.

In total, 568 differential splicing SE events were identified in DM1 NT dCas13 vs WT NT dCas13 comparison (Figure 6.30A), of which 518 (91%) were expressed in DM1 dCas13 treated vs WT NT dCas13 comparison and were therefore subsequently included in our assessment of dCas13 treatment efficiency. Hierarchical clustering using these 518 splicing events revealed DM1 dCas13 treated samples to cluster in between WT NT dCas13 samples and DM1 NT dCas13 samples (Figure 6.30A). Therefore, DM1 dCas13 treated samples had an intermediate splicing profile between WT NT dCas13 and DM1 NT dCas13 samples. This suggests dCas13 treatment reversed, to some extent, DM1-associated mis-spliced events.

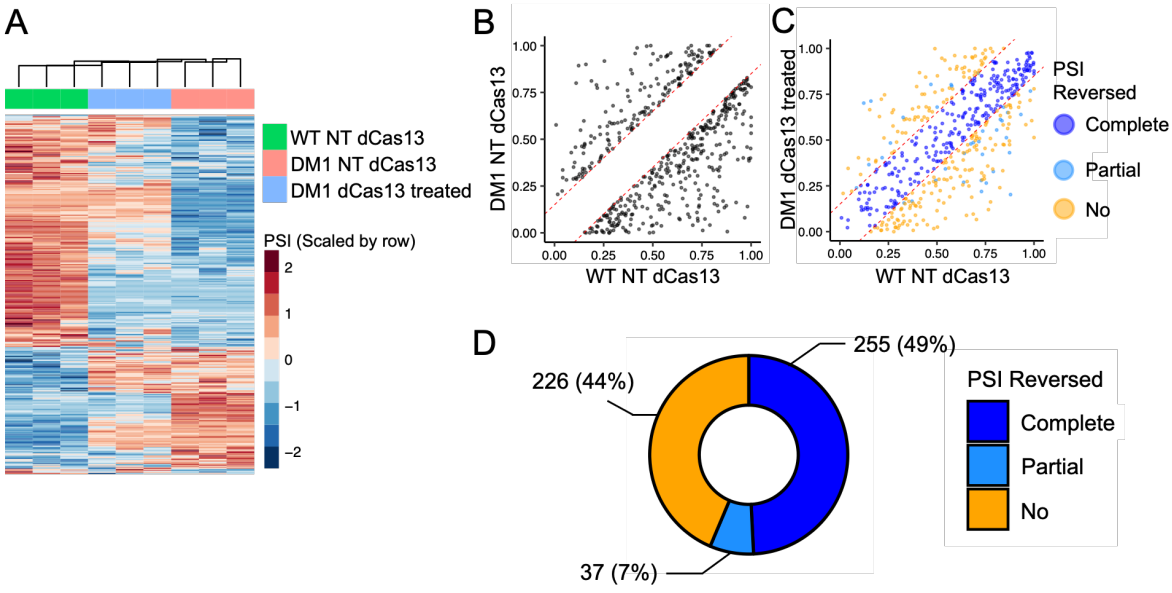


Figure 6.30: Assessing dCas13 treatment efficiency in reversing DM1-associated SE mis-spliced events identified from DM1 NT dCas13 vs WT NT dCas13 comparison and expressed in DM1 dCas13 treated vs WT NT dCas13 comparison. (A) Hierarchical clustering using DM1-associated mis-spliced events. **(B)** DM1-associated mis-spliced events identified from DM1 NT dCas13 vs WT NT

dCas13 comparison. **(C)** The same splicing events in (B) demonstrated for DM1 dCas13 treated vs WT NT dCas13 comparison. **(D)** Number of DM1-associated mis-spliced events stratified into complete, partial, or no reversal when treated with dCas13.

To quantify the extent of treatment-related reversal of DM1-associated mis-spliced events, we stratified each of the 518 DM1-associated mis-spliced event into complete, partial or no reversal. Complete reversal is defined as |ΔPSI| > 15 between DM1 NT ASO/dCas13 vs WT NT ASO/dCas13 comparison but |ΔPSI| < 15 between DM1 ASO/dCas13 treated vs WT NT ASO/dCas13 comparison. For example, *MBNL1* exon 7 demonstrated complete reversal when treated with dCas13 (Figures 6.31A). Partial reversal is defined as |ΔPSI| > 15 between DM1 NT ASO/dCas13 vs WT NT ASO/dCas13 comparison but |ΔPSI| > 15 between DM1 ASO/dCas13 treated vs DM1 NT ASO/dCas13 comparison. For example, *DMD* exon 78 demonstrated partial reversal when treated with dCas13 (Figure 6.31B). No reversal is defined as |ΔPSI| > 15 between DM1 NT ASO/dCas13 vs WT NT ASO/dCas13 comparison but |ΔPSI| > 15 between DM1 ASO/dCas13 treated vs WT NT ASO/dCas13 comparison and |ΔPSI| < 15 between DM1 ASO/dCas13 treated vs DM1 NT ASO/dCas13 comparison. For example, *DMD* exon 78 demonstrated no reversal when treated with low-dose ASO (Figure 6.31C). We selected a threshold of |ΔPSI| > 15 to enable comparison of our method to that reported by a previous study (Batra et al., 2017).

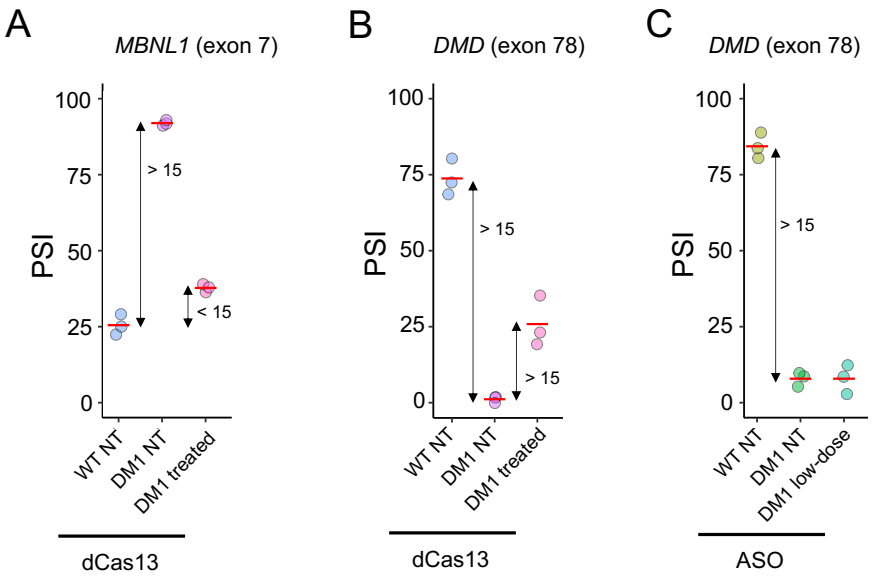


Figure 6.31: Graphical illustration of the definition of ASO/dCas13 treatment-related reversal rate of DM1-associated mis-spliced events. (A-C) Definition for **(A)** complete, **(B)** partial, and **(C)** no reversal of DM1-associated mis-spliced events by ASO/dCas13 treatment.

dCas13 treatment successfully reversed 255 (49%) DM1-associated mis-spliced events, while 37 (7%) mis-spliced events were partially reversed, and 226 (44%) mis-spliced events were not reversed (Figure 6.30B-D). The relatively smaller number of partially reversed compared to completely reversed DM1-associated mis-spliced events suggests that the treatment-related reversal of mis-spliced events is an "all-or-nothing" phenomenon.

Assessment of reversal of DM1-associated mis-spliced events by dCas13 treatment in splicing event types other than SE demonstrated similar reversal rates of ~50% for MXE, RI, A5SS, A3SS, AFE, and ALE (Figure 6.32A- F). This suggests that dCas13 treatment reversed mis-splicing for all splicing event types to similar extents in general.

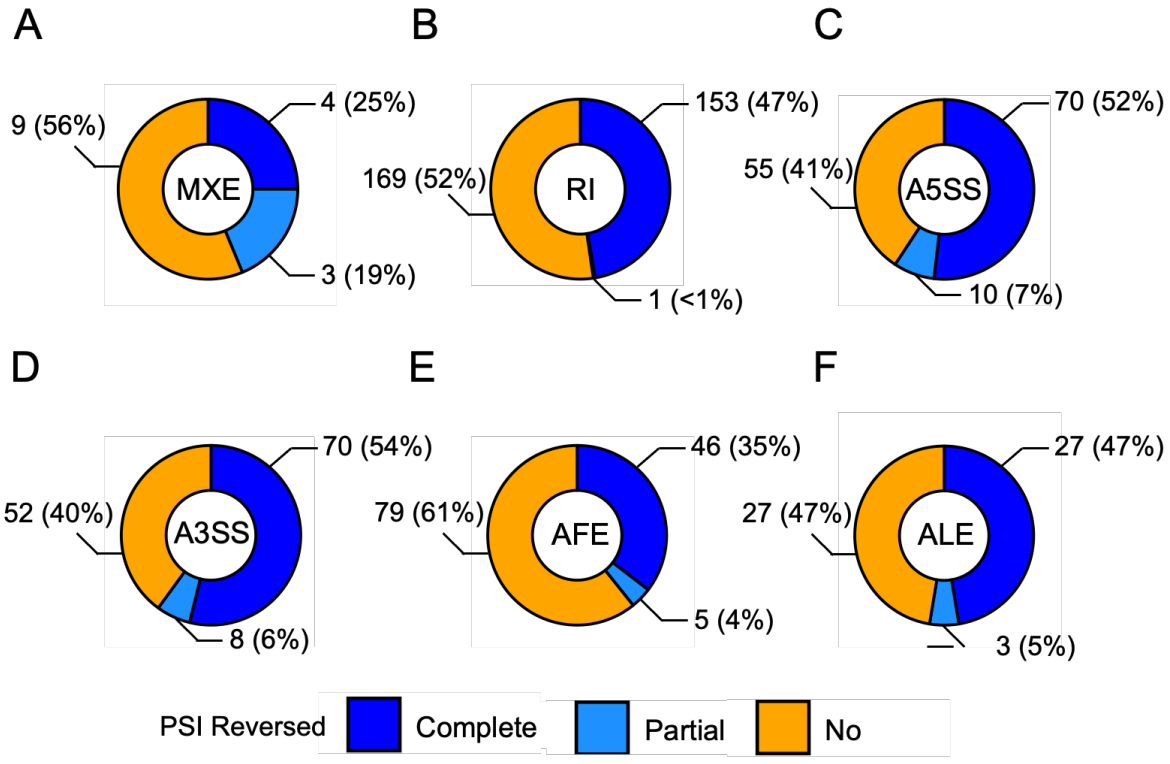


Figure 6.32: Assessing dCas13 treatment efficiency in reversing DM1-associated mis-spliced events, other than SE splicing event type, identified from DM1 NT dCas13 vs WT NT dCas13 comparison. (A-F) dCas13 treatment efficiency in reversing DM1-associated **(A)** MXE, **(B)** RI, **(C)** A5SS, **(D)** A3SS, **(E)** AFE, and **(F)** ALE mis-spliced events.

In total, 827 differentially spliced SE events were identified in DM1 NT ASO vs WT NT ASO comparison (Figure 6.28B), of which 777 (94%) were expressed in DM1 high-dose ASO vs WT NT ASO comparison and were therefore subsequently included in our assessment of high-dose ASO treatment efficiency. Hierarchical clustering using these 777 splicing events revealed high-dose ASO treated samples did not cluster in close proximity with WT NT ASO samples (Figure 6.33A). This suggests that high-dose ASO treatment was largely unsuccessful in reversing DM1-associated mis-spliced events. Indeed, only 141 (18%) mis-spliced events were completely reversed, and another 20 (3%) mis-spliced events were partially reversed, whereas majority, specifically 616 (79%), of mis-spliced events had no reversal upon high-dose ASO treatment (Figures 6.33B-C).

