UNIVERSITY OF CALIFORNIA
SANTA CRUZ

# Identifying Splice Variants that Contribute to Disease: The Trials, Tribulations and Triumphs of Finding a Needle in a Haystack

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

**Megan Durham**

September 2023

The dissertation of Megan Durham is approved:

Professor Angela Brooks, Chair

Professor Lindsay Hinck

Professor Melissa Jurica

Peter Biehl
Dean of Graduate Studies

## Table of Contents

**List of Figures**

## List of Tables

## Abstract

### Identifying Splice Variants that Contribute to Disease: The Trials, Tribulations and Triumphs of Finding a Needle in a Haystack

**Megan Durham**

While it's known that the dysregulation of splicing contributes to disease, identifying specific splicing alterations is a bottleneck for current research. On an individual scale, mutations in signaling cues disrupt canonical splicing and force the production of an aberrant splicing event. On a global scale, splicing factor mutations result in widespread splicing dysregulation and create a subset of isoforms that contribute to disease progression. In both cases, identifying specific splicing aberrations that drive disease is important in developing targeted therapies, and can greatly improve the lives of patients bearing these mutations. To study novel deleterious splicing events, scientists need the appropriate tools to identify and functionally characterize these isoforms. In chapter 2 of this dissertation, I found that using CRISPR/Cas9 to express aberrant exon skipping events from their endogenous promoter did not produce enough aberrant isoform to confirm the oncogenic potential in two functional assays. Because exon skipping events, alone, were not sufficient to confer an oncogenic phenotype, I investigated the level of a known oncogenic isoform, *MET* exon 14 skipping (*METΔ14*), expressed in lung adenocarcinoma primary samples in chapter 3. I found that *METΔ14* is overexpressed in an allele-specific manner, and *METΔ14* overexpression is required to activate the receptor. In chapter 4, through a collaboration with Dr. Esther Obeng's lab, I performed long read sequencing in cell lines expressing mutant *SF3B1*, a common splicing factor mutation in Myelodysplastic Syndrome (MDS). This collaboration uncovered previously unannotated isoforms with potential implication in MDS. Overall, the work in this dissertation outlines the trials and tribulations of using the correct tools to functionally characterize aberrant splicing events, and describes triumphs of discovering how *METΔ14* is expressed in lung cancer tumors and isoforms implicated in MDS-associated anemia requiring further characterization.

## Dedication

To the village that believed in me.

## Acknowledgements

This journey has been challenging, and without the help of my advocates, family, and friends, I’m not sure how I would have made it this far. Thank you to Angela, who showed me that being kind is the best approach. Thank you to my thesis committee, Lindsay and Melissa, for your countless pep talks and encouragement. You always knew what to say to me when I needed to hear it. A special thank you to Lindsay, who was a constant source of support throughout these last few years. A thank you to my mom and dad, it was a treat to have my parents in the same town where I attended graduate school. You are my biggest cheerleaders, and without you none of this would have been possible. I am blessed to have an incredible support system, and I am incredibly grateful for all the friendships I’ve built in graduate school. Thank you to Laura, who is my adventure buddy and one of the kindest people I have ever met. I’m so excited for all our future outdoor adventures. Thank you to Hannah, who is always down to commiserate and/or celebrate over drinks, and would do anything for her friends. Thank you to my lab mates Cindy and Colette, it was beautiful how we helped each other and had each other's backs in graduate school. I’m so excited to see all the wonderful things you will accomplish. I wouldn’t have made it this far without Eva and Namrita, whose scientific expertise is matched by their kind and tenacious nature. I will always cherish what I learned from you. A thank you to Dori and Celine, who were such amazing and supportive friends through it all. Finally, a huge and incredibly warm thank you to Mikey, who stood by my side through thick and thin. You are my rock, my love, my mishi, and I can’t imagine making this journey without you.

# **Chapter 1**: **Introduction**

## *Pre-mRNA splicing is an essential step in gene expression*

Every cell in our body contains the same genetic textbook providing the instructions for life: DNA. While this genetic textbook is a fundamental ingredient in each cell, how DNA is used differs in different cell types (Tung et al. 2020; Uhlén et al. 2015). Several factors determine how genes are regulated from DNA, and this regulation is necessary not only for cellular diversity, but also in response to developmental and environmental cues, and cellular homeostasis. While DNA is essential, it takes a passive role as a genetic blueprint to create messenger RNA (mRNA) which will be used as a template for eventual protein synthesis. Gene expression begins with transcription, where pre-mRNA is copied from DNA, followed by translation, which generates protein. Between transcription and translation, pre-mRNA splicing is an essential processing step that prepares pre-mRNA for protein synthesis.

The long, unprocessed pre-mRNA sequence is a combination of introns, or sequences not destined to code for protein, and exons, functionally relevant sequences necessary for protein production. During pre-mRNA splicing, introns are removed from the sequence and the exons are ligated together to make a mature mRNA transcript. This process is carried out by the spliceosome, a complex molecular machine composed of five small nuclear ribonucleoproteins (snRNPs) and nearly 300 associated proteins (Rappsilber et al. 2002). This strict coordination of splicing is tightly regulated in our cells, and thus it is no surprise that its dysregulation is known to contribute to disease (Kelemen et al. 2013; Singh and Cooper 2012).

### *Trans-acting factors recognize cis-regulatory elements to coordinate splicing*

The spliceosome is composed of *trans*-acting factors, which recognize nucleic acid and bind sequences called *cis*-acting elements, or important signaling cues which define exons during splicing. These *cis*-acting elements define exons through splice sites, which are di-nucleotides pairs adjacent to each exon. The 5’ splice site is the upstream splice site, containing a GU, while the 3’ splice site is downstream, containing an AG. The polypyrimidine tract and branchpoint sequence (BPS) are key intronic motifs in 3’ splice site recognition and are located upstream of the 3’ splice site. Recognition of these sequences by different components of the spliceosome is essential for the first step of pre-mRNA splicing (Will and Lührmann 2011). In addition to these key cis-acting elements, auxiliary sequences called enhancers and silencers are present in both introns and exons which help promote or discourage recognition of certain splice sites. In introns, these are intronic splicing enhancers and silencers (ISEs and ISSs), and in exons, exonic splicing enhancers and silencers (ESEs and ESSs). ESEs in particular are present in most exons and understood to play dominant roles in constitutive splicing, while silencers are important in the selection of weaker splice sites (Liu et al. 1998; Schaal and Maniatis 1999; Wang et al. 2015).

These *cis*-acting elements are recognized by *trans*-acting factors. The core of the spliceosome is composed of five snRNPs, known as U1, U2, U4, U5 and U6. U1 snRNP binds the 5’ splice site (Mount et al. 1983; Zhuang and Weiner 1986), followed by 3’ splice site recognition which requires the BPS, polypyrimidine tract, and the 3’ splice site. The U2 snRNP binds the branch point sequence The U2 snRNP binds the branch point sequence (Black et al. 1985; Wu and Manley 1989), which is feasible through spliceosome factor 3b1 (SF3B1), the largest subunit in the SF3B complex and component of the U2 snRNP (Gozani et al. 1996; Krämer 1996). U2AF2 binds the polypyrimidine tract (Zamore et al. 1992), and U2AF1 binds the 3' splice site (Wu et al. 1999). Finally, U4/U5/U6 snRNPs form the tri-snRNP, which is recruited to assembled splicing machinery and is responsible for spliceosome rearrangements necessary for intron excision and exon ligation (Will and Lührmann 2011). In addition to the core spliceosome machinery, auxiliary silencer and enhancer sequences are recognized by factors like heterogeneous nuclear ribonucleoproteins (hnRNPs) and serine/arginine (SR) proteins, respectively (Wu and Maniatis 1993). In order for splicing to function properly, all these *cis*-acting and *trans*-acting factors must perfectly recognize each other (Figure 1.1). This is required to produce necessary isoforms in different cell types, in response to different stimuli, or at different developmental stages (Steward et al. 2022; Liu et al. 2022; Fiszbein et al. 2016; Wang et al. 2008; Dillman et al. 2013; Buljan et al. 2012)

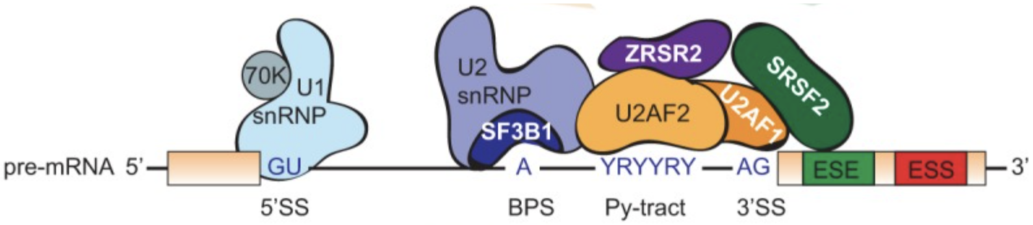


**Figure 1.1: Spliceosomal *trans*-acting factors binding corresponding *cis*-acting elements in early spliceosomal assembly.**

During exon definition, U1 snRNP binds the 5' splice site (5'SS), U2 snRNP binds the branchpoint sequence (BPS) via SF3B1, U2AF2 binds the polypyrimidine tract (Py-tract), and U2AF1 binds the 3' splice site (3'SS) dinucleotide. Auxiliary proteins bind Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS) within the exon to promote or discourage splice site usage. Not shown: Intronic Splicing Enhancers (ISE) and Intronic Splicing Silencers (ISS). Figure from Anczuków and Krainer 2016 © The RNA Society

### *Alternative splicing increases protein diversity*

While constitutive splicing removes introns and ligates exons in the order in which they appear in a gene, alternative splicing deviates from this preferred sequence and uses different, weaker splice sites to create different isoforms from one genetic location. The resulting isoforms are determined by the additive ability of *trans*-acting factors binding *cis*-acting elements in the pre-mRNA (Goldammer et al. 2018; Wang et al. 2015). Genome wide studies suggest 92-95% of human genes undergo alternative splicing (Wang et al. 2008; Pan et al. 2008), and increasing evidence throughout the years has shown alternative splicing vastly increases the diversity of the proteome by creating multiple proteins from one genomic origin (Black 2003; Nilsen and Graveley 2010).

This is illustrated through the comparison of Ensembl genes and transcripts between the simple eukaryotic organism, *C. elegans*, and the complex organism of humans. This analysis revealed a relatively similar gene count between eukaryotes, but a vastly different number of total transcripts as the complexity of organisms increased. This suggests that an increasing number of mRNA transcripts are associated with organismal complexity (Figure 1.2). This evidence is further supported through an analysis of transcripts in human tissues (GTEx) which identified ~7 transcripts per protein coding gene (Tung et al. 2020)

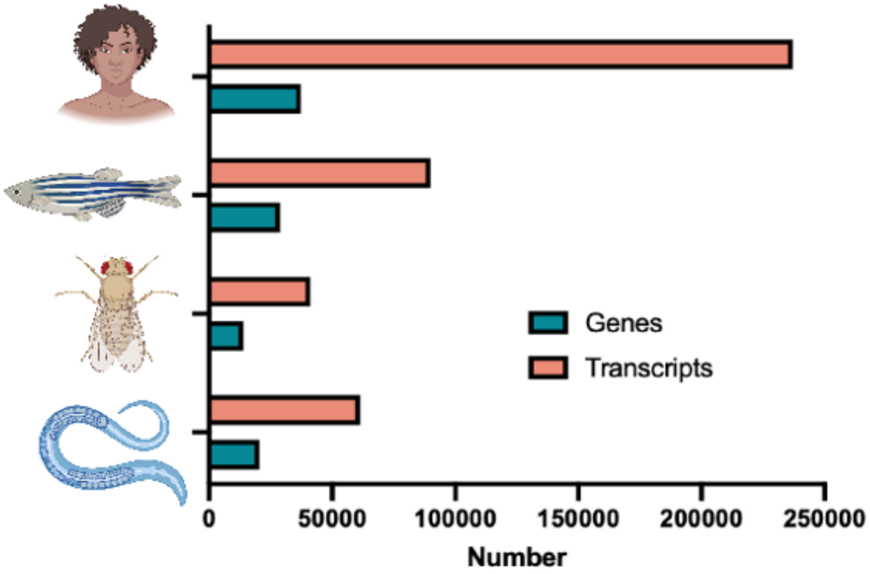


**Figure 1.2: Organismal complexity correlates with the number of transcripts, not the number of genes.**

The number of genes and transcripts from Ensembl for humans, fish, flies and C. elegans. Figure created by Angela Brooks, using data from Ensembl in Dec 2021.

Alternative splicing patterns can be classified into five different classes: exon skipping, intron retention, mutually exclusive exons, alternative 5’ splice site choice, and alternative 3’ splice site choice (Figure 1.3). Exon skipping removes an entire exon from an mRNA transcript, and is the most common alternative splicing event in humans totaling about 42% compared to other splicing events (Barbosa-Morais et al. 2012; Kim et al. 2007) (Figure 1.4 - far right column). These events will create isoforms lacking that exon, referred to as “Δ” followed by the exon number, along with the wild type isoform (ex: *METΔ14*). Alternative 5’ and 3’ splice site choice is the second most abundant splicing decision in humans, occurring about 25% of the time for both alternative 5’ and 3’ splice sites, while intron retention events occur less than 10% of the time (Figure 1.4). These splicing events are key in gene regulation, however, mutations in splicing factors or exon-defining cues may disrupt alternative splicing into something sinister (Kong-Beltran et al. 2006; Lu et al. 2017; Turpin et al. 2016; Smith et al. 2020).