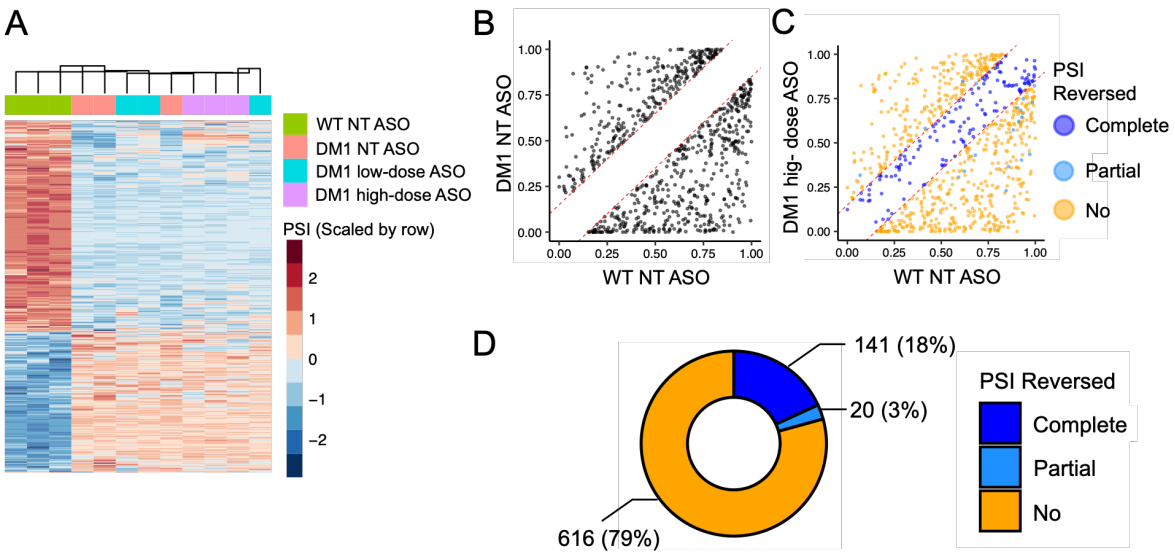


Figure 6.33: Assessing high-dose ASO treatment efficiency in reversing DM1-associated SE mis-spliced events identified from DM1 NT ASO vs WT NT ASO

comparison and expressed in DM1 high-dose ASO vs WT NT ASO comparison. **(A)** Hierarchical clustering using DM1-associated mis-spliced events. **(B)** DM1-associated mis-spliced events identified from DM1 NT ASO vs WT NT ASO comparison. **(C)** The same splicing events in (B) demonstrated for DM1 high-dose ASO treated vs WT NT ASO comparison. **(D)** Number of DM1-associated mis-spliced events stratified into complete, partial, or no reversal when treated with high-dose ASO.

Given the minimal reversal of DM1-associated mis-spliced events by high-dose ASO treatment, it was therefore not surprising that low-dose ASO treatment similar demonstrated minimal reversal of DM1-associated mis-spliced events. Of the 827 differential splicing SE events that were identified in DM1 NT ASO vs WT NT ASO comparison (Figure 6.28B), 780 (94%) were expressed in DM1 low-dose ASO treated vs WT NT ASO comparison and were therefore subsequently included in our assessment of low-dose ASO treatment efficiency. Hierarchical clustering using these 780 splicing events revealed low-dose ASO treated samples did not cluster in close proximity with WT NT ASO samples (Figure 6.34A). This suggests that low dose ASO treatment was largely unsuccessful in reversing DM1-associated mis-spliced events. Indeed, only 120 (15%) mis-spliced events were completely reversed, and another 14 (2%) mis-spliced events were partially reversed, whereas majority, specifically 646 (83%), of mis-spliced events had no reversal upon low-dose ASO treatment (Figures 6.34B-C).

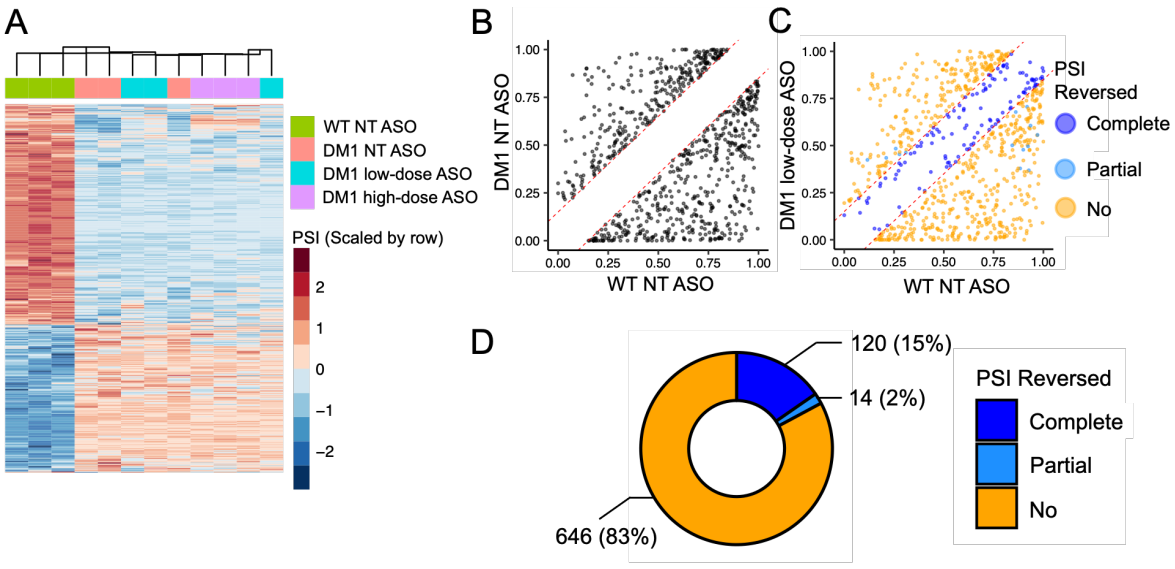


Figure 6.34: Assessing low-dose ASO treatment efficiency in reversing DM1-associated SE mis-spliced events identified from DM1 NT ASO vs WT NT ASO comparison and expressed in DM1 low-dose ASO vs WT NT ASO comparison. **(A)** Hierarchical clustering using DM1-associated mis-spliced events. **(B)** DM1-associated mis-spliced events identified from DM1 NT ASO vs WT NT ASO comparison. **(C)** The same splicing events in (B) demonstrated for DM1 low-dose ASO treated vs WT NT ASO comparison. **(D)** Number of DM1-associated mis-spliced events stratified into complete, partial, or no reversal when treated with low-dose ASO.

Lastly, we investigated if dCas13 and ASO treatments led to reversal of DM1-associated down-regulation of selected muscle-related genes (Batra et al., 2017). dCas13 treatment led to increased gene expression of *DMPK*, *MYOG*, *MYH3*, and *MYH1*, but not *MYOD1*, compared to DM1 NT dCas13 sample group, though the expression levels were not reversed to levels that seen in WT NT dCas13 sample group (Figure 6.35A). Therefore, dCas13 treatment led to partial reversal of DM1-associated down-regulation of muscle-related genes. On the other hand, both high- and low-dose ASO treatments did not lead to marked increased in gene expression of any of the selected muscle-related genes compared to DM1 NT ASO sample group (Figure 6.35B). Therefore, dCas13 treatment was more effective in reversing the DM1-associated down-regulation of muscle-related genes compared to ASO treatment.

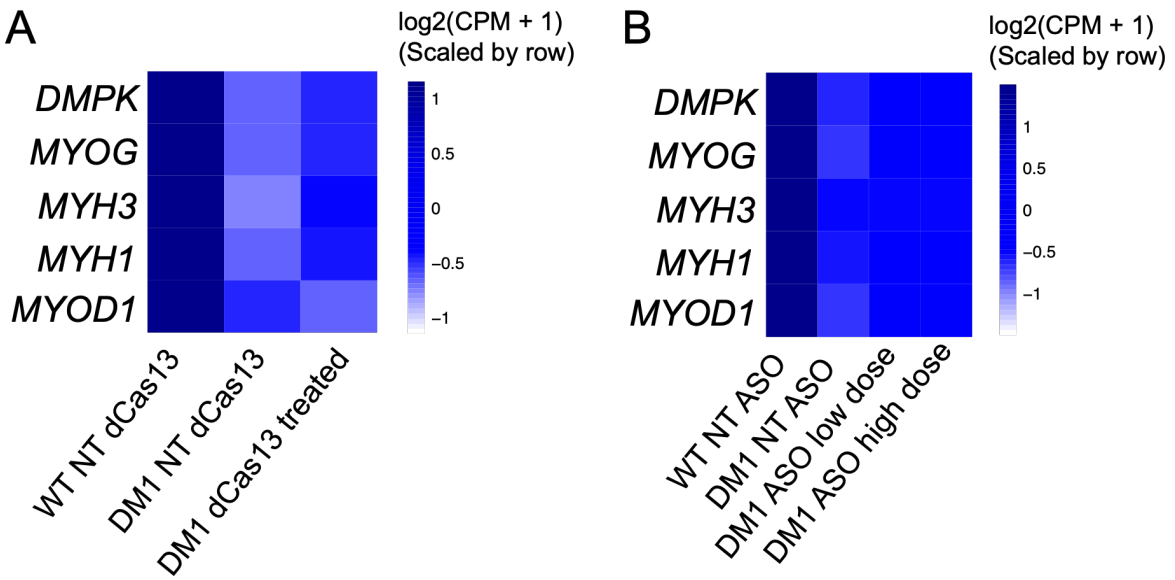


Figure 6.35: Normalised gene expression values of selected muscle-related genes across the different sample groups. Gene expression profile of **(A)** dCas13 treated and **(B)** ASO treated sample groups.

Taken together, dCas13 treatment demonstrated better efficiency, compared to ASO treatment in terms of reversing of DM1-associated mis-spliced events and reversing DM1-associated down-regulation of muscle-related genes. Nevertheless, it is noteworthy that these conclusions were based on DM1 dCas13/ASO treated vs WT NT dCas13/ASO comparison. More robust conclusions may be drawn from performing DM1 dCas13/ASO treated vs WT dCas13/ASO treated comparison (Batra et al., 2017). The latter sample group was unfortunately not included in our study design. This approach would enable us to cancel out differentially spliced events driven by off-target effects of dCas13/ASO treatment. Therefore, we anticipate the true reversal rate of dCas13/ASO treatment to be higher than that reported in our study.

7 Discussion

7.1 Overview and implication

We have developed and benchmarked a computational pipeline consisting of three modules for single-cell splicing analysis, namely MARVEL, VALERIE, and IMPACT (Figure 7.1). MARVEL enables comprehensive characterisation of the splicing landscape for scRNA-seq data generated from plate-based (e.g., Smart-seq2) and droplet-based (e.g., 10x Genomics) platforms to reveal novel biological insights. VALERIE enables visual-based validation of candidate spliced genes identified by MARVEL to identify true positive results. IMPACT enables clinically relevant and druggable biomarkers from highly confident spliced genes validated by VALERIE.

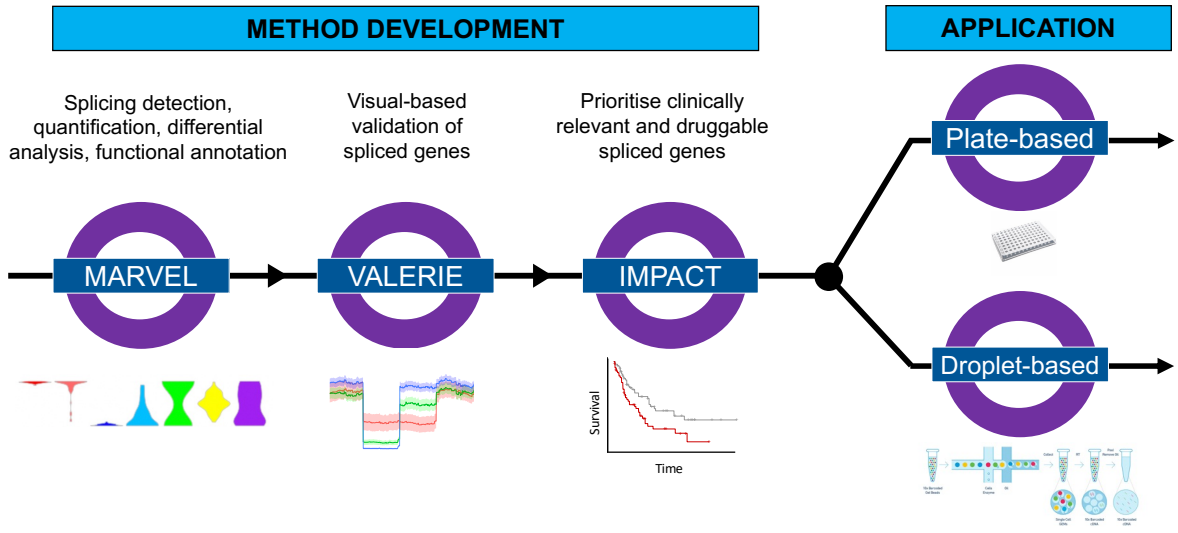


Figure 7.1: Overview of our computational pipeline for single-cell splicing analyses. The computational framework consists of three modules, namely MARVEL, VALERIE, and IMPACT, and applies to scRNA-seq data generated from both plate- and droplet-based platforms.

The implication of the work is to interrogate the single-cell splicing landscape to reveal novel biological insights into health and disease states. We have demonstrated our computational framework for plate-based (e.g., Smart-seq2) and droplet-based (e.g., 10x Genomics) datasets. We also demonstrated our computational framework on homogeneous (i.e., cell lines) and heterogeneous (i.e., haematopoietic stem and progenitor cells) samples. Finally, while our computational

framework focuses on the splicing analysis of single cells, we have also demonstrated its capability to analyse bulk samples.

We believe that our computational framework will continue to be prospectively applied to RNA-seq datasets generated from studies with different disease models and different biological questions.

7.2. Limitations and future work: From correlation to "mechanism"

Ideally, several limitations may need to be addressed prior to the wider application of our computational pipeline. In particular, additional features may be added to MARVEL and IMPACT to improve our ability to identify consequential and actionable spliced genes for downstream experimental studies.

7.2.1 Splicing-mediated activation of protein function

Differential splicing analysis typically returns hundreds, and sometimes thousands, of differentially spliced genes. Therefore, functional annotation of differentially spliced genes would be of particular importance to understand the consequence of aberrant splicing on gene and protein function.

MARVEL currently offers pathway enrichment analysis and nonsense-mediated decay (NMD) prediction for functional annotation of differentially spliced genes. Pathway enrichment analysis enables us to broadly identify the genes with similar functions or genes related to similar biochemical pathways that are co-ordinatedly spliced. NMD enables us to identify genes whose protein function may be disrupted by splicing. For example, we identified *SF3B1*^{K666}-related NMD of *MAP3K7* and *SRSF2*^{P95}-related NMD of *EZH2*.

However, MARVEL does not identify genes whose protein function may be hyper-activated by splicing. This may occur when the insertion (splicing in) or removal (splicing out) of an alternative exon does not interrupt the open reading frame (ORF) of the gene. For example, the insertion of an alternative exon in *GNAS* leads to hyperactivation of the G protein activation and signalling (Wheeler et al., 2022). Another example is the insertion of an alternative exon in *IRAK4* that leads to hyperactivation of NF-κB signalling via increased phosphorylation of IRAK1, p65, p38, and JNK (Smith et al., 2019).

To identify splicing-mediated hyper-activation of protein function, MARVEL may incorporate *in silico* protein structure prediction and sequence-based prediction of

protein-protein interaction. The former approach predicted the inclusion (splicing in) of the alternative exon in *GNAS* led to the expression of a hinge-like domain. This hinge-like domain led *GNAS* to have a higher affinity for GTP and consequently led to hyperactivation of the G protein signalling pathway (Wheeler et al., 2022). The latter approach predicted the inclusion (splicing in) of the alternative exon in *IRAK4* led to the expression of the N-terminal death domain. This domain was predicted to bind to MyD88 and IRAK1 and consequently the phosphorylation of these proteins and ultimately hyperactivation of the NF-κB signalling pathway (Smith et al., 2019).

Taken together, NMD prediction may aid in identifying splicing-mediated deactivation of protein function, while *in silico* protein structural prediction and sequenced-based prediction of protein-protein interactions may aid in identifying splicing-mediated hyperactivation of protein function. Collectively, these features may enable prediction of the functional consequence of aberrant splicing on the protein level. In cancer, the former feature would be relevant in identifying tumour suppressors, while the latter would identify oncogenes.

7.2.2 Sequence motif analysis

MARVEL can identify alternative exons with significantly increased or decreased percent spliced-in (PSI) values in one cell group against another cell group, for example, in cells with genetic variants in splicing factors against cells with wildtype splicing factors. However, it would be of particular interest to attribute a given differentially spliced exon to a given splicing factor. This may increase our confidence in identifying candidate spliced genes for experimental validation.

Each splicing factor, i.e., *SF3B1*, *SRSF2*, and *U2AF1*, is associated with a specific sequence motif. For example, *SRSF2*^{P95} is associated with inclusion (splicing in) of exons with CCNG sequence motif, whereas it is associated with skipping (splicing out) of exons with GGNG sequence motif (Shiozawa et al., 2018). *U2AF1*^{S34} is associated with the inclusion of exon with [C/A]AG sequence motif at the acceptor splice site, whereas it is associated with the exclusion of exon with [T/C]AG sequence motif at the acceptor splice site. Aside from SF3B1, SRSF2, and U2AF1, other RNA-binding proteins (RBPs) that do not commonly harbour genetic variants in myeloid neoplasm but nevertheless may be differentially expressed, have specific sequence

motif (Dominguez et al., 2018; Shiozawa et al., 2018). For example, *RBM39* is up-regulated in AML patients relative to healthy controls (E. Wang et al., 2019).

7.2.3 RNA-splicing factor interaction

IMPACT pre-processes and incorporates external myeloid neoplasm datasets to validate novel differentially spliced genes identified from MARVEL.

Currently, differentially spliced genes from MARVEL are validated using The Cancer Genome Atlas (TCGA) acute myeloid leukaemia (AML) and BeatAML cohorts incorporated into IMPACT. We have shown that splicing events that were statistically significant in our single-cell datasets and successfully validated TCGA and BeatAML were likely to be true positive results. For example, *SF3B1*^{K666} HSC/MEP-associated mis-spliced events that were successfully validated in BeatAML were enriched for A3SSs located within 10-30bp upstream of the canonical splice sites. This is consistent with the binding preference of *SF3B1*^{MUT} at 10-30bp upstream of the canonical splice sites (Alsafadi et al., 2016).

One possible way to confidently attribute a given differentially spliced exon to a specific splicing factor is by using eCLIP-sequencing. eCLIP-sequencing can map the binding sites of splicing factors and RNA-binding proteins (RBPs) to their target RNAs (Van Nostrand et al., 2016). eCLIP-sequencing data has been generated for *SF3B1*^{WT} (K. Wang et al., 2019), *SRSF2*^{P95} (Wheeler et al., 2022), *U2AF1*^{S34} (Wheeler et al., 2022), and *U2AF1*^{Q157} (Biancon et al., 2021).

IMPACT will be able to pre-process and incorporate these published eCLIP-sequencing datasets that indicate the genomic loci bounded by specific splicing factors.

7.2.4 Essential isoform screen

High throughput CRISPR and RNAi screens can identify genes that are essential to cancer cell survival. The Cancer Dependency Map Project generated genetic vulnerability data for various cancer types (Tsherniak et al., 2017). There also exists genetic vulnerability data for specific cancer types, such as acute myeloid leukaemia (AML) (Tzelepis et al., 2016).

Broadly, these screens categorise genes into non-essential genes, non-specific essential genes, and cancer-specific genes (Harman et al., 2021). Non-essential

genes are not required for cancer cell survival. Non-specific essential genes are required for cancer cell survival across most cancer types, such as *MYC*. Cancer-specific genes are required for cancer cell survival for specific cancer types, such as *RUNX1* in MV4-11 and MOLM-13 cell lines.

To date, most genetic vulnerability data are available for genes but not for the isoform expression (Davies et al., 2021). Isoform-specific vulnerability data may be of particular interest, especially in myeloid neoplasm whereby >50% of patients have genetic variants in splicing factor genes.

IMPACT will be able to pre-process and incorporate publicly available CRISPR and RNAi screens to identify isoforms that are essential to cancer cell survival.

7.2.5 Updated framework for prioritising candidate spliced genes

We believe the aforementioned new features for MARVEL and IMPACT, together with existing features, will enable us to identify more reliable and actionable candidate spliced genes (Figure 7.2).

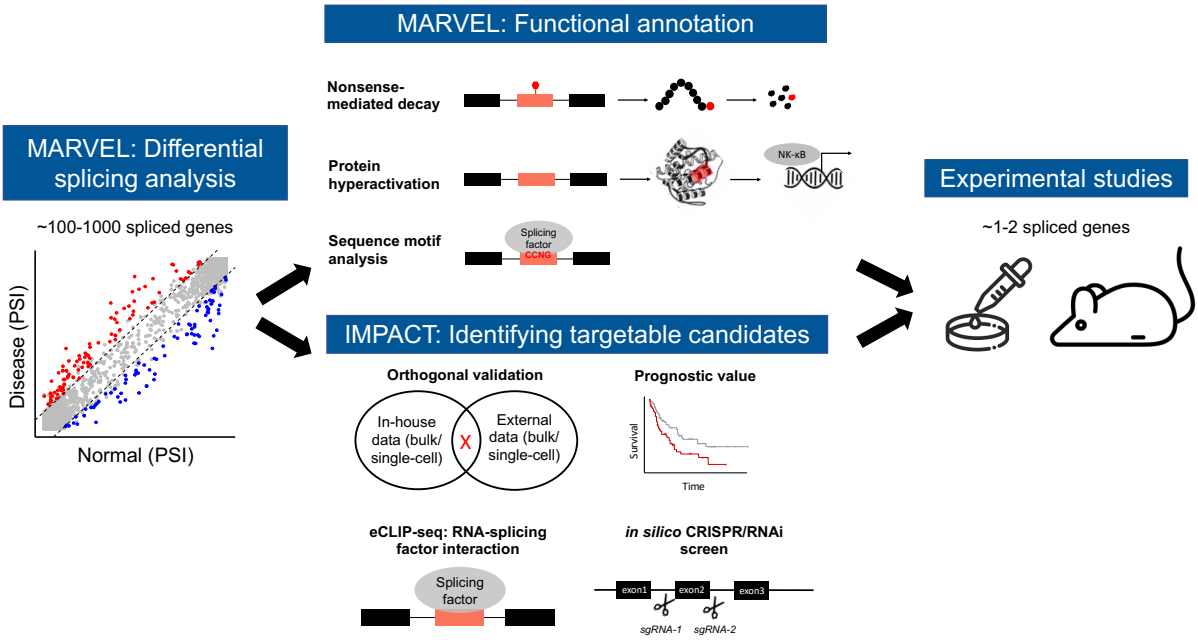


Figure 7.2: An updated pipeline (middle) for prioritising candidate spliced genes identified from transcriptome-wide differential splicing analysis (from MARVEL; left) for downstream experimental studies (right). Features already available are NMD prediction by MARVEL, and orthogonal validation and survival analysis by

IMPACT. Additional features to be provided by MARVEL are predicting impact of splicing on protein function and to identify splicing factor binding sites based on sequence motifs. Additional annotations to be provided by IMPACT are splicing factor binding sites based on eCLIP experiments and essential splicing events (events in which cancer cells depend on for survival) based on *in silico* CRISPR screening.

7.3 Potential applications: From bench to biological insights

With the establishment of a splicing analysis framework applicable to both single-cell and bulk RNA-seq that integrates both splicing and gene expression data, we can proceed with the application of our framework to answer splicing-oriented biological questions that may be relevant for discovering novel and improved therapies.

The applications detailed below aim to answer alternative splicing biological questions that have hitherto not been thoroughly addressed by the literature.

7.3.1 Influence of epigenetics on RNA mis-splicing

In myeloid neoplasm, genetic variants in splicing factors and epigenetic regulators such as *IDH1*, *TET2*, and *AXSL1* have been frequently reported (Cancer Genome Atlas Research et al., 2013; Tyner et al., 2018).