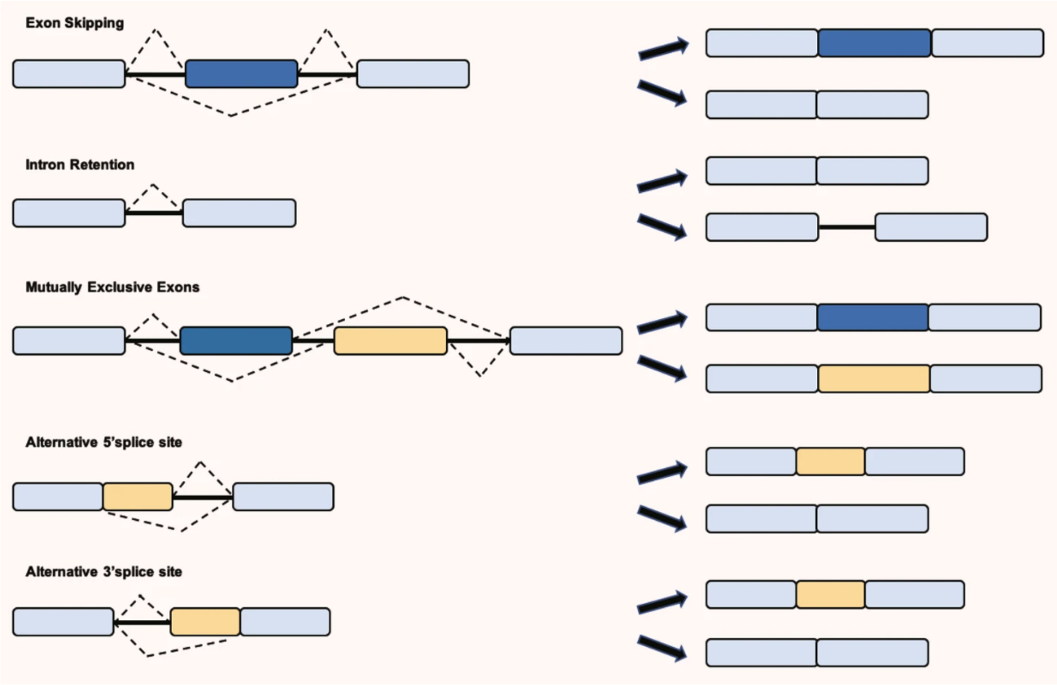


**Figure 1.3: Five major classes of Alternative Splicing events.**

Exon skipping skips the entire exon, and intron retention leads to the inclusion of the entire intron in the mature mRNA. In mutually exclusive splicing events, the spliceosome decides to differentially include one exon over the other, but never both. For alternative 5' or 3' splice site choice, weaker splice sites within the intron are used increasing the length of the exon. Figure created by Zhang et al. 2021 © Springer Nature Limited.

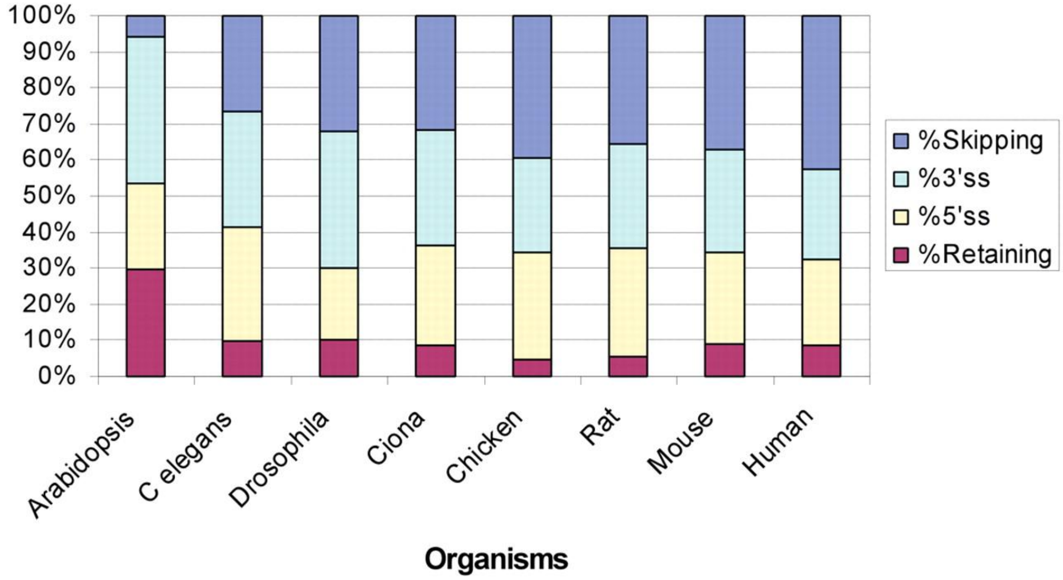


**Figure 1.4: Quantification of alternative Splicing in Eukaryotes.**

Distributions of four classes of alternative splicing events in a variety of eukaryotes determined from the gene-oriented clusters of mRNAs from the UniGene database. Human distributions of splicing events shown at the right-most column. Figure taken from Kim et al. 2007 © 2006 Eddo Kim, Alon Magen, and Gil Ast.

*Aberrant splicing: The dysregulation of splicing contributes to disease*

While alternative splicing is an elegant dance of splicing factors binding sequences in the pre-mRNA to dictate splicing fate, aberrant splicing twists destiny to create deleterious isoforms with grave consequences. Mutations in splicing machinery or nucleic acid sequence prevent *trans*-acting factors from recognizing their *cis-acting* elements. This can dysregulate splicing globally or generate specific deleterious isoforms which are a common occurrence in neurological disease, myelodysplastic syndromes, and many types of cancer (Jiang et al. 2023; Anczuków and Krainer 2016; Sveen et al. 2015; Feng and Xie 2013; Malcovati et al. 2011; Papaemmanuil et al. 2011)

These splicing aberrations result in two consequences: overall transcript degradation via premature stop codons and subsequent Nonsense Mediated Decay (NMD), or creation of a functional aberrant protein. While the generation of premature termination codons by alternative splicing and NMD is a known mechanism for gene regulation (Weischenfeldt et al. 2012), dysregulated protein degradation due to splicing mutations can lead to aberrant down regulation of canonical genes and isoforms (Darman et al. 2015; Alsafadi et al. 2016; Obeng et al. 2016). If these splicing changes do not disrupt the reading frame, this can create a functional protein. The consequences of these truncated proteins include heightened transformation ability, atypical cell development, and other deleterious effects (Kong-Beltran et al. 2006; Turpin et al. 2016; Clough et al. 2022; Tam et al. bioRxiv ).

### *Aberrant splicing: Mutations in trans-acting splicing factors*

The dysregulation of splicing factors can result in global splicing alterations, as this affects many pre-mRNAs that require splicing. Typically, this dysregulation occurs through two different mechanisms: aberrant splicing factor expression and splicing factor mutations. Aberrant expression of proteins associated with core spliceosomal machinery or splicing enhancers and silencers are known to contribute to human disease like cancer (Li et al. 2023; Sveen et al. 2015). For example, overexpression of U2 spliceosomal component SF3B1, which ultimately contributes to 3' splice site recognition, is overexpressed and implicated in the aggressiveness of hepatocellular carcinoma (López-Cánovas et al. 2021). Overexpression of SRSF3, a positive regulator of splice site choice, favored exon 9 inclusion in pyruvate kinase, taking a positive role in cancer-specific energy metabolism (Kuranaga et al. 2018). Additionally, hnRNPH, a suppressor of splice site choice, is upregulated in malignant gliomas and is known to contribute to their progression by promoting splicing of a malignant, ligand-independent tyrosine kinase receptor isoform (Lefave et al. 2011).

While expression changes disrupt splicing through tipping the scale of splicing factor concentrations, mutations in splicing factors are also known to disrupt splicing through their ability to interact with their respective pre-mRNA binding sites. Several high-throughput sequencing studies and reviews highlight that splicing factor mutations are frequent in a variety of cancers and hematological disorders (Kandoth et al. 2013; Watson et al. 2013; Yoshida et al. 2011; Chen et al. 2021). For example, U2AF1, which directly recognizes the 3' splice site AG (Wu et al. 1999), is frequently mutated in its zinc finger domain at position 34 (Brooks et al. 2014; Yoshida et al. 2011), impairing its ability to interact with the 3' splice site (Yoshida et al. 2015). Overall *U2AF1* mutations force global aberrant 3' splice site selection and impact the splicing of several known cancer genes (Brooks et al. 2014; Esfahani et al. 2019). Another example are hotspot mutations in *SF3B1*, which are frequent in chronic lymphocytic leukemia, uveal melanoma, breast cancer, and myelodysplastic syndromes (MDS) (Wang et al. 2011; Quesada et al. 2011; Rossi et al. 2011; Landau et al. 2015; Martin et al. 2013; Harbour et al. 2013; Furney et al. 2013; Maguire et al. 2015; Pereira et al. 2016; Fu et al. 2017; Garcia-Manero et al. 2020; Papaemmanuil et al. 2011)*.* SF3B1 is a component of the U2 snRNP and plays a critical role in recognizing the branch point sequence prior to 3' splice site selection (Krämer 1996). However, mutations in *SF3B1* are known to promote cryptic 3' splice site selection through the selection of a non-canonical Adenosine-rich branch point sequence (Darman et al. 2015; Alsafadi et al. 2016) (Figure 1.5). For MDS, this cryptic splice site choice has resulted in several known splicing consequences in genes related to red blood cell maturation, which have been implicated in MDS-associated anemia (Clough et al. 2022; Tam et al. ; Lieu et al. 2022; Dolatshad et al. 2016). As splicing factor dysregulation compels splicing alterations on a global scale, many specific isoform consequences as they relate to disease remain unknown. As a result, a more thorough investigation of aberrantly produced isoforms is necessary, with the goal of finding targetable proteins with existing therapies.

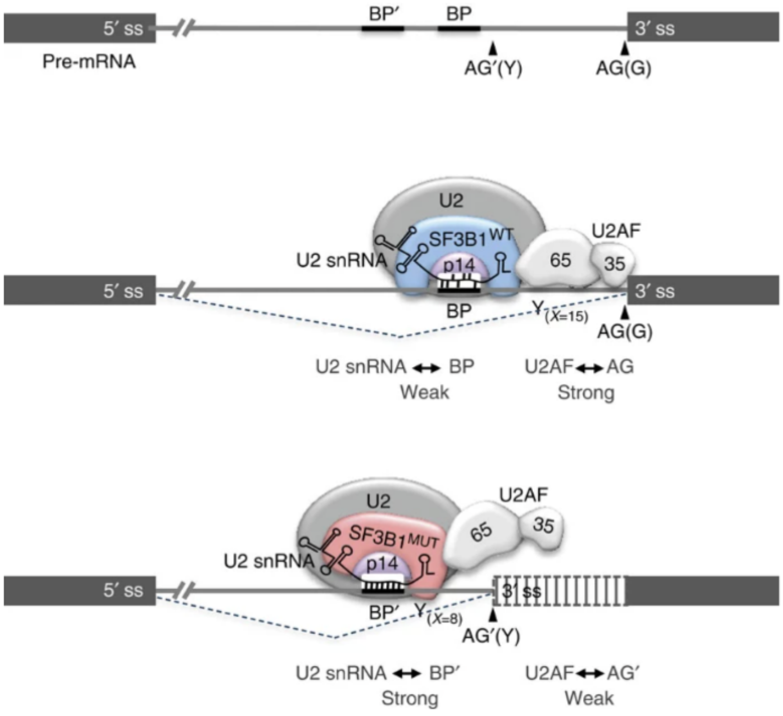


**Figure 1.5: Mutant *SF3B1* leads to cryptic 3' splice site choice through alternate branch point selection.**

In wild type SF3B1, U2AF binds the canonical 3' splice site (AG(G)) and the U2 snRNP recognizes the canonical branch point (BP) through SF3B1. In mutant

SF3B1, this mutation is anticipated to induce a conformational change that causes U2 snRNP to recognize a different branch point (BP') ultimately leading to an aberrant upstream 3' splice site choice (AG'(Y)). Figure from Alsafadi et al. 2016 © Springer Nature Limited.

*Aberrant splicing: Mutations in cis-acting sequence elements*

Mutations in cis-acting elements disrupt exon recognition in pre-mRNA, which forces the production of an aberrantly spliced transcript. Nearly one third of all disease-causing mutations are estimated to occur in these cis-acting sequence cues, and therefore have the potential to disrupt splicing (Singh and Cooper 2012). These mutations can occur in either introns or exons, and may disrupt existing splice sites, create new splice sites, or activate cryptic splice sites. They can also influence binding of splicing enhancers and repressors, and dysregulate the inclusion of certain splicing events (Anna and Monika 2018).

Commonly, these mutations disrupt canonical splice sites, which often result in exon skipping events (Anna and Monika 2018; Kong-Beltran et al. 2006; Smith et al. 2020; Lu et al. 2017; Frampton et al. 2015). One of the most well characterized aberrant splicing events from cis-acting mutations is *MET* exon 14 skipping (*METΔ14*). MET is a receptor in the Ras-MAPK pathway and plays an essential role in a number of critical cellular processes such as cell proliferation, survival, motility, and morphogenesis (Birchmeier et al. 2003; Organ and Tsao 2011). This receptor is activated through binding of its ligand, HGF, and this signal is terminated through MET's degradation. The negative regulatory region of *MET* is encoded by exon 14, and contains a binding site for the E3-ubiquitin ligase, Cbl, at position Y1003 (Peschard et al. 2001). While the inclusion of exon 14 is critical for MET's regulation typically exon 14 mutations force its exclusion from the mature mRNA (*METΔ14*) (Figure 1.6), resulting in a protein with an extended life span in the membrane and prolonged proliferative signaling that drives cancer (Kong-Beltran et al. 2006; Lu et al. 2017; Ma et al. 2005; Ma et al. 2003). *METΔ14* mutations are observed in 2.8% of cases of lung adenocarcinomas (Lu et al. 2017), at a relatively high frequency considering *MET* is estimated to be mutated in 4-6% of lung adenocarcinomas (Caso et al. 2020). While *METΔ14* is most common in lung adenocarcinomas, it has also been identified in other lung neoplasms and brain gliomas (Frampton et al. 2015). Interestingly, a recently identified mutation MET Y1003* replaces a tyrosine with a premature termination codon. While anticipated to reduce overall MET levels through NMD, this mutation produces *METΔ14.* Through computational modeling it is anticipated that this mutation disrupts the binding sequences of several splicing enhancers, disrupting recognition of exon 14 and forcing its exclusion (Cancer Genome Atlas Research Network ...).