It is conceivable that the presence of genetic variants in both splicing factor and epigenetic regulator genes may engender a different RNA mis-splicing profile compared to when genetic variants are present in either splicing factor or epigenetic regulator genes alone. Indeed, AMLs with genetic variants in *IDH2* and *SRSF2* have different splicing profiles compared to AMLs with genetic variants in either *IDH2* or *SRSF2* alone (Yoshimi et al., 2019). The presence of the CCNG sequence motif in *INTS3* exon 4 was associated with increased inclusion (splicing in) of this exon in *SRSF2*^{P95} AMLs. On the other hand, AMLs with *IDH2* genetic variants had increased DNA methylation, including at *INTS3* exon 4. AMLs with genetic variants in both *IDH2* and *SRSF2* had increased *INTS3* exon 4 inclusion (splicing in) compared to AMLs with genetic variants in either *IDH2* or *SRSF2* alone.

Given the influence of DNA methylation and RNA mis-splicing, it is conceivable that genetic variants in other epigenetic regulators, other than *IDH2*, may also

collaborate with genetic variants in splicing factors. One such epigenetic regulator is *TET2*.

TET2 promotes DNA de-methylation by converting 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC). Myeloid neoplasm patients with genetic variants in *TET2* have uniformly low levels of 5hmC, and the decrease in 5hmC was associated with global hypomethylation at CpG sites (Ko et al., 2010). This is in contrast to myeloid neoplasm patients with genetic variants in *IDH2,* which demonstrated DNA hypermethylation (Yoshimi et al., 2019). Therefore, it is conceivable that myeloid neoplasm patients with genetic variants in both *SRSF2* and *TET2* may have different mis-splicing profiles compared to patients with genetic variants in both *SRSF2* and *IDH2*.

We have characterised the RNA mis-splicing landscape of five *SRSF2*^{P95} myelodysplastic syndrome (MDS) patients and one healthy donor (see "section 6.3.2: Single-cell analysis of *SRSF2*-mutant MDS patients"). While several of these patients had genetic variants in *SRSF2* and epigenetic regulators *TET2* and *ASXL1*, we did not scrutinise the influence of the genetic variants in these epigenetic regulators on the mis-splicing profile engendered by *SRSF2*^{P95}. This was due to the small number of patients included in this pilot phase of the study.

To this end, we plan to sequence additional patients to increase our statistical power for investigating differences in mis-splicing profile between MDS patients with genetic variants in *SRSF2* against patients with genetic variants in *SRSF2* and epigenetic regulators *TET2* and *ASXL1* (Figure 7.3).

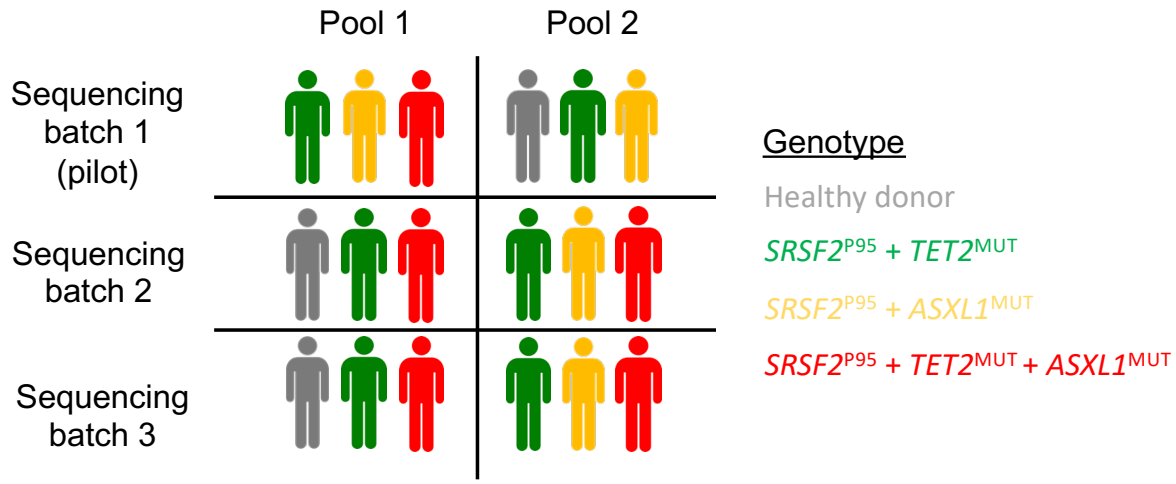


Figure 7.3: MDS cohort to be included for 10x Genomics sequencing to study the effect of genetic variants in epigenetic regulators on *SRSF2*^{P95}-associated mis-splicing. Sequencing batch 1 was our pilot study, and preliminary results for this batch have been described in Section 6.3.2.

7.3.2 Splicing-based stratification of myeloid neoplasms

The most well characterised splicing factor hotspot variants in myeloid neoplasm are *SF3B1*^{K700}, *SRSF2*^{P95}, and *U2AF1*^{S34}. *U2AF1*^{Q157} was also reported in myeloid neoplasm (Shiozawa et al., 2018; Yoshimi et al., 2019), but the mis-splicing profile engendered by this hotspot variant has yet to be comprehensively characterised. This may be partly due to the rarity of this hotspot variant and the relatively small myeloid neoplasm cohorts (~100-200 patients) available to date (Shiozawa et al., 2018; Tyner et al., 2018). Indeed, of the 200 acute myeloid leukaemia (AML) patients in TCGA, only one *U2AF1*^{Q157} patient was identified (Ilagan et al., 2015).

BeatAML is a recent massive publicly available AML cohort consisting of 672 tumour specimens collected from 562 patients (Tyner et al., 2018), of which *U2AF1*^{Q157} genetic variant was identified in 12 specimens (Gao et al., 2013). This may allow us to characterise the mis-splicing profile engendered by *U2AF1*^{Q157} in clinical samples for the first time. We also have scRNA-seq data for phenotypically defined haematopoietic stem cells (HSCs) derived from *U2AF1*^{Q157} myeloproliferative neoplasm (MPN) patients. The analysis of *U2AF1*^{Q157} HSCs may identify mis-spliced genes associated with disease development and progression.

U2AF1^{Q157} myeloid neoplasm patients have worse survival outcomes compared to *U2AF1*^{WT} patients (Tefferi et al., 2018). The more aggressive clinical phenotype engendered by *U2AF1*^{Q157} may offer an opportunity to identify a mis-splicing signature to stratify high-risk myeloid neoplasm patients, irrespective of the presence of genetic variants in splicing factor genes (Anande et al., 2020) (Figure 7.4).

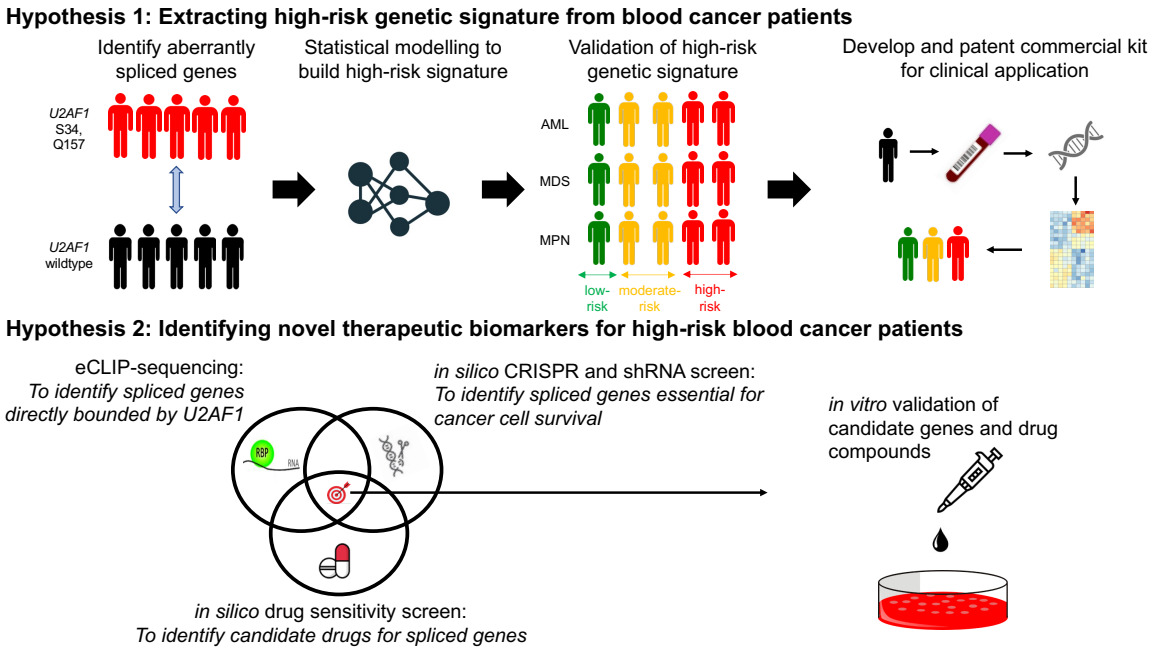


Figure 7.4: Identifying *U2AF1*^{Q157} signature for risk stratification of myeloid neoplasms. Hypothesis 1 focuses on developing the *U2AF1*^{Q157} signature. Hypothesis 2 focuses on candidate spliced gene selection and experimental studies to identify mis-spliced genes that may potentially explain the aggressive clinical phenotype engendered by *U2AF1*^{Q157}.

7.3.3 Splicing-induced neo-antigens for immunotherapy

We have investigated mis-splicing attributed to *trans*-acting genetic variants in splicing factors *SF3B1*, *SRSF2*, and *U2AF1* in myeloid neoplasms, and mis-splicing attributed to dysregulated gene expression of the splicing factor *MBNL1* in myotonic dystrophy type 1 (DM1) patients.

The third mechanism by which mis-splicing may occur is by *cis*-acting genetic variants, particularly in cancers. Somatic variants may introduce novel splice sites and consequently alter the open reading frame (ORF) of the isoform. The change in ORF may introduce a premature stop codon (PTC) and consequently subject the isoform to nonsense-mediated decay (NMD). The resulting degraded peptides are presented on the cancer cell surface. These neo-antigens may elicit an immune response against the cancer cells (Smart et al., 2018).

The effect of *cis*-acting somatic variants on mis-splicing and ultimately neo-antigen generation was studied across 32 solid cancer types from TCGA (Jayasinghe

et al., 2018; Kahles et al., 2018), but the effect of *cis*-acting somatic variants on mis-splicing has not been investigated in blood cancers such as acute myeloid leukaemia (AML). It is noteworthy that *trans*-acting somatic variants on *SF3B1* on mis-splicing and neo-antigen generation have been characterised in myeloproliferative neoplasm (MPN) patients (Schischlik et al., 2019). This hints at mis-splicing, either by *cis*- or *trans*-acting variants, as a potential source of neo-antigens in blood cancers.

Therefore, *cis*-acting somatic variants may also have the potential to generate splicing-induced neo-antigens in myeloid neoplasms. Here, the NMD prediction algorithm by MARVEL will be useful in predicting spliced genes subjected to NMD, and by extension, neo-antigen creation. The characterisation of the repertoire of splicing-induced neo-antigen in myeloid neoplasm from publicly available datasets such as BeatAML and TCGA may aid in assessing the potential utility of immunotherapy in myeloid neoplasm patients (Figure 7.5). Neo-antigens are associated with immune response, and this immune signature may serve as a potential biomarker for the immunotherapy (Wen & Leong, 2019; Wen et al., 2016).