Another well characterized aberrantly spliced oncogene created through splice site mutations is *ERBB2* exon 16 skipping (*ERBB2Δ16*) (Turpin et al. 2016; Smith et al. 2020; Shi et al. 2020). ERBB2 is a receptor in the Ras-MAPK pathway, and activating mutations in this receptor are found in several different cancer types (Oh and Bang 2020). For *ERBB2Δ16*, splice site mutations force exon 16 exclusion, which generates an uneven number of cysteines in each monomer. This forces monomers to form disulfide bonds between two monomers, stabilizing the *ERBB2Δ16* tyrosine kinase signaling through constitutive dimerization (Siegel et al. 1999). Like *METΔ14*, this constitutively active variant is implicated in different cancer types, such as breast cancer, non-small cell lung cancer, rectal cancer, and ovarian cancer (Turpin et al. 2016; Smith et al. 2020; Shi et al. 2020). Additionally, while *METΔ14* is a relatively common splicing aberration, *ERBB2Δ16* is rare, present 0.046% of some cohorts (Shi et al. 2020). Being able to identify these aberrant splicing events is important because this can lead to better therapies that will expand treatment options for patients.

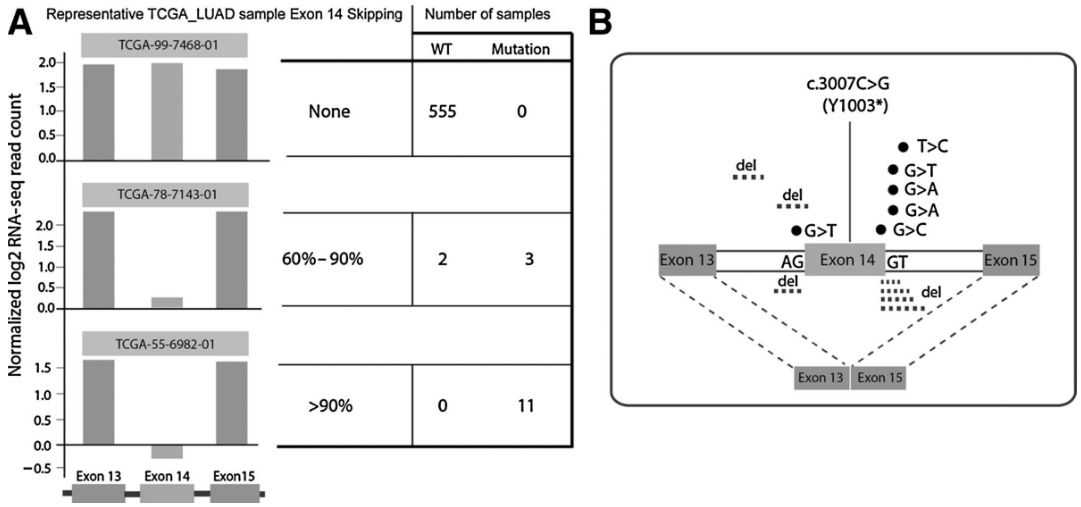


**Figure 1.6: *MET* exon 14 skipping due to somatic mutations at splicing cues.**

A) Lung Adenocarcinoma (LUAD) samples with varied *MET* exon 14 skipping and number of samples with wild type and mutant MET in each group. B) Indels seen near exon 14 in LUAD samples with mRNA evidence of exon 14 skipping. Figure from Lu et al. 2017 © American Association for Cancer Research.

### CRISPR/Cas9 as a tool to study aberrant exon skipping events

Typically, single aberrantly spliced variants are studied through their overexpression (Smith et al. 2020; Alajati et al. 2013; Suzawa et al. 2019). However, researchers express these splicing variants on the longest coding sequence, and the longest coding sequence may not be the appropriate context for these splicing events. In fact, about 50% of transcripts found to be dominantly expressed in tissues are not the longest coding sequence, and therefore are not the ‘canonical’ transcript as denoted by UniProt (Gonzàlez-Porta et al. 2013). Additionally, modest doses of oncogenes can drive clonal outgrowth and modulate drug response (Bielski et al. 2018), suggesting overexpression is not necessary to induce a phenotype.

Because of this, CRISPR-based expression is gaining traction to study these aberrant splicing events. Ever since the application of clustered regularly interspaced short palindromic repeats, or CRISPR, in eukaryotic cells, this tool has revolutionized genomic editing for eukaryotic model systems (Mali et al. 2013; Cong et al. 2013). This system works by programming a single guide RNA (sgRNA) to target a specific region in the genome, which guides a protein called Cas9 to induce a double stranded break at that location (Jinek et al. 2012). If no homology repair template is added, the DNA is restored through the indel-forming repair process of non-homologous end joining (NHEJ) (Mali et al. 2013; Cong et al. 2013). While this is commonly used for gene knock-out through targeting the N-terminal coding exons (Wang et al. 2014; Wang et al. 2015) this system can also be applied to force aberrant exon skipping events. Aberrant exon skipping events remove entire regulatory regions and functional domains, and in-frame exon skipping events can create gain-of-function oncogenic drivers (Kong-Beltran et al. 2006; Lu et al. 2017; Turpin et al. 2016; Smith et al. 2020). To force an exon skipping event using CRISPR, an sgRNA is programmed to target either the 5’ or 3’ splice site, and Cas9 induces a double stranded break at this location. The error-prone repair process of NHEJ introduces indels at splice sites, rendering them unrecognizable to the spliceosome and forcing an aberrant exon skipping event (Figure 1.7). This has been used in several studies of *METΔ14* (Lu et al. 2017; Togashi et al. 2015; Wang et al. 2022), which uncouples the transforming ability of *MET* overexpression with the catalytic ability of *METΔ14* (Lu et al. 2017). Because CRISPR-based expression disrupts splice site signals which commonly produce exon skipping events (Anna and Monika 2018; Kong-Beltran et al. 2006; Smith et al. 2020; Lu et al. 2017; Frampton et al. 2015), and uses the endogenous promoter to express the variant, this method is ideal to study exon skipping in model systems.

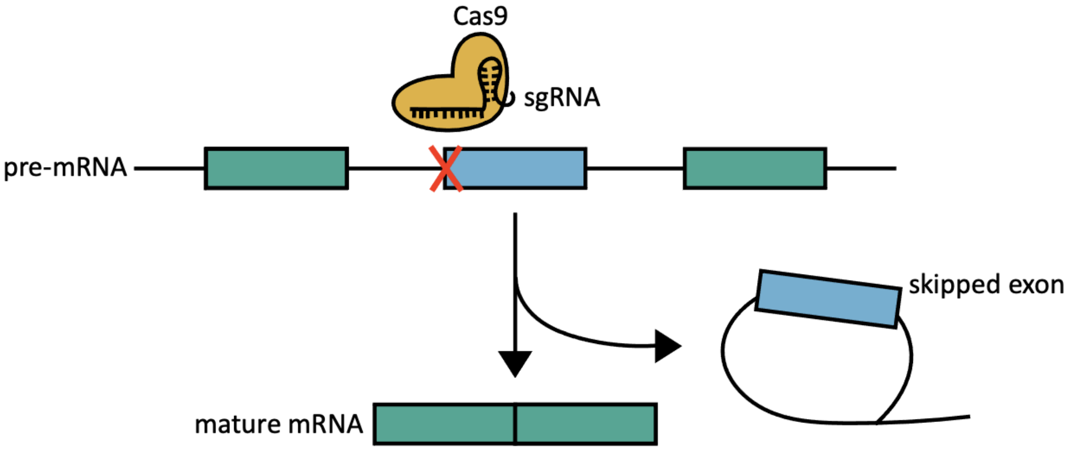


**Figure 1.7: Using CRISPR/Cas9 to force exon skipping events**

A splice site targeting sgRNA guides Cas9 to create indels at splice sites. This renders splicing signals unrecognizable to the spliceosome, which forces an exon skipping event in the mature mRNA.

*Necessity of identifying new splicing events*

Due to the dedication of scientists and medical doctors whose mission is to tackle splicing-induced disease, there are several treatment options available which greatly improve the lives of those afflicted by these conditions. One approach to correct aberrant splicing events is gene therapy approaches, such as Antisense Oligonucleotides (ASOs). ASOs are short synthetic DNA molecules complementary to the pre-mRNA sequence to alter splicing. This can occur through binding to splice sites or enhancer and silencer elements in order to prevent spliceosome interaction with these sites, and produce the desired splicing outcome (Suñé-Pou et al. 2020). In December 2016, FDA approved Spinraza$^{TM}$ (nusinersen) for treatment of Spinal Muscular Atrophy (SMA) (Lorson et al. 1999; Monani et al. 1999). Affecting 1 in 10,000 infants (Pearn 1978), SMA is marked by motor neuron depletion in the spinal cord, leading to early death without aggressive intervention. This disease is caused by inefficient SMA levels, which can be bolstered through selective splicing in of exon 7 of SMN2 (Porensky and Burghes 2013). Expanding from this use case, ASOs have promise for fast personalized treatments of rare diseases. As an example, a 6 year old girl presenting with progressed cerebral and cerebellar atrophy exhibited missplicing of *MFSD8* due to a cryptic splice site in intron 6. Scientists were able to develop an ASO and obtain FDA approval in less than a year to give this patient a chance at life (Kim et al. 2019). This example highlights how versatile and customizable these therapies can be to target splicing aberrations.

In addition to splicing modulators, small molecule inhibitors target and prevent disease progression in patients expressing the protein products of aberrant splicing events. For instance, *METΔ14* has several treatments currently used in clinical practice to target cancer. Crizotinib, a general receptor tyrosine kinase inhibitor, has been shown to shrink tumor volume in *METΔ14* driven tumors (Lu et al. 2017). Additionally, several drugs are currently FDA approved to target tumors expressing *METΔ14*, including capmatinib which was approved by the FDA in 2020, and tepotinib, approved in 2021 (Mathieu et al. 2022).

These cases illustrate the importance of identifying splicing mutations to expand treatment for patients with no other options. With the advent of next generation sequencing we can determine gene expression and subtle mutations due to its depth and breadth. Expanding on this tool kit, long read sequencing provides isoform-level information, giving researchers power to identify deleterious splicing events and isoforms not afforded by short read sequencing. Due to this increased power and understanding of aberrant splicing in disease, can greatly improve the lives of patients and their loved ones afflicted with these diseases.

## References

Alajati, Abdullah, Nina Sausgruber, Nicola Aceto, Stephan Duss, Sophie Sarret, Hans Voshol, Debora Bonenfant, and Mohamed Bentires-Alj. 2013. “Mammary Tumor Formation and Metastasis Evoked by a HER2 Splice Variant.” Cancer Research 73 (17): 5320–27.

Alsafadi, Samar, Alexandre Houy, Aude Battistella, Tatiana Popova, Michel Wassef, Emilie Henry, Franck Tirode, et al. 2016. “Cancer-Associated SF3B1 Mutations Affect Alternative Splicing by Promoting Alternative Branchpoint Usage.” Nature Communications 7 (1): 1–12.

Anczuków, Olga, and Adrian R. Krainer. 2016. “Splicing-Factor Alterations in Cancers.” RNA 22 (9): 1285–1301.

Anna, Abramowicz, and Gos Monika. 2018. “Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation.” Journal of Applied Genetics 59 (3): 253–68.

Barbosa-Morais, Nuno L., Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, et al. 2012. “The Evolutionary Landscape of Alternative Splicing in Vertebrate Species.” Science 338 (6114): 1587–93.

Birchmeier, Carmen, Walter Birchmeier, Ermanno Gherardi, and George F. Vande Woude. 2003. “Met, Metastasis, Motility and More.” Nature Reviews. Molecular Cell Biology 4 (12): 915–25.

Black, D. L., B. Chabot, and J. A. Steitz. 1985. “U2 as Well as U1 Small Nuclear Ribonucleoproteins Are Involved in Premessenger RNA Splicing.” Cell 42 (3): 737–50.

Black, Douglas L. 2003. “Mechanisms of Alternative Pre-Messenger RNA Splicing.” Annual Review of Biochemistry 72 (February): 291–336.

Brooks, Angela N., Peter S. Choi, Luc de Waal, Tanaz Sharifnia, Marcin Imielinski, Gordon Saksena, Chandra Sekhar Pedamallu, et al. 2014. “A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events.” PloS One 9 (1): e87361.

Buljan, Marija, Guilhem Chalancon, Sebastian Eustermann, Gunter P. Wagner, Monika Fuxreiter, Alex Bateman, and M. Madan Babu. 2012. “Tissue-Specific Splicing of Disordered Segments That Embed Binding Motifs Rewires Protein Interaction Networks.” Molecular Cell 46 (6): 871–83.

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Profiling of Lung Adenocarcinoma.” Nature 511 (7511): 543–50.

Caso, Raul, Francisco Sanchez-Vega, Kay See Tan, Brooke Mastrogiacomo, Jian Zhou, Gregory D. Jones, Bastien Nguyen, et al. 2020. “The Underlying Tumor Genomics of Predominant Histologic Subtypes in Lung Adenocarcinoma.” Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 15 (12): 1844–56.

Chen, Sisi, Salima Benbarche, and Omar Abdel-Wahab. 2021. “Splicing Factor Mutations in Hematologic Malignancies.” Blood 138 (8): 599–612.

Clough, Courtnee A., Joseph Pangallo, Martina Sarchi, Janine O. Ilagan, Khrystyna North, Rochelle Bergantinos, Massiel C. Stolla, et al. 2022. “Coordinated Missplicing of TMEM14C and ABCB7 Causes Ring Sideroblast Formation in SF3B1-Mutant Myelodysplastic Syndrome.” Blood 139 (13): 2038–49.

Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” Science 339 (6121): 819–23.

Darman, Rachel B., Michael Seiler, Anant A. Agrawal, Kian H. Lim, Shouyong Peng, Daniel Aird, Suzanna L. Bailey, et al. 2015. “Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3’ Splice Site Selection through Use of a Different Branch Point.” Cell Reports 13 (5): 1033–45.

Dillman, Allissa A., David N. Hauser, J. Raphael Gibbs, Michael A. Nalls, Melissa K. McCoy, Iakov N. Rudenko, Dagmar Galter, and Mark R. Cookson. 2013. “mRNA Expression, Splicing and Editing in the Embryonic and Adult Mouse Cerebral Cortex.” Nature Neuroscience 16 (4): 499–506.

Dolatshad, H., A. Pellagatti, F. G. Liberante, M. Llorian, E. Repapi, V. Steeples, S. Roy, et al. 2016. “Cryptic Splicing Events in the Iron Transporter ABCB7 and Other Key Target Genes in SF3B1-Mutant Myelodysplastic Syndromes.” Leukemia 30 (12): 2322–31.

Esfahani, Mohammad S., Luke J. Lee, Young-Jun Jeon, Ryan A. Flynn, Henning Stehr, Angela B. Hui, Noriko Ishisoko, et al. 2019. “Functional Significance of U2AF1 S34F Mutations in Lung Adenocarcinomas.” Nature Communications 10 (1): 1–13.

Feng, Dairong, and Jiuyong Xie. 2013. “Aberrant Splicing in Neurological Diseases.” Wiley Interdisciplinary Reviews. RNA 4 (6): 631–49.

Fiszbein, Ana, Luciana E. Giono, Ana Quaglino, Bruno G. Berardino, Lorena Sigaut, Catalina von Bilderling, Ignacio E. Schor, et al. 2016. “Alternative Splicing of G9a Regulates Neuronal Differentiation.” Cell Reports 14 (12): 2797–2808.

Frampton, Garrett M., Siraj M. Ali, Mark Rosenzweig, Juliann Chmielecki, Xinyuan Lu, Todd M. Bauer, Mikhail Akimov, et al. 2015. “Activation of MET via Diverse Exon 14 Splicing Alterations Occurs in Multiple Tumor Types and Confers Clinical Sensitivity to MET Inhibitors.” Cancer Discovery 5 (8): 850–59.

Fu, Xing, Ming Tian, Jia Gu, Teng Cheng, Ding Ma, Ling Feng, and Xing Xin. 2017. “SF3B1 Mutation Is a Poor Prognostic Indicator in Luminal B and Progesterone Receptor-Negative Breast Cancer Patients.” Oncotarget 8 (70): 115018–27.

Furney, Simon J., Malin Pedersen, David Gentien, Amaury G. Dumont, Audrey Rapinat, Laurence Desjardins, Samra Turajlic, et al. 2013. “SF3B1 Mutations Are Associated with Alternative Splicing in Uveal Melanoma.” Cancer Discovery 3 (10): 1122–29.

Goldammer, Gesine, Alexander Neumann, Miriam Strauch, Michaela Müller-McNicoll, Florian Heyd, and Marco Preußner. 2018. “Characterization of Cis-Acting Elements That Control Oscillating Alternative Splicing.” RNA Biology 15 (8): 1081–92.

Gonzàlez-Porta, Mar, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. 2013. “Transcriptome Analysis of Human Tissues and Cell Lines Reveals One Dominant Transcript per Gene.” Genome Biology 14 (7): R70.

Gozani, O., R. Feld, and R. Reed. 1996. “Evidence That Sequence-Independent Binding of Highly Conserved U2 snRNP Proteins Upstream of the Branch Site Is Required for Assembly of Spliceosomal Complex A.” Genes & Development 10 (2): 233–43.

Harbour, J. William, Elisha D. O. Roberson, Hima Anbunathan, Michael D. Onken, Lori A. Worley, and Anne M. Bowcock. 2013. “Recurrent Mutations at Codon 625 of the Splicing Factor SF3B1 in Uveal Melanoma.” Nature Genetics 45 (2): 133–35.

Jiang, Moqin, Meng Chen, Qian Liu, Zhiling Jin, Xiangdong Yang, and Weifeng Zhang. 2023. “SF3B1 Mutations in Myelodysplastic Syndromes: A Potential Therapeutic Target for Modulating the Entire Disease Process.” Frontiers in Oncology 13 (March): 1116438.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity.” Science 337 (6096): 816–21.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. “Mutational Landscape and Significance across 12 Major Cancer Types.” Nature 502 (7471): 333–39.

Kelemen, Olga, Paolo Convertini, Zhaiyi Zhang, Yuan Wen, Manli Shen, Marina Falaleeva, and Stefan Stamm. 2013. “Function of Alternative Splicing.” Gene 514 (1): 1–30.

Kim, Eddo, Alon Magen, and Gil Ast. 2007. “Different Levels of Alternative Splicing among Eukaryotes.” Nucleic Acids Research 35 (1): 125–31.

Kim, Jinkuk, Chunguang Hu, Christelle Moufawad El Achkar, Lauren E. Black, Julie Douville, Austin Larson, Mary K. Pendergast, et al. 2019. “Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease.” The New England Journal of Medicine 381 (17): 1644–52.

Kong-Beltran, Monica, Somasekar Seshagiri, Jiping Zha, Wenjing Zhu, Kaumudi Bhawe, Nerissa Mendoza, Thomas Holcomb, et al. 2006. “Somatic Mutations Lead to an Oncogenic Deletion of Met in Lung Cancer.” Cancer Research 66 (1): 283–89.

Krämer, A. 1996. “The Structure and Function of Proteins Involved in Mammalian Pre-mRNA Splicing.” Annual Review of Biochemistry 65: 367–409.

Kuranaga, Yuki, Nobuhiko Sugito, Haruka Shinohara, Takuya Tsujino, Kohei Taniguchi, Kazumasa Komura, Yuko Ito, Tomoyoshi Soga, and Yukihiro Akao. 2018. “SRSF3, a Splicer of the PKM Gene, Regulates Cell Growth and Maintenance of Cancer-Specific Energy Metabolism in Colon Cancer Cells.” International Journal of Molecular Sciences 19 (10). https://doi.org/10.3390/ijms19103012.

Landau, Dan A., Eugen Tausch, Amaro N. Taylor-Weiner, Chip Stewart, Johannes G. Reiter, Jasmin Bahlo, Sandra Kluth, et al. 2015. “Mutations Driving CLL and Their Evolution in Progression and Relapse.” Nature 526 (7574): 525–30.

Lefave, Clare V., Massimo Squatrito, Sandra Vorlova, Gina L. Rocco, Cameron W. Brennan, Eric C. Holland, Ying-Xian Pan, and Luca Cartegni. 2011. “Splicing Factor hnRNPH Drives an Oncogenic Splicing Switch in Gliomas.” The EMBO Journal 30 (19): 4084–97.

Li, Dianyang, Wenying Yu, and Maode Lai. 2023. “Targeting Serine- and Arginine-Rich Splicing Factors to Rectify Aberrant Alternative Splicing.” Drug Discovery Today 28 (9): 103691.

Lieu, Yen K., Zhaoqi Liu, Abdullah M. Ali, Xin Wei, Alex Penson, Jian Zhang, Xiuli An, et al. 2022. “SF3B1 Mutant-Induced Missplicing of MAP3K7 Causes Anemia in Myelodysplastic Syndromes.” Proceedings of the National Academy of Sciences of the United States of America 119 (1). https://doi.org/10.1073/pnas.2111703119.

Liu, H. X., M. Zhang, and A. R. Krainer. 1998. “Identification of Functional Exonic Splicing Enhancer Motifs Recognized by Individual SR Proteins.” Genes & Development 12 (13): 1998–2012.

Liu, Xiao-Xiao, Qian-Huan Guo, Wei-Bo Xu, Peng Liu, and Kang Yan. 2022. “Rapid Regulation of Alternative Splicing in Response to Environmental Stresses.” Frontiers in Plant Science 13 (March): 832177.

López-Cánovas, Juan L., Mercedes Del Rio-Moreno, Helena García-Fernandez, Juan M. Jiménez-Vacas, M. Trinidad Moreno-Montilla, Marina E. Sánchez-Frias, Víctor Amado, et al. 2021. “Splicing Factor SF3B1 Is Overexpressed and Implicated in the Aggressiveness and Survival of Hepatocellular Carcinoma.” Cancer Letters 496 (January): 72–83.

Lorson, C. L., E. Hahnen, E. J. Androphy, and B. Wirth. 1999. “A Single Nucleotide in the SMN Gene Regulates Splicing and Is Responsible for Spinal Muscular Atrophy.” Proceedings of the National Academy of Sciences of the United States of America 96 (11): 6307–11.

Lu, Xinyuan, Nir Peled, John Greer, Wei Wu, Peter Choi, Alice H. Berger, Sergio Wong, et al. 2017. “MET Exon 14 Mutation Encodes an Actionable Therapeutic Target in Lung Adenocarcinoma.” Cancer Research 77 (16): 4498–4505.

Ma, Patrick C., Ramasamy Jagadeeswaran, Simha Jagadeesh, Maria S. Tretiakova, Vidya Nallasura, Edward A. Fox, Mark Hansen, et al. 2005. “Functional Expression and Mutations of c-Met and Its Therapeutic Inhibition with SU11274 and Small Interfering RNA in Non–Small Cell Lung Cancer.” Cancer Research 65 (4): 1479–88.

Ma, Patrick C., Takashi Kijima, Gautam Maulik, Edward A. Fox, Martin Sattler, James D. Griffin, Bruce E. Johnson, and Ravi Salgia. 2003. “C-MET Mutational Analysis in Small Cell Lung Cancer: Novel Juxtamembrane Domain Mutations Regulating Cytoskeletal Functions.” Cancer Research 63 (19): 6272–81.

Maguire, Sarah L., Andri Leonidou, Patty Wai, Caterina Marchiò, Charlotte Ky Ng, Anna Sapino, Anne-Vincent Salomon, Jorge S. Reis-Filho, Britta Weigelt, and Rachael C. Natrajan. 2015. “SF3B1 Mutations Constitute a Novel Therapeutic Target in Breast Cancer.” The Journal of Pathology 235 (4): 571–80.

Malcovati, Luca, Elli Papaemmanuil, David T. Bowen, Jacqueline Boultwood, Matteo G. Della Porta, Cristiana Pascutto, Erica Travaglino, et al. 2011. “Clinical Significance of SF3B1 Mutations in Myelodysplastic Syndromes and Myelodysplastic/myeloproliferative Neoplasms.” Blood 118 (24): 6239–46.

Mali, Prashant, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. 2013. “RNA-Guided Human Genome Engineering via Cas9.” Science 339 (6121): 823–26.

Martin, Marcel, Lars Maßhöfer, Petra Temming, Sven Rahmann, Claudia Metz, Norbert Bornfeld, Johannes van de Nes, et al. 2013. “Exome Sequencing Identifies Recurrent Somatic Mutations in EIF1AX and SF3B1 in Uveal Melanoma with Disomy 3.” Nature Genetics 45 (8): 933–36.

Mathieu, Luckson N., Erin Larkins, Oladimeji Akinboro, Pourab Roy, Anup K. Amatya, Mallorie H. Fiero, Pallavi S. Mishra-Kalyani, et al. 2022. “FDA Approval Summary: Capmatinib and Tepotinib for the Treatment of Metastatic NSCLC Harboring MET Exon 14 Skipping Mutations or Alterations.” Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 28 (2): 249–54.

Monani, U. R., C. L. Lorson, D. W. Parsons, T. W. Prior, E. J. Androphy, A. H. Burghes, and J. D. McPherson. 1999. “A Single Nucleotide Difference That Alters Splicing Patterns Distinguishes the SMA Gene SMN1 from the Copy Gene SMN2.” Human Molecular Genetics 8 (7): 1177–83.

Mount, S. M., I. Pettersson, M. Hinterberger, A. Karmas, and J. A. Steitz. 1983. “The U1 Small Nuclear RNA-Protein Complex Selectively Binds a 5’ Splice Site in Vitro.” Cell 33 (2): 509–18.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” Nature 463 (7280): 457–63.

Obeng, Esther A., Ryan J. Chappell, Michael Seiler, Michelle C. Chen, Dean R. Campagna, Paul J. Schmidt, Rebekka K. Schneider, et al. 2016. “Physiologic Expression of Sf3b1K700E Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation.” Cancer Cell 30 (3): 404–17.

Oh, Do-Youn, and Yung-Jue Bang. 2020. “HER2-Targeted Therapies - a Role beyond Breast Cancer.” Nature Reviews. Clinical Oncology 17 (1): 33–48.

Organ, Shawna Leslie, and Ming-Sound Tsao. 2011. “An Overview of the c-MET Signaling Pathway.” Therapeutic Advances in Medical Oncology 3 (1 Suppl): S7–19.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” Nature Genetics 40 (12): 1413–15.

Papaemmanuil, E., M. Cazzola, J. Boultwood, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, et al. 2011. “Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts.” The New England Journal of Medicine 365 (15): 1384–95.

Pearn, J. 1978. “Incidence, Prevalence, and Gene Frequency Studies of Chronic Childhood Spinal Muscular Atrophy.” Journal of Medical Genetics 15 (6): 409–13.

Pereira, Bernard, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, et al. 2016. “The Somatic Mutation Profiles of 2,433 Breast Cancers Refine Their Genomic and Transcriptomic Landscapes.” Nature Communications 7 (1): 1–16.

Peschard, P., T. M. Fournier, L. Lamorte, M. A. Naujokas, H. Band, W. Y. Langdon, and M. Park. 2001. “Mutation of the c-Cbl TKB Domain Binding Site on the Met Receptor Tyrosine Kinase Converts It into a Transforming Protein.” Molecular Cell 8 (5): 995–1004.