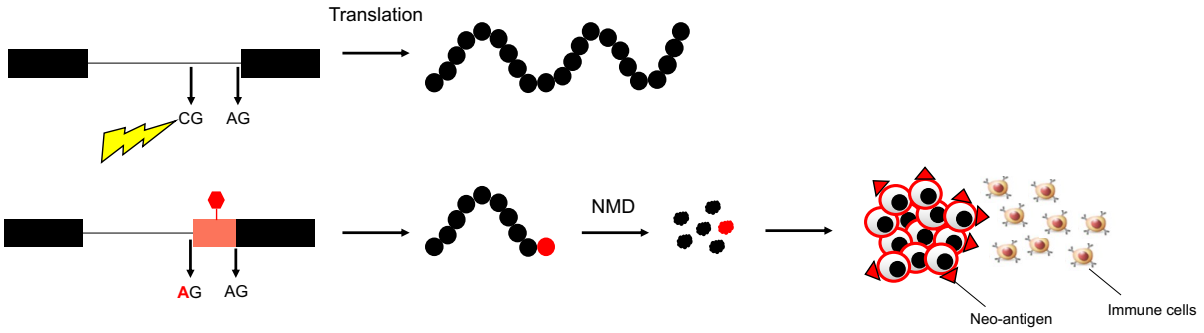


Figure 7.5: *cis*-acting somatic variant in cancer leading to aberrant splicing, neo-antigen presentation, and immune response. These variants may introduce a cryptic splice site and disrupt the ORF of the isoform, ultimately leading to NMD and neo-antigen presentation on the cancer cell surface and immune response against these neo-antigens.

7.4 Conclusion

We have developed a computational framework for single-cell splicing analysis and demonstrated its application on plate- and droplet-based RNA-seq datasets. Our

framework will provide the foundation for adding new functionalities to more comprehensively characterise the splicing landscape and identify reliable candidate spliced genes for downstream experimental studies. We anticipate our framework to be prospectively applied to reveal biological insights in both health and disease states, and across different scientific disciplines, including haematology, immunology, drug discovery, and basic science.

8 References

A Victor Hoffbrand, D. R. H., David M Keeling, Atul B Mehta. (2016). *Postgraduate Haematology* (7th ed.): Wiley Blackwell.

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., . . . Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism, 4*(1), 36. doi:10.1186/2040-2392-4-36

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., . . . Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature, 500*(7463), 415-421. doi:10.1038/nature12477

Alsafadi, S., Houy, A., Battistella, A., Popova, T., Wassef, M., Henry, E., . . . Stern, M. H. (2016). Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun, 7*, 10615. doi:10.1038/ncomms10615

Anande, G., Deshpande, N. P., Mareschal, S., Batcha, A. M. N., Hampton, H. R., Herold, T., . . . Pimanda, J. E. (2020). RNA Splicing Alterations Induce a Cellular Stress Response Associated with Poor Prognosis in Acute Myeloid Leukemia. *Clin Cancer Res, 26*(14), 3597-3607. doi:10.1158/1078-0432.CCR-20-0184

Arandel, L., Polay Espinoza, M., Matloka, M., Bazinet, A., De Dea Diniz, D., Naouar, N., . . . Furling, D. (2017). Immortalized human myotonic dystrophy muscle cell lines to assess therapeutic compounds. *Dis Model Mech, 10*(4), 487-497. doi:10.1242/dmm.027367

Bamopoulos, S. A., Batcha, A. M. N., Jurinovic, V., Rothenberg-Thurley, M., Janke, H., Ksienzyk, B., . . . Herold, T. (2020). Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with acute myeloid leukemia. *Leukemia, 34*(10), 2621-2634. doi:10.1038/s41375-020-0839-4

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., . . . Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature, 483*(7391), 603-607. doi:10.1038/nature11003

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., . . . Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and

lineage dependencies targeted by small molecules. *Cell, 154*(5), 1151-1161. doi:10.1016/j.cell.2013.08.003

Batra, R., Nelles, D. A., Pirie, E., Blue, S. M., Marina, R. J., Wang, H., . . . Yeo, G. W. (2017). Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. *Cell, 170*(5), 899-912 e810. doi:10.1016/j.cell.2017.07.010

Bergot, T., Lippert, E., Douet-Guilbert, N., Commet, S., Corcos, L., & Bernard, D. G. (2020). Human Cancer-Associated Mutations of SF3B1 Lead to a Splicing Modification of Its Own RNA. *Cancers (Basel), 12*(3). doi:10.3390/cancers12030652

Biancon, G., Joshi, P., Zimmer, J. T., Hunck, T., Gao, Y., Lessard, M. D., . . . Halene, S. (2021). Multi-omics profiling of U2AF1 mutants dissects pathogenic mechanisms affecting RNA granules in myeloid malignancies. *bioRxiv*, 2021.2004.2022.441020. doi:10.1101/2021.04.22.441020

Bondu, S., Alary, A. S., Lefevre, C., Houy, A., Jung, G., Lefebvre, T., . . . Fontenay, M. (2019). A variant erythroferrone disrupts iron homeostasis in SF3B1-mutated myelodysplastic syndrome. *Sci Transl Med, 11*(500). doi:10.1126/scitranslmed.aav5467

Bonnal, S. C., Lopez-Oreja, I., & Valcarcel, J. (2020). Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol, 17*(8), 457-474. doi:10.1038/s41571-020-0350-x

Bragulat, M., Meyer, M., Macias, S., Camats, M., Labrador, M., & Vilardell, J. (2010). RPL30 regulation of splicing reveals distinct roles for Cbp80 in U1 and U2 snRNP cotranscriptional recruitment. *RNA, 16*(10), 2033-2041. doi:10.1261/rna.2366310

Brooks, A. N., Choi, P. S., de Waal, L., Sharifnia, T., Imielinski, M., Saksena, G., . . . Meyerson, M. (2014). A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS One, 9*(1), e87361. doi:10.1371/journal.pone.0087361

Buchner, D. A., Trudeau, M., & Meisler, M. H. (2003). SCNM1, a putative RNA splicing factor that modifies disease severity in mice. *Science, 301*(5635), 967-969. doi:10.1126/science.1086187

Buen Abad Najar, C. F., Yosef, N., & Lareau, L. F. (2020). Coverage-dependent bias creates the appearance of binary splicing in single cells. *Elife, 9*. doi:10.7554/eLife.54603

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., . . . Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun, 8*, 16027. doi:10.1038/ncomms16027

Cancer Genome Atlas Research, N., Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., . . . Eley, G. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med, 368*(22), 2059-2074. doi:10.1056/NEJMoa1301689

Castro, M. A., Oliveira, M. I., Nunes, R. J., Fabre, S., Barbosa, R., Peixoto, A., . . . Carmo, A. M. (2007). Extracellular isoforms of CD6 generated by alternative splicing regulate targeting of CD6 to the immunological synapse. *J Immunol, 178*(7), 4351-4361. doi:10.4049/jimmunol.178.7.4351

Cesana, M., Guo, M. H., Cacchiarelli, D., Wahlster, L., Barragan, J., Doulatov, S., . . . Daley, G. Q. (2018). A CLK3-HMGA2 Alternative Splicing Axis Impacts Human Hematopoietic Stem Cell Molecular Identity throughout Development. *Cell Stem Cell, 22*(4), 575-588 e577. doi:10.1016/j.stem.2018.03.012

Chatrikhi, R., Mallory, M. J., Gazzara, M. R., Agosto, L. M., Zhu, W. S., Litterman, A. J., . . . Lynch, K. W. (2019). RNA Binding Protein CELF2 Regulates Signal-Induced Alternative Polyadenylation by Competing with Enhancers of the Polyadenylation Machinery. *Cell Rep, 28*(11), 2795-2806 e2793. doi:10.1016/j.celrep.2019.08.022

Chen, L., Kostadima, M., Martens, J. H. A., Canu, G., Garcia, S. P., Turro, E., . . . Rendon, A. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. *Science, 345*(6204), 1251033. doi:10.1126/science.1251033

Chen, Y., Zheng, Y., Gao, Y., Lin, Z., Yang, S., Wang, T., . . . Tong, M. H. (2018). Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res, 28*(9), 879-896. doi:10.1038/s41422-018-0074-y

Cheng, A. W., Shi, J., Wong, P., Luo, K. L., Trepman, P., Wang, E. T., . . . Lodish, H. F. (2014). Muscleblind-like 1 (Mbnl1) regulates pre-mRNA alternative splicing during terminal erythropoiesis. *Blood, 124*(4), 598-610. doi:10.1182/blood-2013-12-542209

Cho, D. H., & Tapscott, S. J. (2007). Myotonic dystrophy: emerging mechanisms for DM1 and DM2. *Biochim Biophys Acta, 1772*(2), 195-204. doi:10.1016/j.bbadis.2006.05.013

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., . . . Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol, 31*(3), 213-219. doi:10.1038/nbt.2514

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., . . . Chang, H. Y. (2018). The chromatin accessibility landscape of primary human cancers. *Science, 362*(6413). doi:10.1126/science.aav1898

Crane, E. K., Kwan, S. Y., Izaguirre, D. I., Tsang, Y. T., Mullany, L. K., Zu, Z., . . . Wong, K. K. (2015). Nutlin-3a: A Potential Therapeutic Opportunity for TP53 Wild-Type Ovarian Carcinomas. *PLoS One, 10*(8), e0135101. doi:10.1371/journal.pone.0135101

Cui, K., Liu, C., Li, X., Zhang, Q., & Li, Y. (2020). Comprehensive characterization of the rRNA metabolism-related genes in human cancer. *Oncogene, 39*(4), 786-800. doi:10.1038/s41388-019-1026-9

Dalton, W. B., Helmenstine, E., Pieterse, L., Li, B., Gocke, C. D., Donaldson, J., . . . DeZern, A. E. (2020). The K666N mutation in SF3B1 is associated with increased progression of MDS and distinct RNA splicing. *Blood Adv, 4*(7), 1192-1196. doi:10.1182/bloodadvances.2019001127

Davies, R., Liu, L., Taotao, S., Tuano, N., Chaturvedi, R., Huang, K. K., . . . Rosenbluh, J. (2021). CRISPRi enables isoform-specific loss-of-function screens and identification of gastric cancer-specific isoform dependencies. *Genome Biol, 22*(1), 47. doi:10.1186/s13059-021-02266-6

Dehghannasiri, R., Olivieri, J. E., Damljanovic, A., & Salzman, J. (2021). Specific splice junction detection in single cells with SICILIAN. *Genome Biology, 22*(1), 219. doi:10.1186/s13059-021-02434-8

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software, 64*(4), 1 - 34. doi:10.18637/jss.v064.i04

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Dolatshad, H., Pellagatti, A., Fernandez-Mercado, M., Yip, B. H., Malcovati, L., Attwood, M., . . . Boultwood, J. (2015). Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. *Leukemia, 29*(8), 1798. doi:10.1038/leu.2015.178

Dolatshad, H., Pellagatti, A., Liberante, F. G., Llorian, M., Repapi, E., Steeples, V., . . . Boultwood, J. (2016). Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. *Leukemia, 30*(12), 2322-2331. doi:10.1038/leu.2016.149

Dolgalev, I., & Tikhonova, A. N. (2021). Connecting the Dots: Resolving the Bone Marrow Niche Heterogeneity. *Front Cell Dev Biol, 9*, 622519. doi:10.3389/fcell.2021.622519

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., . . . Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell, 70*(5), 854-867 e859. doi:10.1016/j.molcel.2018.05.001

Dowd, C. (2020). A New ECDF Two-Sample Test Statistic. *arXiv*.