Porensky, Paul N., and Arthur H. M. Burghes. 2013. “Antisense Oligonucleotides for the Treatment of Spinal Muscular Atrophy.” Human Gene Therapy 24 (5): 489–98.

Quesada, Víctor, Laura Conde, Neus Villamor, Gonzalo R. Ordóñez, Pedro Jares, Laia Bassaganyas, Andrew J. Ramsay, et al. 2011. “Exome Sequencing Identifies Recurrent Mutations of the Splicing Factor SF3B1 Gene in Chronic Lymphocytic Leukemia.” Nature Genetics 44 (1): 47–52.

Rappsilber, Juri, Ursula Ryder, Angus I. Lamond, and Matthias Mann. 2002. “Large-Scale Proteomic Analysis of the Human Spliceosome.” Genome Research 12 (8): 1231–45.

Rossi, Davide, Alessio Bruscaggin, Valeria Spina, Silvia Rasi, Hossein Khiabanian, Monica Messina, Marco Fangazio, et al. 2011. “Mutations of the SF3B1 Splicing Factor in Chronic Lymphocytic Leukemia: Association with Progression and Fludarabine-Refractoriness.” Blood 118 (26): 6904–8.

Schaal, T. D., and T. Maniatis. 1999. “Multiple Distinct Splicing Enhancers in the Protein-Coding Sequences of a Constitutively Spliced Pre-mRNA.” Molecular and Cellular Biology 19 (1): 261–73.

Shi, Lei, Caihua Xu, Yutong Ma, Qiuxiang Ou, Xue Wu, Songhua Lu, Yang Shao, Renhua Guo, and Jinliang Kong. 2020. “Clinical Significance of ERBB2 Exon 16 Skipping: Analysis of a Real-World Retrospective Observational Cohort Study.” ESMO Open 5 (6): e000985.

Singh, Ravi K., and Thomas A. Cooper. 2012. “Pre-mRNA Splicing in Disease and Therapeutics.” Trends in Molecular Medicine 18 (8): 472–82.

Smith, Harvey W., Lei Yang, Chen Ling, Arlan Walsh, Victor D. Martinez, Jonathan Boucher, Dongmei Zuo, et al. 2020. “An ErbB2 Splice Variant Lacking Exon 16 Drives Lung Carcinoma.” Proceedings of the National Academy of Sciences 117 (33): 20139–48.

Steward, Rachel A., Maaike A. de Jong, Vicencio Oostra, and Christopher W. Wheat. 2022. “Alternative Splicing in Seasonal Plasticity and the Potential for Adaptation to Environmental Change.” Nature Communications 13 (1): 1–12.

Suñé-Pou, Marc, María J. Limeres, Cristina Moreno-Castro, Cristina Hernández-Munain, Josep M. Suñé-Negre, María L. Cuestas, and Carlos Suñé. 2020. “Innovative Therapeutic and Delivery Approaches Using Nanotechnology to Correct Splicing Defects Underlying Disease.” Frontiers in Genetics 11 (July): 731.

Suzawa, Ken, Michael Offin, Adam J. Schoenfeld, Andrew J. Plodkowski, Igor Odintsov, Daniel Lu, William W. Lockwood, et al. 2019. “Acquired MET Exon 14 Alteration Drives Secondary Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor in EGFR-Mutated Lung Cancer.” JCO Precision Oncology 3 (May). https://doi.org/10.1200/PO.19.00011.

Sveen, A., S. Kilpinen, A. Ruusulehto, R. A. Lothe, and R. I. Skotheim. 2016. “Aberrant RNA Splicing in Cancer; Expression Changes and Driver Mutations of Splicing Factor Genes.” Oncogene 35 (19): 2413–27.

Tam, Annie S., Shuhe Tsai, Emily Yun-Chia Chang, Veena Mathew, Alynn Shanks, T. Roderick Docking, Arun Kumar, Delphine G. Bernard, Aly Karsan, and Peter C. Stirling. n.d. “DYNLL1 Mis-Splicing Is Associated with Replicative Genome Instability in SF3B1 Mutant Cells.” https://doi.org/10.1101/2021.05.26.445839.

Togashi, Yosuke, Hiroshi Mizuuchi, Shuta Tomida, Masato Terashima, Hidetoshi Hayashi, Kazuto Nishio, and Tetsuya Mitsudomi. 2015. “MET Gene Exon 14 Deletion Created Using the CRISPR/Cas9 System Enhances Cellular Growth and Sensitivity to a MET Inhibitor.” Lung Cancer  90 (3): 590–97.

Tung, Kuo-Feng, Chao-Yu Pan, Chao-Hsin Chen, and Wen-Chang Lin. 2020. “Top-Ranked Expressed Gene Transcripts of Human Protein-Coding Genes Investigated with GTEx Dataset.” Scientific Reports 10 (1): 16245.

Turpin, J., C. Ling, E. J. Crosby, Z. C. Hartman, A. M. Simond, L. A. Chodosh, J. P. Rennhack, et al. 2016. “The ErbB2ΔEx16 Splice Variant Is a Major Oncogenic Driver in Breast Cancer That Promotes a pro-Metastatic Tumor Microenvironment.” Oncogene 35 (47): 6053–64.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. “Proteomics. Tissue-Based Map of the Human Proteome.” Science 347 (6220): 1260419.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” Nature 456 (7221): 470–76.

Wang, Feng, Yang Liu, Wanglong Qiu, Elaine Shum, Monica Feng, Dejian Zhao, Deyou Zheng, Alain Borczuk, Haiying Cheng, and Balazs Halmos. 2022. “Functional Analysis of MET Exon 14 Skipping Alteration in Cancer Invasion and Metastatic Dissemination.” Cancer Research 82 (7): 1365–79.

Wang, Lili, Michael S. Lawrence, Youzhong Wan, Petar Stojanov, Carrie Sougnez, Kristen Stevenson, Lillian Werner, et al. 2011. “SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia.” The New England Journal of Medicine 365 (26): 2497–2506.

Wang, Tim, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. 2015. “Identification and Characterization of Essential Genes in the Human Genome.” Science 350 (6264): 1096–1101.

Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. “Genetic Screens in Human Cells Using the CRISPR-Cas9 System.” Science 343 (6166): 80–84.

Wang, Yan, Jing Liu, B. O. Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, et al. 2015. “Mechanism of Alternative Splicing and Its Regulation.” Biomedical Reports 3 (2): 152–58.

Watson, Ian R., Koichi Takahashi, P. Andrew Futreal, and Lynda Chin. 2013. “Emerging Patterns of Somatic Mutations in Cancer.” Nature Reviews. Genetics 14 (10): 703–18.

Weischenfeldt, Joachim, Johannes Waage, Geng Tian, Jing Zhao, Inge Damgaard, Janus Schou Jakobsen, Karsten Kristiansen, Anders Krogh, Jun Wang, and Bo T. Porse. 2012. “Mammalian Tissues Defective in Nonsense-Mediated mRNA Decay Display Highly Aberrant Splicing Patterns.” Genome Biology 13 (5): R35.

Will, Cindy L., and Reinhard Lührmann. 2011. “Spliceosome Structure and Function.” Cold Spring Harbor Perspectives in Biology 3 (7). https://doi.org/10.1101/cshperspect.a003707.

Wu, J., and J. L. Manley. 1989. “Mammalian Pre-mRNA Branch Site Selection by U2 snRNP Involves Base Pairing.” Genes & Development 3 (10): 1553–61.

Wu, J. Y., and T. Maniatis. 1993. “Specific Interactions between Proteins Implicated in Splice Site Selection and Regulated Alternative Splicing.” Cell 75 (6): 1061–70.

Wu, S., C. M. Romfo, T. W. Nilsen, and M. R. Green. 1999. “Functional Recognition of the 3’ Splice Site AG by the Splicing Factor U2AF35.” Nature 402 (6763): 832–35.

Yoshida, Hisashi, Sam-Yong Park, Takashi Oda, Taeko Akiyoshi, Mamoru Sato, Mikako Shirouzu, Kengo Tsuda, et al. 2015. “A Novel 3’ Splice Site Recognition by the Two Zinc Fingers in the U2AF Small Subunit.” Genes & Development 29 (15): 1649–60.

Yoshida, Kenichi, Masashi Sanada, Yuichi Shiraishi, Daniel Nowak, Yasunobu Nagata, Ryo Yamamoto, Yusuke Sato, et al. 2011. “Frequent Pathway Mutations of Splicing Machinery in Myelodysplasia.” Nature 478 (7367): 64–69.

Zamore, P. D., J. G. Patton, and M. R. Green. 1992. “Cloning and Domain Structure of the Mammalian Splicing Factor U2AF.” Nature 355 (6361): 609–14.

Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. “Alternative Splicing and Cancer: A Systematic Review.” Signal Transduction and Targeted Therapy 6 (1): 78.

Zhuang, Y., and A. M. Weiner. 1986. “A Compensatory Base Change in U1 snRNA Suppresses a 5’ Splice Site Mutation.” Cell 46 (6): 827–35.

# Chapter 2: Using ssCRISPR to identify and validate novel aberrant exon skipping events driving LUAD

## Abstract

While it’s known that aberrant splicing contributes to cancer, the identification of specific oncogenic splicing events is a bottleneck for cancer research. To this end we developed **s**plice **s**ite CRISPR (ssCRISPR) which combines a computational pipeline to identify aberrant exon skipping events in cancer sequencing data, and CRISPR/Cas9 to force the production of those exon skipping events in a lung cell line model. Through a computational analysis of whole-exome and matched RNA sequencing data from lung adenocarcinomas (LUAD), we identified 994 exon skipping events with the potential to contribute to oncogenesis. Among these candidates, we identified five high-interest in-frame skipped exons from genes within the Ras-MAPK pathway, a pathway frequently mutated in cancer. I hypothesize that a subset of the 994 exon skipping events contribute to the initiation and maintenance of LUAD tumors. To answer this question I used ssCRISPR in a high-throughput format coupled with low attachment growth to enrich for transformed cells. For the high-interest Ras-MAPK candidates, I performed ssCRISPR on an individual scale to create lung cell lines expressing each candidate isoform. I found that likely due to the low expression of these variants from their endogenous promoter, CRISPR-based expression is not sufficient to functionally validate these potentially oncogenic isoforms.

**Introduction:**

A complex mutational landscape is a hallmark of cancer. These cancer-associated alterations are composed of driver mutations, which drive carcinogenesis, and passenger mutations, which are generally considered not to contribute to disease (Vogelstein et al. 2013; Tomasetti et al. 2013). In the case of lung adenocarcinoma (LUAD), the average sample from The Cancer Genome Atlas (TCGA) has about 280 mutations per tumor (TCGA) However, it is estimated that there are three driver mutations per LUAD tumor (Tomasetti et al. 2015). These oncogenic driver mutations make for attractive drug targets, enabling better targeted cancer therapies for LUAD patients who would otherwise rely on chemotherapy (Herbst et al. 2018). A recently appreciated class of oncogenic drivers are the result of aberrant splicing. Typically, mutations at splice sites disrupt the spliceosome's ability to identify the exon, forcing an aberrant exon skipping event (Kong-Beltran et al. 2006; Lu et al. 2017; Turpin et al. 2016; Smith et al. 2020). Two well characterized lung adenocarcinoma cancer drivers are created this way: *MET* exon 14 skipping (*METΔ14*) and *ERBB2* exon 16 skipping (*ERBB2Δ16*) (Smith et al. 2020; Kong-Beltran et al. 2006; Lu et al. 2017). Furthermore, patients with *METΔ14* mutations are known to respond favorably to drugs targeting the protein product of the *METΔ14* oncogene (Mathieu et al. 2022; Lu et al. 2017).

While individual events are well characterized, a systematic interrogation of the contribution of aberrant splicing events driving cancer needs to be explored. Due to the high somatic mutation frequency in the exomes of lung cancer tumors (Lawrence et al. 2013), LUAD is an excellent model system to identify novel aberrant exon skipping events resulting from somatic mutations at splice sites and within splicing machinery. We focused on exon skipping as this is the most common splicing event in humans (Kim et al. 2007). Furthermore, exon skipping can drastically impact isoforms, as in frame exon skipping can remove entire functional domains or regulatory binding sites which alter canonical protein function (Kong-Beltran et al. 2006; Lu et al. 2017; Turpin et al. 2016; Smith et al. 2020). While recent work computationally predicts the consequences of skipped exons in cancer (Kim et al. 2020), these isoforms are not biologically validated to confirm their oncogenic potential.

To identify and validate oncogenic drivers of LUAD resulting from aberrant exon splicing events, we developed ssCRISPR: a high-throughput screening method using CRISPR/Cas9 to force the production of exon skipping events. Using a computational analysis of whole exome and matched RNA-seq data from LUADs, we identified 994 potentially oncogenic exon skipping events. The oncogenic potential of these candidates were measured using a Growth in Low Attachment Assay (GILA), which enriches for cells expressing oncogenes (Rotem et al. 2015; Izar and Rotem 2016). Of these skipping events, we identified five candidates in the Ras-MAPK pathway, which is commonly altered in LUAD (Herbst et al. 2018). Among these Ras-MAPK candidates, we identified *METΔ14* and *ERBB2Δ16*, confirming our computational analysis can detect oncogenic exon skipping events. The remaining candidates were three novel in-frame exon skipping events: *MTORΔ12, BRAFΔ13, and NF1Δ46* (Table 2.1). Because these candidates were in a pathway of interest, they were interrogated more rigorously using a single-guide ssCRISPR approach to cause exon skipping. I hypothesized that a subset of these exon skipping events drive tumorigenesis. While the original intention was to reveal the extent to which aberrant splicing contributes to cancer development, I found instead that low dosage of these exon skipping events produced by CRISPR-based expression was not sufficient to confer an oncogenic phenotype in the GILA assay. I suggest overexpression of these variants coupled with a GILA assay is the optimal way to characterize the oncogenic potential of these exon skipping events.