Efremova, M., Vento-Tormo, M., Teichmann, S. A., & Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc, 15*(4), 1484-1506. doi:10.1038/s41596-020-0292-x

Erben, L., He, M. X., Laeremans, A., Park, E., & Buonanno, A. (2018). A Novel Ultrasensitive In Situ Hybridization Approach to Detect Short Sequences and Splice Variants with Cellular Resolution. *Mol Neurobiol, 55*(7), 6169-6181. doi:10.1007/s12035-017-0834-6

Falcao, A. M., Meijer, M., Scaglione, A., Rinwa, P., Agirre, E., Liang, J., . . . Castelo-Branco, G. (2019). PAD2-Mediated Citrullination Contributes to Efficient Oligodendrocyte Differentiation and Myelination. *Cell Rep, 27*(4), 1090-1102 e1010. doi:10.1016/j.celrep.2019.03.108

Falcao, A. M., van Bruggen, D., Marques, S., Meijer, M., Jakel, S., Agirre, E., . . . Castelo-Branco, G. (2018). Disease-specific oligodendrocyte lineage cells arise in multiple sclerosis. *Nat Med, 24*(12), 1837-1844. doi:10.1038/s41591-018-0236-y

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., . . . Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol, 16*, 278. doi:10.1186/s13059-015-0844-5

Fu, X. D., & Maniatis, T. (1992). The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc Natl Acad Sci U S A, 89*(5), 1725-1729. doi:10.1073/pnas.89.5.1725

Gaiti, F., Chamely, P., Hawkins, A. G., Cortés-López, M., Swett, A. D., Ganesan, S., . . . Landau, D. A. (2022). Single-cell multi-omics defines the cell-type specific impact of splicing aberrations in human hematopoietic clonal outgrowths. *bioRxiv*, 2022.2006.2008.495292. doi:10.1101/2022.06.08.495292

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., . . . Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal, 6*(269), pl1. doi:10.1126/scisignal.2004088

Ghandi, M., Huang, F. W., Jane-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., 3rd, . . . Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature, 569*(7757), 503-508. doi:10.1038/s41586-019-1186-3

Ghasemi, N., Razavi, S., & Nikzad, E. (2017). Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses and Cell-Based Therapy. *Cell J, 19*(1), 1-10. doi:10.22074/cellj.2016.4867

Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A. G., . . . Mead, A. J. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat Med, 23*(6), 692-702. doi:10.1038/nm.4336

Graf, E. M., Bock, M., Heubach, J. F., Zahanich, I., Boxberger, S., Richter, W., . . . Ravens, U. (2005). Tissue distribution of a human Ca v 1.2 alpha1 subunit splice variant with a 75 bp insertion. *Cell Calcium, 38*(1), 11-21. doi:10.1016/j.ceca.2005.03.005

Grancharova, T., Gerbin, K. A., Rosenberg, A. B., Roco, C. M., Arakaki, J. E., DeLizo, C. M., . . . Gunawardane, R. N. (2021). A comprehensive analysis of gene expression changes in a high replicate and open-source dataset of

differentiating hiPSC-derived cardiomyocytes. *Sci Rep, 11*(1), 15845. doi:10.1038/s41598-021-94732-1

Graubert, T. A., Shen, D., Ding, L., Okeyo-Owuor, T., Lunn, C. L., Shao, J., . . . Walter, M. J. (2011). Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet, 44*(1), 53-57. doi:10.1038/ng.1031

Grinfeld, J., Nangalia, J., Baxter, E. J., Wedge, D. C., Angelopoulos, N., Cantrill, R., . . . Campbell, P. J. (2018). Classification and Personalized Prognosis in Myeloproliferative Neoplasms. *N Engl J Med, 379*(15), 1416-1430. doi:10.1056/NEJMoa1716614

Grosso, A. R., Martins, S., & Carmo-Fonseca, M. (2008). The emerging role of splicing factors in cancer. *EMBO Rep, 9*(11), 1087-1093. doi:10.1038/embor.2008.189

Group, P. T. C., Calabrese, C., Davidson, N. R., Demircioglu, D., Fonseca, N. A., He, Y., . . . Consortium, P. (2020). Genomic basis for RNA alterations in cancer. *Nature, 578*(7793), 129-136. doi:10.1038/s41586-020-1970-0

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., . . . Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol*. doi:10.1038/nbt.4259

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramskold, D., Hendriks, G. J., Larsson, A. J. M., . . . Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat Biotechnol, 38*(6), 708-714. doi:10.1038/s41587-020-0497-0

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., & Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature, 504*(7480), 389-393. doi:10.1038/nature12831

Harman, J. R., Thorne, R., Jamilly, M., Tapia, M., Crump, N. T., Rice, S., . . . Milne, T. A. (2021). A KMT2A-AFF1 gene regulatory network highlights the role of core transcription factors and reveals the regulatory logic of key downstream target genes. *Genome Res*. doi:10.1101/gr.268490.120

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res, 22*(9), 1760-1774. doi:10.1101/gr.135350.111

Hasmad, H. N., Lai, K. N., Wen, W. X., Park, D. J., Nguyen-Dumont, T., Kang, P. C. E., . . . Teo, S. H. (2016). Evaluation of germline BRCA1 and BRCA2 mutations

in a multi-ethnic Asian cohort of ovarian cancer patients. *Gynecol Oncol, 141*(2), 318-322. doi:10.1016/j.ygyno.2015.11.001

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., & Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat Commun, 9*(1), 619. doi:10.1038/s41467-018-02866-0

Heaton, H., Talman, A. M., Knights, A., Imaz, M., Gaffney, D. J., Durbin, R., . . . Lawniczak, M. K. N. (2020). Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat Methods, 17*(6), 615-620. doi:10.1038/s41592-020-0820-1

Hentges, L. D., Sergeant, M. J., Cole, C. B., Downes, D. J., Hughes, J. R., & Taylor, S. (2022). LanceOtron: a deep learning peak caller for genome sequencing experiments. *Bioinformatics*. doi:10.1093/bioinformatics/btac525

Hu, N., Kim, E., Antoury, L., Li, J., Gonzalez-Perez, P., Rutkove, S. B., & Wheeler, T. M. (2021). Antisense oligonucleotide and adjuvant exercise therapy reverse fatigue in old mice with myotonic dystrophy. *Mol Ther Nucleic Acids, 23*, 393-405. doi:10.1016/j.omtn.2020.11.014

Hu, Y., Wang, K., & Li, M. (2020). Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS Comput Biol, 16*(6), e1007925. doi:10.1371/journal.pcbi.1007925

Huang, Y., & Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol, 18*(1), 123. doi:10.1186/s13059-017-1248-5

Huang, Y., & Sanguinetti, G. (2021). BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biol, 22*(1), 251. doi:10.1186/s13059-021-02461-5

Hwang, M., Jun, D. W., Kang, E. H., Yoon, K. A., Cheong, H., Kim, Y. H., . . . Kim, S. (2019). EI24, as a Component of Autophagy, Is Involved in Pancreatic Cell Proliferation. *Front Oncol, 9*, 652. doi:10.3389/fonc.2019.00652

Ilagan, J. O., Ramakrishnan, A., Hayes, B., Murphy, M. E., Zebari, A. S., Bradley, P., & Bradley, R. K. (2015). U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res, 25*(1), 14-26. doi:10.1101/gr.181016.114

Inoue, D., Polaski, J. T., Taylor, J., Castel, P., Chen, S., Kobayashi, S., . . . Abdel-Wahab, O. (2021). Minor intron retention drives clonal hematopoietic disorders

and diverse cancer predisposition. *Nat Genet, 53*(5), 707-718. doi:10.1038/s41588-021-00828-9

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res, 21*(7), 1160-1167. doi:10.1101/gr.110882.110

Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., . . . Ding, L. (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep, 23*(1), 270-281 e273. doi:10.1016/j.celrep.2018.03.052

Kahles, A., Lehmann, K. V., Toussaint, N. C., Huser, M., Stark, S. G., Sachsenberg, T., . . . Ratsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell, 34*(2), 211-224 e216. doi:10.1016/j.ccell.2018.07.001

Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, 2021.2005.2005.442755. doi:10.1101/2021.05.05.442755

Kanagal-Shamanna, R., Montalban-Bravo, G., Sasaki, K., Darbaniyan, F., Jabbour, E., Bueso-Ramos, C., . . . Garcia-Manero, G. (2021). Only SF3B1 mutation involving K700E independently predicts overall survival in myelodysplastic syndromes. *Cancer, 127*(19), 3552-3565. doi:10.1002/cncr.33745

Kanumilli, S., Tringham, E. W., Payne, C. E., Dupere, J. R., Venkateswarlu, K., & Usowicz, M. M. (2006). Alternative splicing generates a smaller assortment of CaV2.1 transcripts in cerebellar Purkinje cells than in the cerebellum. *Physiol Genomics, 24*(2), 86-96. doi:10.1152/physiolgenomics.00149.2005

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., . . . University of California Santa, C. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res, 31*(1), 51-54. doi:10.1093/nar/gkg129

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods, 7*(12), 1009-1015. doi:10.1038/nmeth.1528

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., . . . Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell, 161*(5), 1187-1201. doi:10.1016/j.cell.2015.04.044

Ko, M., Huang, Y., Jankowska, A. M., Pape, U. J., Tahiliani, M., Bandukwala, H. S., . . . Rao, A. (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature, 468*(7325), 839-843. doi:10.1038/nature09586

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., . . . Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods, 16*(12), 1289-1296. doi:10.1038/s41592-019-0619-0

Kosti, A., de Araujo, P. R., Li, W. Q., Guardia, G. D. A., Chiou, J., Yi, C., . . . Penalva, L. O. F. (2020). The RNA-binding protein SERBP1 functions as a novel oncogenic factor in glioblastoma by bridging cancer metabolism and epigenetic regulation. *Genome Biol, 21*(1), 195. doi:10.1186/s13059-020-02115-y

Kumazaki, T., Mitsui, Y., Hamada, K., Sumida, H., & Nishiyama, M. (1999). Detection of alternative splicing of fibronectin mRNA in a single cell. *J Cell Sci, 112 (Pt 10)*, 1449-1453. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10212139

Kunkele, A., De Preter, K., Heukamp, L., Thor, T., Pajtler, K. W., Hartmann, W., . . . Schulte, J. H. (2012). Pharmacological activation of the p53 pathway by nutlin-3 exerts anti-tumoral effects in medulloblastomas. *Neuro Oncol, 14*(7), 859-869. doi:10.1093/neuonc/nos115

Kurilov, R., Haibe-Kains, B., & Brors, B. (2020). Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci Rep, 10*(1), 2849. doi:10.1038/s41598-020-59656-2

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol, 9*(8), e1003118. doi:10.1371/journal.pcbi.1003118

Lee, S. C., North, K., Kim, E., Jang, E., Obeng, E., Lu, S. X., . . . Abdel-Wahab, O. (2018). Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell, 34*(2), 225-241 e228. doi:10.1016/j.ccell.2018.07.003

Legnini, I., Alles, J., Karaiskos, N., Ayoub, S., & Rajewsky, N. (2019). FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat Methods, 16*(9), 879-886. doi:10.1038/s41592-019-0503-y

Lessi, F., Scatena, C., Aretini, P., Menicagli, M., Franceschi, S., Naccarato, A. G., & Mazzanti, C. M. (2019). Molecular profiling of microinvasive breast cancer microenvironment progression. *J Transl Med, 17*(1), 187. doi:10.1186/s12967-019-1936-x

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, J., Ugalde-Morales, E., Wen, W. X., Decker, B., Eriksson, M., Torstensson, A., . . . Czene, K. (2018). Differential Burden of Rare and Common Variants on Tumor Characteristics, Survival, and Mode of Detection in Breast Cancer. *Cancer Res, 78*(21), 6329-6338. doi:10.1158/0008-5472.CAN-18-1018

Li, J., Wen, W. X., Eklund, M., Kvist, A., Eriksson, M., Christensen, H. N., . . . Czene, K. (2019). Prevalence of BRCA1 and BRCA2 pathogenic variants in a large, unselected breast cancer cohort. *Int J Cancer, 144*(5), 1195-1204. doi:10.1002/ijc.31841

Li, Y., Chen, J., Xu, Q., Han, Z., Tan, F., Shi, T., & Chen, G. (2021). Single-cell transcriptomic analysis reveals dynamic alternative splicing and gene regulatory networks among pancreatic islets. *Sci China Life Sci, 64*(1), 174-176. doi:10.1007/s11427-020-1711-x

Li, Y., Wang, D., Wang, H., Huang, X., Wen, Y., Wang, B., . . . Shi, L. (2021). A splicing factor switch controls hematopoietic lineage specification of pluripotent stem cells. *EMBO Rep, 22*(1), e50535. doi:10.15252/embr.202050535

Liang, Y., Tebaldi, T., Rejeski, K., Joshi, P., Stefani, G., Taylor, A., . . . Halene, S. (2018). SRSF2 mutations drive oncogenesis by activating a global program of

aberrant alternative splicing in hematopoietic cells. *Leukemia, 32*(12), 2659-2671. doi:10.1038/s41375-018-0152-7

Lieu, Y. K., Liu, Z., Ali, A. M., Wei, X., Penson, A., Zhang, J., . . . Mukherjee, S. (2022). SF3B1 mutant-induced missplicing of MAP3K7 causes anemia in myelodysplastic syndromes. *Proc Natl Acad Sci U S A, 119*(1). doi:10.1073/pnas.2111703119

Linker, S. M., Urban, L., Clark, S. J., Chhatriwala, M., Amatya, S., McCarthy, D. J., . . . Bonder, M. J. (2019). Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biol, 20*(1), 30. doi:10.1186/s13059-019-1644-0

Liu, S., Zhou, B., Wu, L., Sun, Y., Chen, J., & Liu, S. (2021). Single-cell differential splicing analysis reveals high heterogeneity of liver tumor-infiltrating T cells. *Sci Rep, 11*(1), 5325. doi:10.1038/s41598-021-84693-w

Liu, W., & Zhang, X. (2020). Single-cell alternative splicing analysis reveals dominance of single transcript variant. *Genomics, 112*(3), 2418-2425. doi:10.1016/j.ygeno.2020.01.014

Liu, Z., & Rabadan, R. (2021). Computing the Role of Alternative Splicing in Cancer. *Trends Cancer, 7*(4), 347-358. doi:10.1016/j.trecan.2020.12.015

Liu, Z., Yoshimi, A., Wang, J., Cho, H., Chun-Wei Lee, S., Ki, M., . . . Rabadan, R. (2020). Mutations in the RNA Splicing Factor SF3B1 Promote Tumorigenesis through MYC Stabilization. *Cancer Discov, 10*(6), 806-821. doi:10.1158/2159-8290.CD-19-1330

Ma, Y., Zhang, H., Yang, X., Li, Y., Guan, J., Lv, Y., . . . Gai, Z. (2019). Establishment of a human induced pluripotent stem cell line (SDQLCHi004-A) from a patient with nemaline myopathy-4 disease carrying heterozygous mutation in TPM2 gene. *Stem Cell Res, 40*, 101559. doi:10.1016/j.scr.2019.101559

Macaulay, I. C., Teng, M. J., Haerty, W., Kumar, P., Ponting, C. P., & Voet, T. (2016). Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat Protoc, 11*(11), 2081-2103. doi:10.1038/nprot.2016.138

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., . . . McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell, 161*(5), 1202-1214. doi:10.1016/j.cell.2015.05.002

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., . . . Koeffler, H. P. (2015). Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun, 6*, 6042. doi:10.1038/ncomms7042

Manipur, I., Granata, I., & Guarracino, M. R. (2019). Exploiting single-cell RNA sequencing data to link alternative splicing and cancer heterogeneity: A computational approach. *Int J Biochem Cell Biol, 108*, 51-60. doi:10.1016/j.biocel.2018.12.015

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., & Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res, 24*(3), 496-510. doi:10.1101/gr.161034.113

Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal, 17*, 10-12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Munoz, J. F., Delorey, T., Ford, C. B., Li, B. Y., Thompson, D. A., Rao, R. P., & Cuomo, C. A. (2019). Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with Candida albicans. *Nat Commun, 10*(1), 1607. doi:10.1038/s41467-019-09599-8

Nakamori, M., Hamanaka, K., Thomas, J. D., Wang, E. T., Hayashi, Y. K., Takahashi, M. P., . . . Mochizuki, H. (2017). Aberrant Myokine Signaling in Congenital Myotonic Dystrophy. *Cell Rep, 21*(5), 1240-1252. doi:10.1016/j.celrep.2017.10.018

Ng, P. S., Wen, W. X., Fadlullah, M. Z., Yoon, S. Y., Lee, S. Y., Thong, M. K., . . . Teo, S. H. (2016). Identification of germline alterations in breast cancer predisposition genes among Malaysian breast cancer patients using panel testing. *Clin Genet, 90*(4), 315-323. doi:10.1111/cge.12735

Nik-Zainal, S., Wedge, D. C., Alexandrov, L. B., Petljak, M., Butler, A. P., Bolli, N., . . . Stratton, M. R. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet, 46*(5), 487-491. doi:10.1038/ng.2955

Ntranos, V., Yi, L., Melsted, P., & Pachter, L. (2019). A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat Methods, 16*(2), 163-166. doi:10.1038/s41592-018-0303-9

Okeyo-Owuor, T., White, B. S., Chatrikhi, R., Mohan, D. R., Kim, S., Griffith, M., . . . Graubert, T. A. (2015). U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia, 29*(4), 909-917. doi:10.1038/leu.2014.303

Ou, M., Zhao, M., Li, C., Tang, D., Xu, Y., Dai, W., . . . Dai, Y. (2021). Single-cell sequencing reveals the potential oncogenic expression atlas of human iPSC-derived cardiomyocytes. *Biol Open, 10*(2). doi:10.1242/bio.053348

Ozaki, H., Hayashi, T., Umeda, M., & Nikaido, I. (2020). Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets. *BMC Genomics, 21*(1), 177. doi:10.1186/s12864-020-6542-z

Pagès, H. A., P.; Gentleman, R.; DebRoy, S. (2021). Biostrings: Efficient manipulation of biological strings. *R package version 2.62.0*.

Pangallo, J., Kiladjian, J. J., Cassinat, B., Renneville, A., Taylor, J., Polaski, J. T., . . . Bradley, R. K. (2020). Rare and private spliceosomal gene mutations drive partial, complete, and dual phenocopies of hotspot alterations. *Blood, 135*(13), 1032-1043. doi:10.1182/blood.2019002894

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., . . . Chronic Myeloid Disorders Working Group of the International Cancer Genome, C. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med, 365*(15), 1384-1395. doi:10.1056/NEJMoa1103283

Park, S. M., Ou, J., Chamberlain, L., Simone, T. M., Yang, H., Virbasius, C. M., . . . Green, M. R. (2016). U2AF35(S34F) Promotes Transformation by Directing Aberrant ATG7 Pre-mRNA 3' End Formation. *Mol Cell, 62*(4), 479-490. doi:10.1016/j.molcel.2016.04.011

Pellagatti, A., Armstrong, R. N., Steeples, V., Sharma, E., Repapi, E., Singh, S., . . . Boultwood, J. (2018). Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood, 132*(12), 1225-1240. doi:10.1182/blood-2018-04-843771

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol, 33*(3), 290-295. doi:10.1038/nbt.3122

Petti, A. A., Williams, S. R., Miller, C. A., Fiddes, I. T., Srivatsan, S. N., Chen, D. Y., . . . Ley, T. J. (2019). A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun, 10*(1), 3660. doi:10.1038/s41467-019-11591-1

Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods, 10*(11), 1096-1098. doi:10.1038/nmeth.2639

Picelli, S., Faridani, O. R., Bjorklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc, 9*(1), 171-181. doi:10.1038/nprot.2014.006

Pimentel, H., Parra, M., Gee, S. L., Mohandas, N., Pachter, L., & Conboy, J. G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res, 44*(2), 838-851. doi:10.1093/nar/gkv1168

Psaila, B., & Mead, A. J. (2019). Single-cell approaches reveal novel cellular pathways for megakaryocyte and erythroid differentiation. *Blood, 133*(13), 1427-1435. doi:10.1182/blood-2018-11-835371

Psaila, B., Wang, G., Rodriguez-Meira, A., Li, R., Heuston, E. F., Murphy, L., . . . Mead, A. J. (2020). Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets. *Mol Cell, 78*(3), 477-492 e478. doi:10.1016/j.molcel.2020.04.008

Puigdevall, P., & Castelo, R. (2018). GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor. *Bioinformatics, 34*(18), 3208-3210. doi:10.1093/bioinformatics/bty311

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y. A., & Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nat Methods, 14*(3), 309-315. doi:10.1038/nmeth.4150

Quesada, V., Conde, L., Villamor, N., Ordonez, G. R., Jares, P., Bassaganyas, L., . . . Lopez-Otin, C. (2011). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet, 44*(1), 47-52. doi:10.1038/ng.1032

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Rahman, M. A., Lin, K. T., Bradley, R. K., Abdel-Wahab, O., & Krainer, A. R. (2020). Recurrent SRSF2 mutations in MDS affect both splicing and NMD. *Genes Dev, 34*(5-6), 413-427. doi:10.1101/gad.332270.119

Rallapalli, R., Strachan, G., Cho, B., Mercer, W. E., & Hall, D. J. (1999). A novel MDMX transcript expressed in a variety of transformed cell lines encodes a truncated protein with potent p53 repressive activity. *J Biol Chem, 274*(12), 8299-8308. doi:10.1074/jbc.274.12.8299

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., . . . Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol, 30*(8), 777-782. doi:10.1038/nbt.2282

Ren, L., Li, J., Wang, C., Lou, Z., Gao, S., Zhao, L., . . . Tang, J. (2021). Single cell RNA sequencing for breast cancer: present and future. *Cell Death Discov, 7*(1), 104. doi:10.1038/s41420-021-00485-1

Ren, X., Deng, R., Zhang, K., Sun, Y., Teng, X., & Li, J. (2018). SpliceRCA: in Situ Single-Cell Analysis of mRNA Splicing Variants. *ACS Cent Sci, 4*(6), 680-687. doi:10.1021/acscentsci.8b00081

Rivera, O. D., Mallory, M. J., Quesnel-Vallieres, M., Chatrikhi, R., Schultz, D. C., Carroll, M., . . . Lynch, K. W. (2021). Alternative splicing redefines landscape of commonly mutated genes in acute myeloid leukemia. *Proc Natl Acad Sci U S A, 118*(15). doi:10.1073/pnas.2014967118

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Rodriguez-Meira, A., Buck, G., Clark, S. A., Povinelli, B. J., Alcolea, V., Louka, E., . . . Mead, A. J. (2019). Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol Cell, 73*(6), 1292-1305 e1298. doi:10.1016/j.molcel.2019.01.009

Roy, A., Wang, G., Iskander, D., O'Byrne, S., Elliott, N., O'Sullivan, J., . . . Thongjuea, S. (2021). Transitions in lineage specification and gene regulatory networks in hematopoietic stem/progenitor cells over human development. *Cell Rep, 36*(11), 109698. doi:10.1016/j.celrep.2021.109698

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol, 33*(5), 495-502. doi:10.1038/nbt.3192

Schischlik, F., Jager, R., Rosebrock, F., Hug, E., Schuster, M., Holly, R., . . . Kralovics, R. (2019). Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood, 134*(2), 199-210. doi:10.1182/blood.2019000519

Schulte-Schrepping, J., Reusch, N., Paclik, D., Bassler, K., Schlickeiser, S., Zhang, B., . . . Deutsche, C.-O. I. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell, 182*(6), 1419-1440 e1423. doi:10.1016/j.cell.2020.08.001

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., . . . Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A, 111*(51), E5593-5601. doi:10.1073/pnas.1419161111

Shiozawa, Y., Malcovati, L., Galli, A., Sato-Otsubo, A., Kataoka, K., Sato, Y., . . . Cazzola, M. (2018). Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat Commun, 9*(1), 3649. doi:10.1038/s41467-018-06063-x

Shirai, C. L., Ley, J. N., White, B. S., Kim, S., Tibbitts, J., Shao, J., . . . Walter, M. J. (2015). Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell, 27*(5), 631-643. doi:10.1016/j.ccell.2015.04.008

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res, 15*(8), 1034-1050. doi:10.1101/gr.3715005

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J. M., Blackburn, J., Barton, K., . . . Swarbrick, A. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun, 10*(1), 3120. doi:10.1038/s41467-019-11049-4

Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., . . . Van Allen, E. M. (2018). Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol, 36*(11), 1056-1058. doi:10.1038/nbt.4239

Smith, M. A., Choudhary, G. S., Pellagatti, A., Choi, K., Bolanos, L. C., Bhagat, T. D., . . . Starczynowski, D. T. (2019). U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies. *Nat Cell Biol, 21*(5), 640-650. doi:10.1038/s41556-019-0314-5

Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., & Yeo, G. W. (2017). Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Mol Cell, 67*(1), 148-161 e145. doi:10.1016/j.molcel.2017.06.003

Springer, J., McGregor, G. P., Fink, L., & Fischer, A. (2003). Alternative splicing in single cells dissected from complex tissues: separate expression of prepro-tachykinin A mRNA splice variants in sensory neurones. *J Neurochem, 85*(4), 882-888. doi:10.1046/j.1471-4159.2003.01720.x