Table 2.1: Five aberrantly spliced RAS hg19 candidates identified from computational analysis

| Gene | ΔExon | Exon Coordinates |
|---|---|---|
| *BRAF* | 13 | chr7:140476712-140476888 |
| *ERBB2* | 16 | chr17:37876040-37876550 |
| *MET* | 14 | chr7:116411903-116412043 |
| *MTOR* | 12 | chr1:11298459-11298674 |
| *NF1* | 46 | chr17:29665722-29665823 |

**Results:**

<u>*ssCRISPR sgRNA library design*</u>

ssCRISPR combines a computational analysis to identify candidates from TCGA data and a CRISPR/Cas9 high-throughput screen to force the production of exon skipping events across a population of immortalized tracheobronchial epithelial cells (AALE) (Lundberg et al. 2002). This computational analysis yielded 994 candidate exon skipping events with the potential to contribute to LUAD tumorigenesis. These candidates were identified from four different criteria: 1) somatic mutations at splice sites, 2) U2AF1 S34F and RBM10 loss-of-function (LOF) mutations (common alterations in LUAD (Imielinski et al.

2012)), 3) driver-negative tumors, and 4) a cohort of exon skipping events identified with Guardant Health using proprietary methods.

We designed the single guide (sgRNA) library with the CRISPOR tool, which maximizes targeting efficiency while minimizing off-target effects (Figure 2.1) (Haeussler et al. 2016). A maximum of three sgRNAs targeted each splice site. To distinguish between exon skipping events that create loss of function (LOF) proteins, we designed control sgRNAs targeting the first exon of each candidate gene. We also designed negative control sgRNAs targeting the center of the candidate exon because this would be less likely to result in exon skipping. Finally, we included 150 negative control sgRNAs which do not have targets in the human genome. As these sgRNAs would not selectively enrich in low attachment, these would behave as passenger mutations and control for exon skipping events with no bearing on oncogenesis.

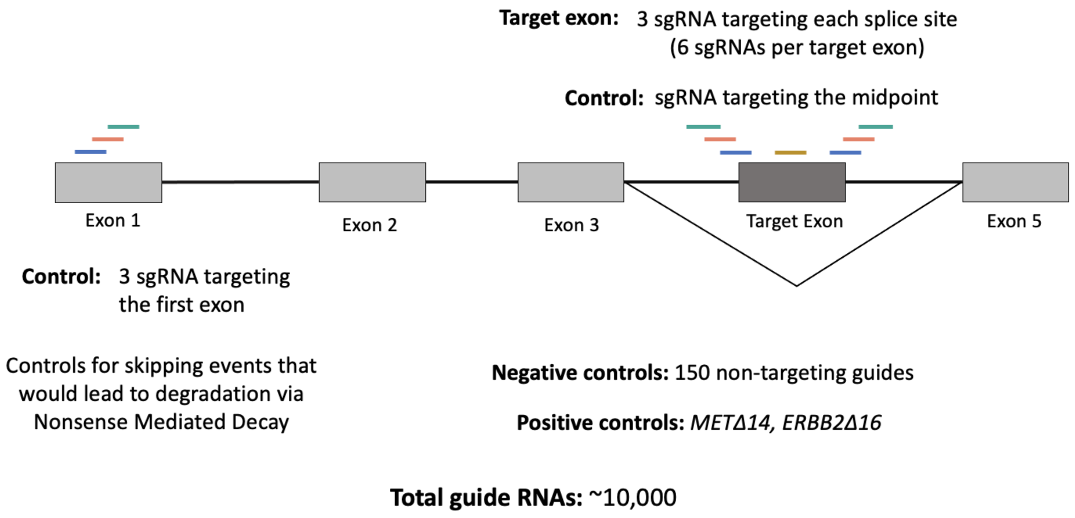

**Target exon:** 3 sgRNA targeting each splice site
(6 sgRNAs per target exon)
**Control:** sgRNA targeting the midpoint
Exon 1
Exon 2
Exon 3
Target Exon
Exon 5
**Control:** 3 sgRNA targeting the first exon
Controls for skipping events that would lead to degradation via Nonsense Mediated Decay
**Negative controls:** 150 non-targeting guides
**Positive controls:** *METΔ14, ERBB2Δ16*
**Total guide RNAs:** ~10,000

**Figure 2.1: ssCRISPR sgRNA Library Design.**

A maximum of three single guide RNAs (sgRNAs) target each splice site of a candidate exon. sgRNAs targeting the first exon control for exon skipping events that would lead to overall decreased gene expression through nonsense mediated decay. 150 non-targeting sgRNAs control for exon skipping events that do not confer a selective advantage in low attachment. sgRNAs targeting candidate exons of *METΔ14* and *ERBB2Δ16* serve as positive controls in the screen.

*Validations of ssCRISPR screen*

I cloned the ssCRISPR sgRNA library using the Agilent SureVector CRISPR Library Cloning System, which uses recombination cloning to insert each sgRNA into a lentiviral backbone. Because cloning imposes a risk of sgRNA dropout, I assessed sgRNA distribution with next-generation sequencing (NGS). I PCR-amplified the sgRNAs using primers flanking the sgRNA sequence. This added NGS-compatible sequences (regions to hybridize with the flow cell and unique barcoding sequences) flanking the sgRNA amplicons. After sequencing this pool with a MiSeq, I used a custom Python script to quantify sgRNAs in the cloned pool. This program revealed a tight distribution of sgRNAs in this cloned library, even better distributed than two other sgRNA libraries used in successful CRISPR screens (Figure 2.2A).

I validated Cas9 activity in AALEs constitutively expressing Cas9 (AALE-Cas9) using an sgRNA targeting the first exon of the non-essential gene *CD44.* Cells that integrated both this *CD44* sgRNA and retained functional Cas9 would result in knockdown of total *CD44* levels. Using a CD44-GFP conjugated antibody and subsequent FACS analysis, I measured a near 50% knockdown of GFP compared to the non-targeting sgRNA control, indicating functional Cas9 is present in AALE-Cas9 (Figure 2.2B-D).

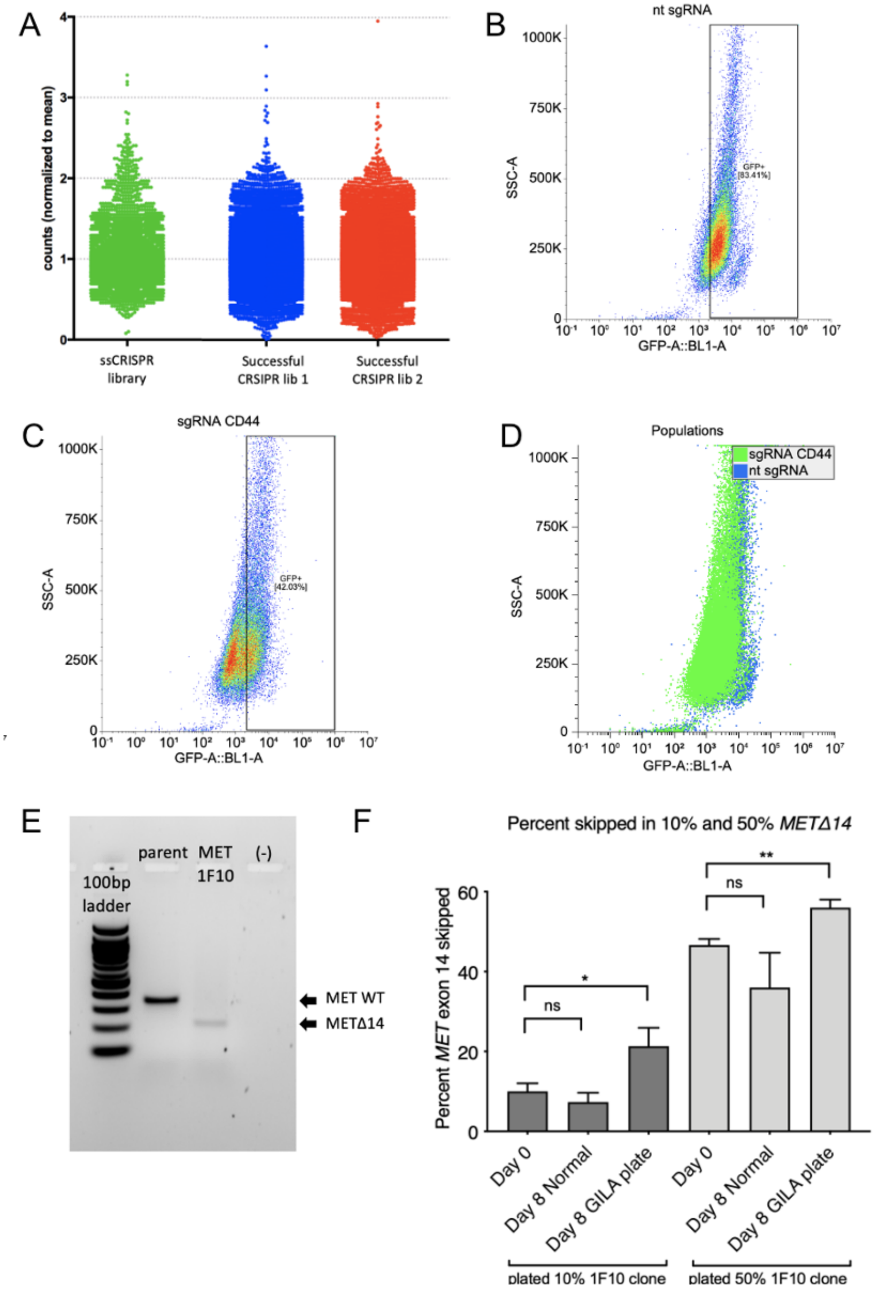


**Figure 2.2: Validations of ssCRISPR sgRNA library, Cas9 functionality, and assay feasibility.**

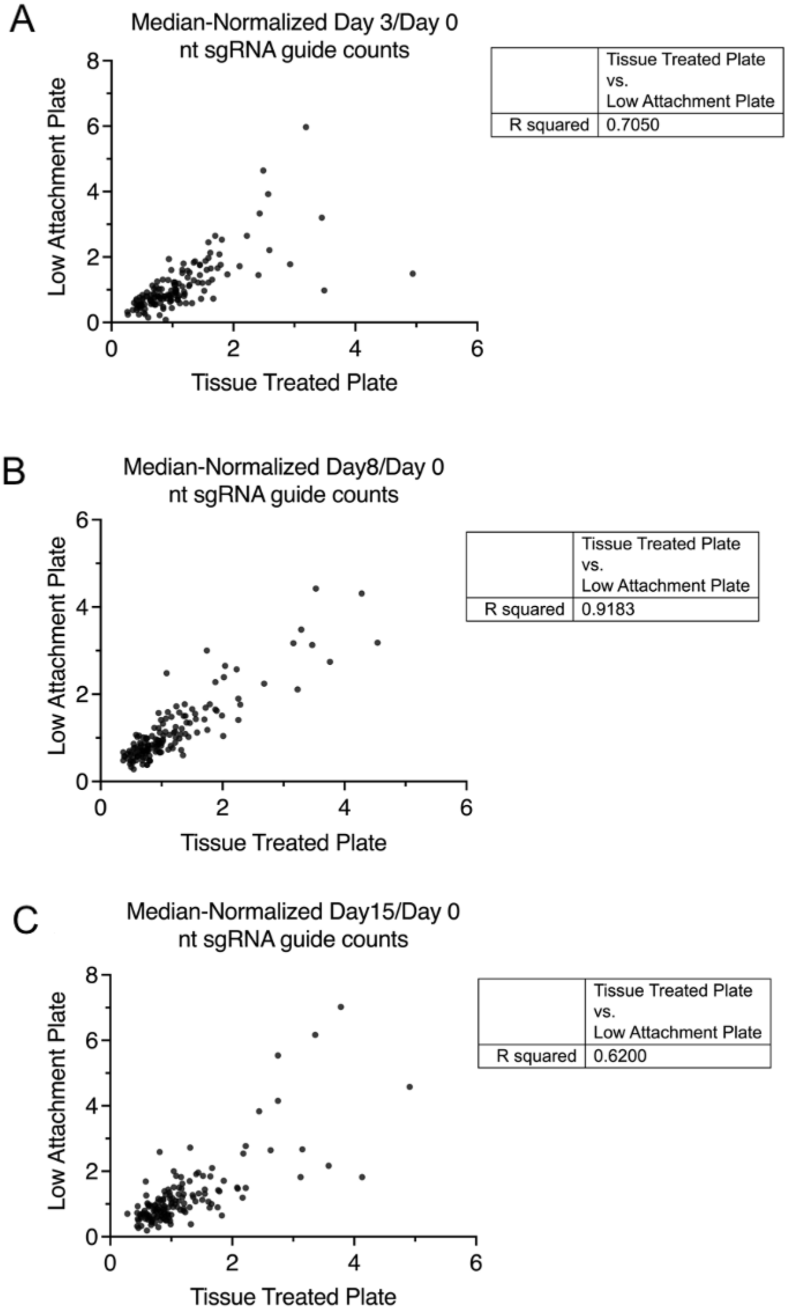
A) Normalized read counts of cloned ssCRISPR sgRNA library compared with two successful sgRNA CRISPR libraries. B) GFP+ cells (boxed region) of CD44-GFP stained cells infected with the non-targeting (nt) sgRNA. C) GFP+

cells (boxed region) of CD44-GFP stained cells infected with the *CD44*-targeting sgRNA (sgRNA CD44) D) Overlay of cell populations for GFP+ cells in the nt sgRNA cells and sgRNA CD44 cells E) RT-PCR using exon 14 spanning primers showing predominant *METΔ14* isoform in MET 1F10 clone. F) Mixing experiment combining 10% MET1F10 or 50% 1F10 clone with the parental line. Day 0 represents the initial *METΔ14* ratio at the day of plating. Day 8 normal is the amount of *METΔ14* at day 8 grown in a tissue treated plate. Day 8 GILA is the amount of *METΔ14* at day 8 grown in a low attachment plate.

To ensure the GILA assay can enrich for cells expressing aberrant exon skipping events created with ssCRISPR, I tested the ability of *METΔ14* to enrich from a mixed population of cells grown in low attachment using an AALE clone expressing only *METΔ14* (MET 1F10) (Figure 2.2E). I mixed MET 1F10 in ratios of 10% and 50% with the parental AALE line. After growth in low attachment for 8 days, I extracted RNA from these mixed populations, generated cDNA, and used exon spanning primers across exon 14 to PCR amplify both wild type *MET* (MET WT - from the parental AALE population) and the *METΔ14* isoform (from MET 1F10). After calculating the percent *METΔ14* in each sample, I identified an enrichment of *METΔ14* from both the 10% and 50% population (Figure 2.2F). This enrichment was not recapitulated in the normal tissue treated plate, confirming low attachment specifically enriches for aberrantly spliced cancer drivers created with ssCRISPR. Furthermore, *METΔ14* enrichment is not dependent on the initial population of MET 1F10, indicating that small populations of oncogenes can enrich in the ssCRIPSR screen. Thus, three criteria suggest that the ssCRISPR screen can generate and enrich for aberrantly spliced oncogenes: 1) the ideal sgRNA distribution in our ssCRISPR library, 2) validation of functional Cas9 in our AALE-Cas9 line, and 3) confirmation of *METΔ14* enrichment in the GILA assay.

### *ssCRISPR screen unable to enrich for aberrant exon skipping events*

I introduced the ssCRISPR sgRNA library in AALE-Cas9 cells using lentivirus, allowing for constitutive sgRNA expression. To ensure a ratio of one sgRNA per cell, I infected at the low multiplicity of infection (MOI) of 0.3, meaning around ~30% of the cells received an sgRNA. This is with the goal of assessing one exon skipping event at a time, as opposed to the combined effects of multiple exon skipping events in one cell. After selection with puromycin for four days, I expanded this infection to establish a 1000x coverage across each condition. Because the ssCRISPR library contained close to 10,000 sgRNAs, I required 10 million cells per time point. Because low attachment growth specifically enriches for oncogenic cells while simultaneously killing normal cells (Rotem et al. 2015 and Figure 2.1 F), I compared the enrichment of ssCRISPR-expressed exon skipping events in a GILA plate (low attachment) relative to their growth in a tissue treated plate (normal growth). In addition to these growth conditions, I obtained a Day 0, Day 3, Day 8 and Day 15 timepoint to measure enrichment of these exon skipping events over time. I obtained the Day 0 timepoint at the day of plating the Day 3, 8, and 15 timepoints, ensuring Day 0 represents the sgRNA population before selection. This resulted in one Day 0 timepoint, Day 3, 8 and 15 “tissue culture plate” timepoints and Day 3, 8 and 15 “low attachment plate” timepoints. I quantified sgRNAs as a metric for exon skipping event enrichment. This is with the assumption that the sgRNA produced an exon skipping event that provided a proliferative advantage in low attachment. Because every sgRNA lentivirally integrated into the cell’s genome with a universal lentiviral backbone, I isolated genomic DNA at the end of each



|  | Tissue Treated Plate<br>vs.<br>Low Attachment Plate |
|---|---|
| R squared | 0.7050 |

|  | Tissue Treated Plate<br>vs.<br>Low Attachment Plate |
|---|---|
| R squared | 0.9183 |

|  | Tissue Treated Plate<br>vs.<br>Low Attachment Plate |
|---|---|
| R squared | 0.6200 |

**Figure 2.3: Expected non-targeting sgRNA enrichment in ssCRISPR screen.**

Non-targeting (nt) sgRNA enrichment determined by dual normalization of raw sgRNA counts to 1) the timepoint median then 2) to the Day 0 timepoint. Normalized sgRNA counts were plotted against each other in Prism compared using a Pearson correlation.

timepoint to prepare sgRNAs for next generation sequencing via PCR amplification.

PCR-induced overamplification can produce false positives by creating biases in sgRNA enrichment. Therefore, I compared sgRNA enrichment of nt sgRNAs between both the tissue treated plate and low attachment plate for each timepoint. Since low attachment compels non-oncogenic cells to perish through time, I hypothesized that these nt sgRNAs would enrich in the tissue treated plate, or remain evenly distributed between both growth conditions. Using a custom Python script to count sgRNAs from Illumina sequencing data, I found these nt sgRNAs were evenly enriched in both conditions, as evidenced by a high correlation score, with slight enrichment in the tissue treated plate (Figure 2.3). These results imply the sequencing strategy is appropriate for amplifying and sequencing sgRNAs from the genome. Additionally, this confirms the low attachment screen does not enrich sgRNAs with no bearing on oncogenesis.

Next I examined sgRNA enrichment of our positive control producing *METΔ14*. Because disruption of different splice sites may be more impactful due to their sequence features, sgRNAs targeting each splice site are colored accordingly (Figure 2.4A). When taking into account sgRNAs targeting both splice sites, there is no significant difference in sgRNA enrichment at any time point (Figure 2.4B). Separating this data per splice site target, the sgRNAs targeting the 5’ splice site did not enrich in any low attachment condition (Figure 2.4C). However, the sgRNAs targeting the 3’ splice site are significantly enriched at Day 15 in low attachment (Figure 2.4D). While I anticipated sgRNA enrichment for *METΔ14,* it is surprising to only see enrichment in the latest timepoint, while typically enrichment is seen after 8 days (Izar and Rotem 2016; Rotem et al. 2015). This suggests the amount of *METΔ14* produced with CRISPR-based expression may be too subtle to detect with sgRNA enrichment. Additionally, due to the unequal enrichment of sgRNAs targeting each splice site (Figure 2.4 C-D), this suggests that to identify novel skipped exons that contribute to LUAD I am unable to increase statistical power by combining all sgRNAs targeting each candidate exon.

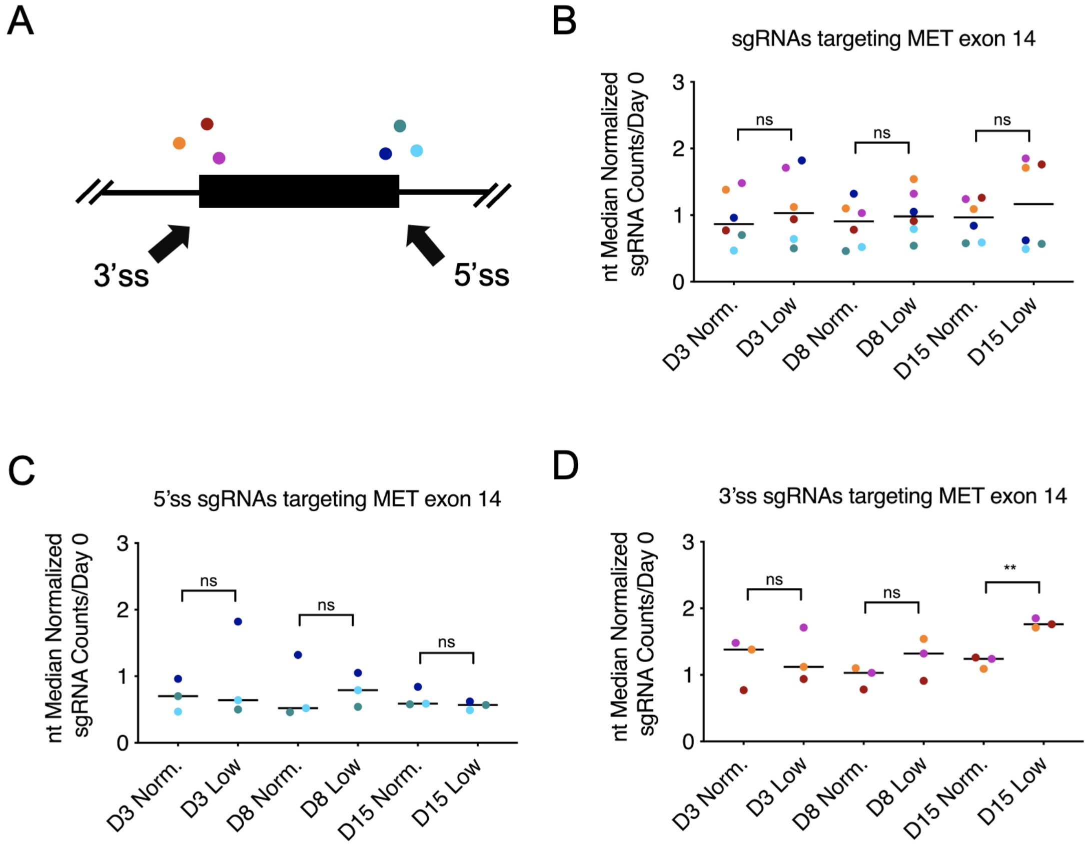


**Figure 2.4: Low sgRNA enrichment for *METΔ14.***

A) Diagram of sgRNAs targeting each exon. Cool-colored dots target the 5' splice site, and warm-colored dots target the 3' splice site. B) Comparison of both 3' and 5' sgRNA enrichment in low attachment compared to 'normal' tissue treated plates. Comparison of sgRNA enrichment for C) 5' splice site-targeting guides and D) 3' splice site-targeting guides.

In contrast to *METΔ14,* no sgRNA targeting *ERBB2Δ16,* the other positive control in the screen, enriched at any timepoint in low attachment for any splice site (Figure 2.5). Additionally, no sgRNAs targeting the high-interest Ras-MAPK candidates, *MTORΔ12, BRAFΔ13,* and *NF1Δ46,* enriched in the ssCRISPR screen (Figure 2.6). While taken at face value this suggests these exon skipping events do not drive oncogenesis, *ERBB2Δ16* is a well-known aberrantly spliced oncogene driving LUAD among other cancers (Smith et al 2020; Turpin et al. 2016). It's possible that CRISPR-based expression of these

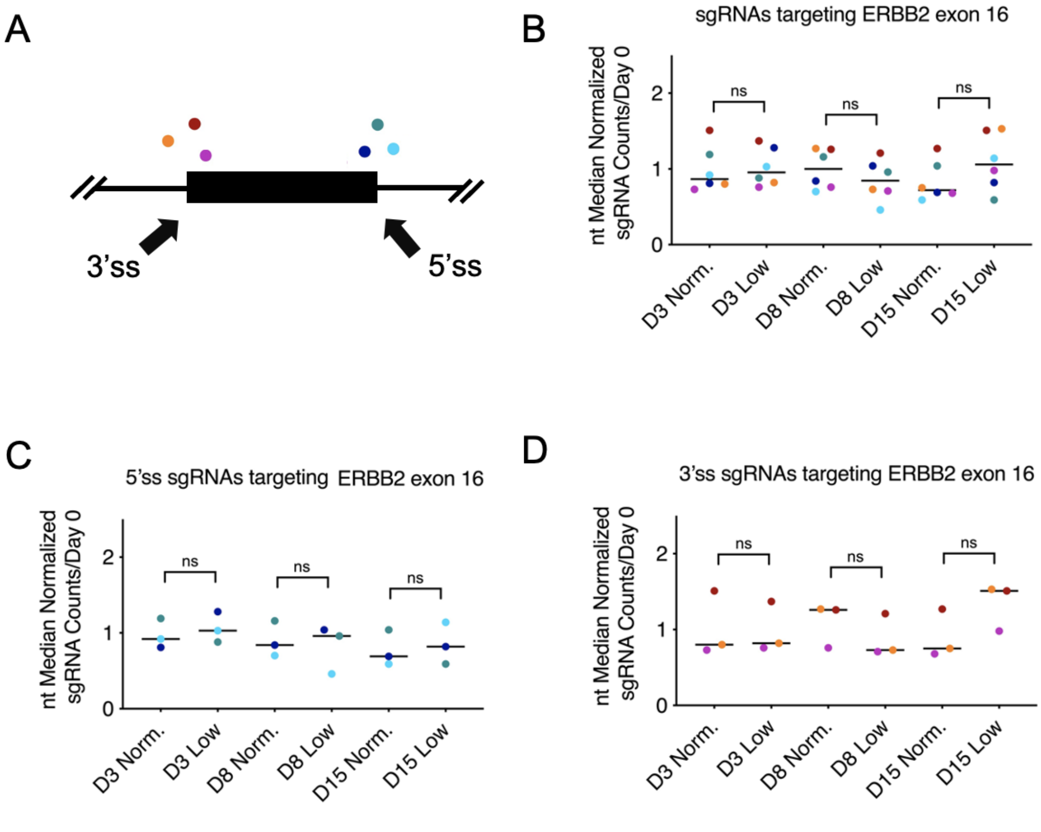


**Figure 2.5: No sgRNA enrichment for *ERBB2Δ16.***

A) Diagram of sgRNAs targeting each exon. Cool-colored dots target the 5’ splice site, and warm-colored dots target the 3’ splice site. B) Comparison of both 3’ and 5’ sgRNA enrichment in low attachment compared to ‘normal’ tissue treated plates. Comparison of sgRNA enrichment for C) 5’ splice site-targeting guides and D) 3’ splice site-targeting guides.