Steinboeck, F., & Kristufek, D. (2005). Identification of the cytolinker protein plectin in neuronal cells - expression of a rodless isoform in neurons of the rat superior cervical ganglion. *Cell Mol Neurobiol, 25*(7), 1151-1169. doi:10.1007/s10571-005-8503-0

Sugiyama, H., Takahashi, K., Yamamoto, T., Iwasaki, M., Narita, M., Nakamura, M., . . . Yamanaka, S. (2017). Nat1 promotes translation of specific proteins that induce differentiation of mouse embryonic stem cells. *Proc Natl Acad Sci U S A, 114*(2), 340-345. doi:10.1073/pnas.1617234114

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun, 11*(1), 1438. doi:10.1038/s41467-020-15171-6

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., . . . Surani, M. A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell, 6*(5), 468-478. doi:10.1016/j.stem.2010.03.015

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., . . . Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods, 6*(5), 377-382. doi:10.1038/nmeth.1315

Tapial, J., Ha, K. C. H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., . . . Irimia, M. (2017). An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res, 27*(10), 1759-1768. doi:10.1101/gr.220962.117

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., . . . Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res, 47*(D1), D941-D947. doi:10.1093/nar/gky1015

Tefferi, A., Finke, C. M., Lasho, T. L., Hanson, C. A., Ketterling, R. P., Gangat, N., & Pardanani, A. (2018). U2AF1 mutation types in primary myelofibrosis: phenotypic and prognostic distinctions. *Leukemia, 32*(10), 2274-2278. doi:10.1038/s41375-018-0078-0

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., . . . Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol, 32*(4), 381-386. doi:10.1038/nbt.2859

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., . . . Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell, 170*(3), 564-576 e516. doi:10.1016/j.cell.2017.06.010

Turan, S., & Bastepe, M. (2013). The GNAS complex locus and human diseases associated with loss-of-function mutations or epimutations within this imprinted gene. *Horm Res Paediatr, 80*(4), 229-241. doi:10.1159/000355384

Tyner, J. W., Tognon, C. E., Bottomly, D., Wilmot, B., Kurtz, S. E., Savage, S. L., . . . Druker, B. J. (2018). Functional genomic landscape of acute myeloid leukaemia. *Nature, 562*(7728), 526-531. doi:10.1038/s41586-018-0623-z

Tzelepis, K., Koike-Yusa, H., De Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O. M., . . . Yusa, K. (2016). A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep, 17*(4), 1193-1205. doi:10.1016/j.celrep.2016.09.079

van der Spek, A., Warner, S. C., Broer, L., Nelson, C. P., Vojinovic, D., Ahmad, S., . . . van Duijn, C. M. (2020). Exome Sequencing Analysis Identifies Rare Variants in ATM and RPL8 That Are Associated With Shorter Telomere Length. *Front Genet, 11*, 337. doi:10.3389/fgene.2020.00337

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., . . . Yeo, G. W. (2016). Robust transcriptome-wide discovery

of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods, 13*(6), 508-514. doi:10.1038/nmeth.3810

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., . . . Kriegstein, A. R. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. *Science, 364*(6441), 685-689. doi:10.1126/science.aav8130

Visconte, V., Avishai, N., Mahfouz, R., Tabarroki, A., Cowen, J., Sharghi-Moshtaghin, R., . . . Tiu, R. V. (2015). Distinct iron architecture in SF3B1-mutant myelodysplastic syndrome patients is linked to an SLC25A37 splice variant with a retained intron. *Leukemia, 29*(1), 188-195. doi:10.1038/leu.2014.170

Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Pawitan, Y., & Rantalainen, M. (2018). Isoform-level gene expression patterns in single-cell RNA-sequencing data. *Bioinformatics, 34*(14), 2392-2400. doi:10.1093/bioinformatics/bty100

Waks, Z., Klein, A. M., & Silver, P. A. (2011). Cell-to-cell variability of alternative RNA splicing. *Mol Syst Biol, 7*, 506. doi:10.1038/msb.2011.32

Wang, E., & Aifantis, I. (2020). RNA Splicing and Cancer. *Trends Cancer, 6*(8), 631-644. doi:10.1016/j.trecan.2020.04.011

Wang, E., Lu, S. X., Pastore, A., Chen, X., Imig, J., Chun-Wei Lee, S., . . . Aifantis, I. (2019). Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. *Cancer Cell, 35*(3), 369-384 e367. doi:10.1016/j.ccell.2019.01.010

Wang, E. T., Cody, N. A., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., . . . Burge, C. B. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell, 150*(4), 710-724. doi:10.1016/j.cell.2012.06.041

Wang, G., Wen, W. X., Mead, A. J., Roy, A., Psaila, B., & Thongjuea, S. (2022). Processing single-cell RNA-seq datasets using SingCellaR. *STAR Protocols, 3*(2), 101266. doi:https://doi.org/10.1016/j.xpro.2022.101266

Wang, K., Yin, C., Du, X., Chen, S., Wang, J., Zhang, L., . . . Cheng, H. (2019). A U2-snRNP-independent role of SF3b in promoting mRNA export. *Proc Natl Acad Sci U S A, 116*(16), 7837-7846. doi:10.1073/pnas.1818835116

Wang, X., Zhang, X., Liu, L., Xiang, M., Wang, W., Sun, X., . . . Liu, X. (2015). Genomic and transcriptomic analysis of the endophytic fungus Pestalotiopsis fici reveals its lifestyle and high potential for synthesis of natural products. *BMC Genomics, 16*, 28. doi:10.1186/s12864-014-1190-9

Warf, M. B., Diegel, J. V., von Hippel, P. H., & Berglund, J. A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proc Natl Acad Sci U S A, 106*(23), 9203-9208. doi:10.1073/pnas.0900342106

Weinstein, L. S., Liu, J., Sakamoto, A., Xie, T., & Chen, M. (2004). Minireview: GNAS: normal and abnormal functions. *Endocrinology, 145*(12), 5459-5464. doi:10.1210/en.2004-0865

Welch, J. D., Hu, Y., & Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Res, 44*(8), e73. doi:10.1093/nar/gkv1525

Wen, W. X., Allen, J., Lai, K. N., Mariapun, S., Hasan, S. N., Ng, P. S., . . . Teo, S. H. (2018). Inherited mutations in BRCA1 and BRCA2 in an unselected multiethnic cohort of Asian patients with breast cancer and healthy controls from Malaysia. *J Med Genet, 55*(2), 97-103. doi:10.1136/jmedgenet-2017-104947

Wen, W. X., & Leong, C. O. (2019). Association of BRCA1- and BRCA2-deficiency with mutation burden, expression of PD-L1/PD-1, immune infiltrates, and T cell-inflamed signature in breast cancer. *PLoS One, 14*(4), e0215381. doi:10.1371/journal.pone.0215381

Wen, W. X., Soo, J. S., Kwan, P. Y., Hong, E., Khang, T. F., Mariapun, S., . . . Teo, S. H. (2016). Germline APOBEC3B deletion is associated with breast cancer risk in an Asian multi-ethnic cohort and with immune cell presentation. *Breast Cancer Res, 18*(1), 56. doi:10.1186/s13058-016-0717-1

Westoby, J., Herrera, M. S., Ferguson-Smith, A. C., & Hemberg, M. (2018). Simulation-based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biol, 19*(1), 191. doi:10.1186/s13059-018-1571-5

Wheeler, E. C., Vora, S., Mayer, D., Kotini, A. G., Olszewska, M., Park, S. S., . . . Papapetrou, E. P. (2022). Integrative RNA-omics Discovers GNAS Alternative Splicing as a Phenotypic Driver of Splicing Factor-Mutant Neoplasms. *Cancer Discov, 12*(3), 836-855. doi:10.1158/2159-8290.CD-21-0508

Wianny, F., Blachere, T., Godet, M., Guillermas, R., Cortay, V., Bourillot, P. Y., . . . Dehay, C. (2016). Epigenetic status of H19/IGF2 and SNRPN imprinted genes in aborted and successfully derived embryonic stem cell lines in non-human primates. *Stem Cell Res, 16*(3), 557-567. doi:10.1016/j.scr.2016.03.002

Wojciechowska, M., Sobczak, K., Kozlowski, P., Sedehizadeh, S., Wojtkowiak-Szlachcic, A., Czubak, K., . . . Brook, J. D. (2018). Quantitative Methods to Monitor RNA Biomarkers in Myotonic Dystrophy. *Sci Rep, 8*(1), 5885. doi:10.1038/s41598-018-24156-x

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol, 19*(1), 15. doi:10.1186/s13059-017-1382-0

Wong, J. J., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W., Huang, Y., . . . Rasko, J. E. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. *Cell, 154*(3), 583-595. doi:10.1016/j.cell.2013.06.052

Wu, L., Zhang, X., Zhao, Z., Wang, L., Li, B., Li, G., . . . Xu, X. (2015). Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. *Gigascience, 4*, 51. doi:10.1186/s13742-015-0091-4

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., . . . Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y), 2*(3), 100141. doi:10.1016/j.xinn.2021.100141

Yao, C., Bora, S. A., Parimon, T., Zaman, T., Friedman, O. A., Palatinus, J. A., . . . Chen, P. (2021). Cell-type-specific immune dysregulation in severely ill COVID-19 patients. *Cell Rep, 34*(13), 108943. doi:10.1016/j.celrep.2021.108943

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., . . . Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Res, 48*(D1), D682-D688. doi:10.1093/nar/gkz966

Yip, B. H., Steeples, V., Repapi, E., Armstrong, R. N., Llorian, M., Roy, S., . . . Boultwood, J. (2017a). The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. *J Clin Invest, 127*(9), 3557. doi:10.1172/JCI96202

Yip, B. H., Steeples, V., Repapi, E., Armstrong, R. N., Llorian, M., Roy, S., . . . Boultwood, J. (2017b). The U2AF1S34F mutation induces lineage-specific splicing alterations in myelodysplastic syndromes. *J Clin Invest, 127*(6), 2206-2221. doi:10.1172/JCI91363

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., . . . Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature, 478*(7367), 64-69. doi:10.1038/nature10496

Yoshimi, A., Lin, K. T., Wiseman, D. H., Rahman, M. A., Pastore, A., Wang, B., . . . Abdel-Wahab, O. (2019). Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. *Nature, 574*(7777), 273-277. doi:10.1038/s41586-019-1618-0

Zhang, J., Ali, A. M., Lieu, Y. K., Liu, Z., Gao, J., Rabadan, R., . . . Manley, J. L. (2019). Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Mol Cell, 76*(1), 82-95 e87. doi:10.1016/j.molcel.2019.07.017

Zhang, J., Kuo, C. C., & Chen, L. (2015). WemIQ: an accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics, 31*(6), 878-885. doi:10.1093/bioinformatics/btu757

Zhang, Z., Hernandez, K., Savage, J., Li, S., Miller, D., Agrawal, S., . . . Grossman, R. L. (2021). Uniform genomic data analysis in the NCI Genomic Data Commons. *Nat Commun, 12*(1), 1226. doi:10.1038/s41467-021-21254-9

Zhao, Y. G., Zhao, H., Miao, L., Wang, L., Sun, F., & Zhang, H. (2012). The p53-induced gene Ei24 is an essential component of the basal autophagy pathway. *J Biol Chem, 287*(50), 42053-42063. doi:10.1074/jbc.M112.415968

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., . . . Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun, 8*, 14049. doi:10.1038/ncomms14049

9 Epilogue

This thesis was submitted on 6th October 2022 and was defended on 28th November 2022. Many thanks to Professor Jim Hughes and Dr. Jyoti Nangalia for being on the examination panel. I hope they enjoyed the viva session as much as I did. Following the successful defence, I started my new role as a Postdoctoral Fellow at the Centre for Genomics Research at AstraZeneca, Cambridge on 19th December 2022. I was fortunate to find a portfolio that enabled me to continue my research on blood cancers. In a world where evidence is being challenged as opinion by the less-informed public, science may be the only sanctuary left for truth. In the words of Maxine Roby, “blood leaves DNA, and DNA leaves no doubt.”

March 2023