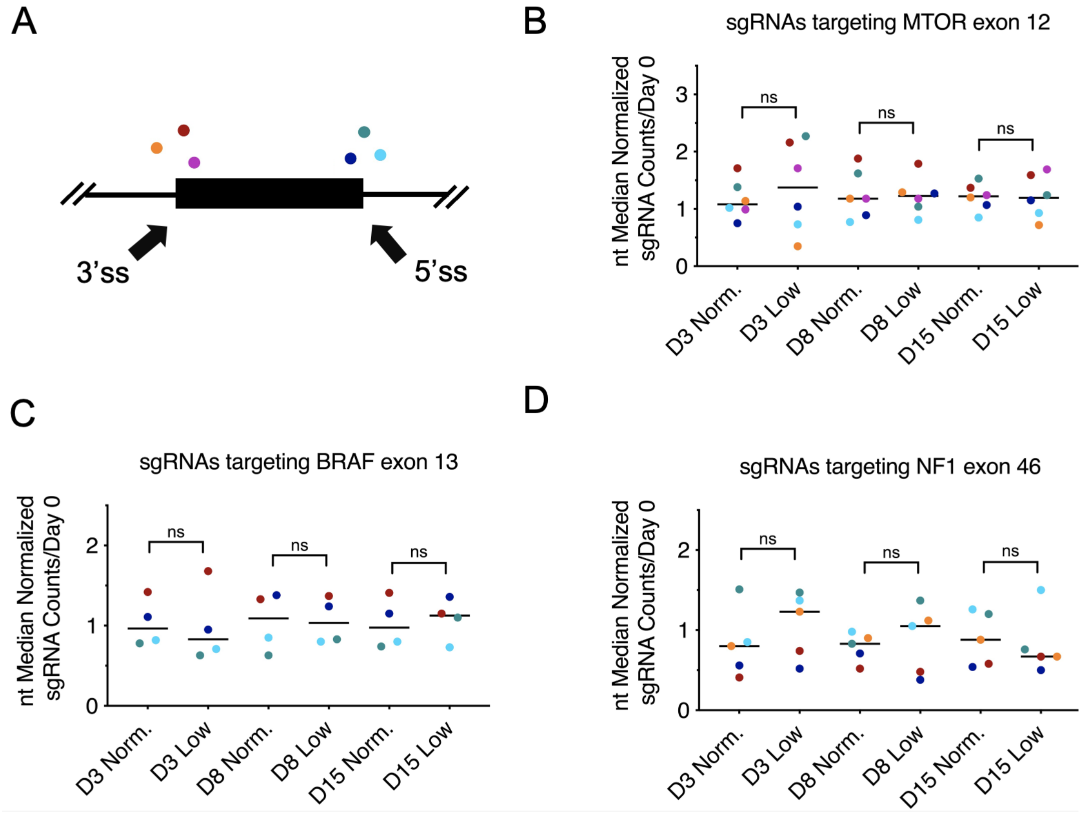


**Figure 2.6: No sgRNA enrichment for *MTORΔ12*, *BRAFΔ13*, or *NF1Δ46.***

A) Diagram of sgRNAs targeting each exon. Comparison of both 3’ and 5’ sgRNA enrichment in low attachment compared to ‘normal’ tissue treated plates for *MTORΔ12* (B), *BRAFΔ13* (C) and *NF1Δ46* (D). An sgRNA quantity lower than 6 indicates less than 3 guides targeting the corresponding splice site.

exon skipping events creates subtle differences in the cell's ability to survive in low attachment, and using sgRNA enrichment as a binary metric does not capture more complex expression changes of these exon skipping events in cells. This is supported by Figure 2.2F, where I determined the *METΔ14* **transcript** enriches in low attachment after 8 days using RT-PCR, as opposed to quantifying by sgRNA enrichment. Unfortunately, due to the nature of the screen it is impossible to use transcript-level quantification to measure abundance of the pooled exon skipping events.

Another drawback with the screen stems from the variability in the effectiveness of splice site targeting sgRNAs. While typically 3 sgRNAs target each splice site totalling 6 per exon, not all sgRNAs will edit the splice site. We constructed these sgRNAs in a high-throughput format with limited targetable genetic real estate, determined by proximity to the splice site and the Cas9 PAM sequence. This means that based on intrinsic factors of the sgRNA sequence (Concordet and Haeussler 2018), some sgRNAs will consequently perform better than others leading many of these sgRNAs to function like nt sgRNAs (Figure 2.7). This makes it difficult to confidently detect GOF exon skipping events which may confer a selective advantage in low attachment, as we rely on the statistical power of the enrichment of multiple sgRNAs targeting the same exon. Due to the inability to confidently enrich the positive control sgRNAs, inaccuracy of using sgRNA enrichment, and variability of sgRNAs, a natural conclusion is the ssCRISPR screen in combination with the GILA assay may not be an appropriate method to identify novel exon skipping events driving LUAD. However, as our computational analysis identified the in-frame candidates *MTORΔ12*, *BRAFΔ13*, and *NF1Δ46*, it calls for more stringent analysis using a single guide approach to cause exon skipping. This would allow us to control for sgRNA variability and use assays that permit more sensitive detection of a transformative phenotype.

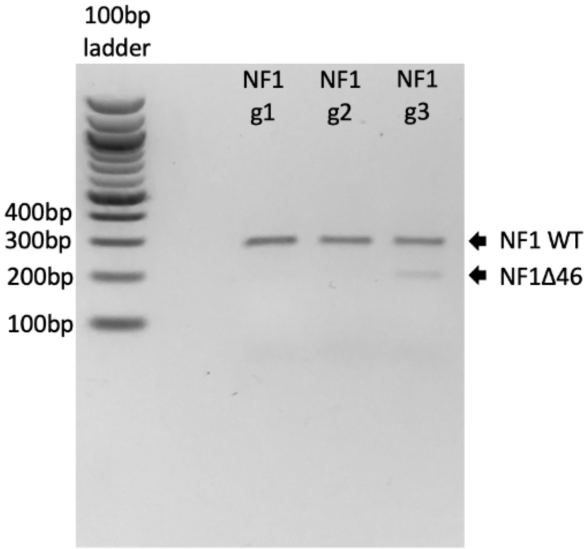


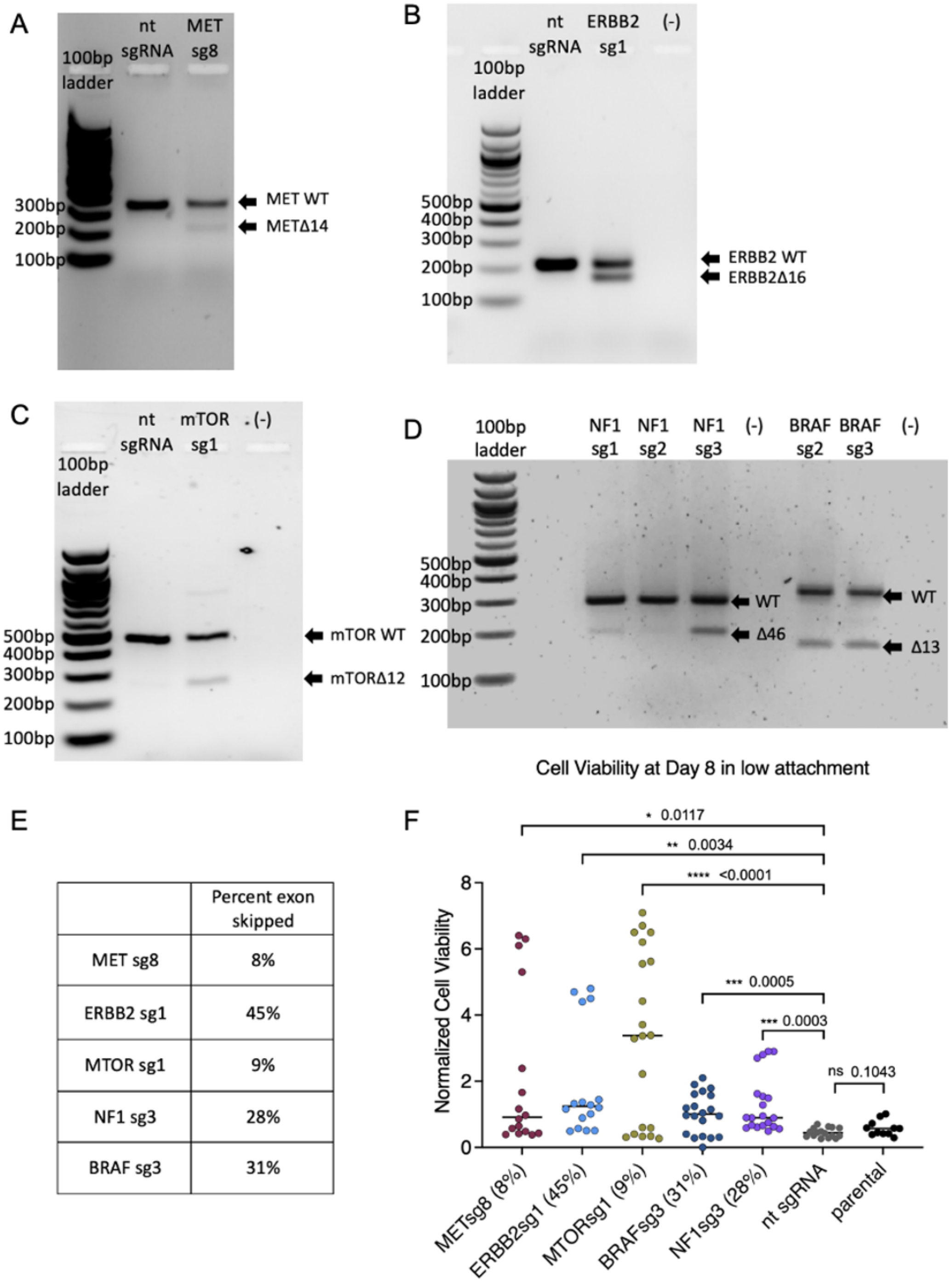
**Figure 2.7: Different sgRNAs have different editing efficiencies.**

Three sgRNAs targeted exon 46 of *NF1* (NF1 g1, NF1 g2 and NF1 g3) to force the production of *NF1Δ46.* I determined ratios of wild type (NF1 WT) and exon 46 skipped isoforms (NF1Δ46) using PCR and primers spanning exon 46.

*AALE ssCRISPR single targets: heterogeneous lines*

I used lentivirus to introduce sgRNAs targeting candidate exons of *MET*, *ERBB2*, *MTOR*, *BRAF* and *NF1* into AALE-Cas9 cells (ex. “MET sg8”). After puromycin selection, each resulting “heterogeneous” cell line expressed varying degrees of candidate exon skipping. Due to the error-prone repair process of non-homologous end joining (NHEJ), the resulting indels at splice sites vary depending on the cell. Therefore, in this heterogeneous cell line I anticipate cells with no splice site editing, resulting in 0% exon skipping, impactful splice site editing, resulting in near 100% exon skipping, and all values in between. Quantifying the amount of exon skipping across this heterogeneous population reflects the average expression of exon skipping events in these cell lines. Using RT-PCR and primers spanning the candidate exons, I identified exon skipping in each candidate (Figure 2.8A-D), and quantified the relative percent of the skipped isoform (Figure 2.8E). I found the sgRNAs had variable success in forcing exon skipping across these heterogeneous populations, resulting in as little as 8% candidate isoform for *METΔ14* to as much as 45% *ERBB2Δ16.*

I measured the oncogenic potential of these heterogeneous cell lines using a GILA assay coupled with a cell viability assay. Overall, these experiments revealed inconsistent cell viability in all heterogeneous ssCRIPSR lines, including the positive controls for targeting exons of *MET* and *ERBB2* (Figure 2.8F). However, all ssCRISPR lines resulted in significantly higher viability in low attachment compared to either the parental line or nt sgRNA negative control. Furthermore, both the Ras-MAPK candidates and *METΔ14* and *ERBB2Δ16* positive controls exhibited a similar pattern of variability, with some replicates associated with high viability, and some replicates associated with viability similar to the negative controls. Overall, this suggests that the Ras-MAPK candidates confer a selective advantage in low attachment, however, this proliferative advantage is likely correlated with the amount of candidate exon skipping, which. Because the majority of replicates function as negative controls, this prevents me from saying with confidence that these exon skipping events allow AALEs to proliferate in low attachment.



|  | Percent exon skipped |
|---|---|
| MET sg8 | 8% |
| ERBB2 sg1 | 45% |
| MTOR sg1 | 9% |
| NF1 sg3 | 28% |
| BRAF sg3 | 31% |

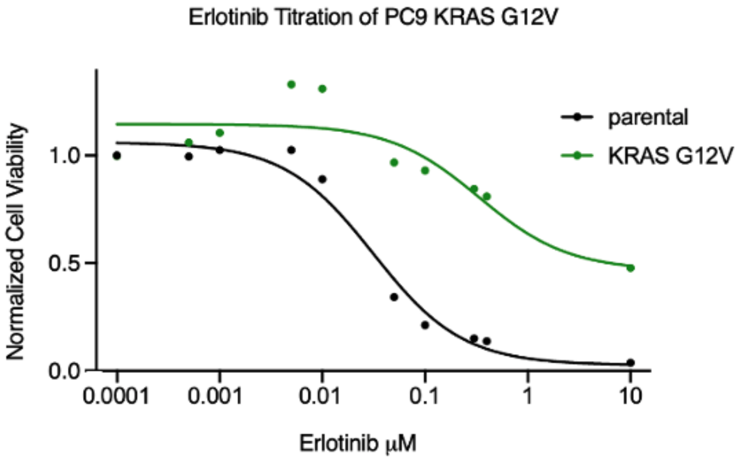
**Figure 2.8: Heterogeneous ssCRISPR-based expression in AALE cells leads to variable oncogenic readout in low attachment assay.**

ssCRISPR generated AALE heterogeneous lines for A) *METΔ14*, B) *ERBB2Δ16*, and C) *MTORΔ12*. D) Multiple ssCRISPR sgRNAs result in variable exon skipping when targeting *NF1* exon 46 and *BRAF* exon 13. I chose NF1 sg3 and BRAF sg3 for further analysis. E) Percent of exon skipped isoform in each cell line quantified from A-D using Image Studio Lite. F) GILA cell viability assay from a combination of 4-5 experiments. Cell viability at day 8 is normalized to values at day 0. Statistics calculated using a Mann-Whitney test. nt sgRNA = non-targeting sgRNA, and sg indicates “single guide” approach.

*PC9 ssCRISPR single targets: heterogeneous lines*

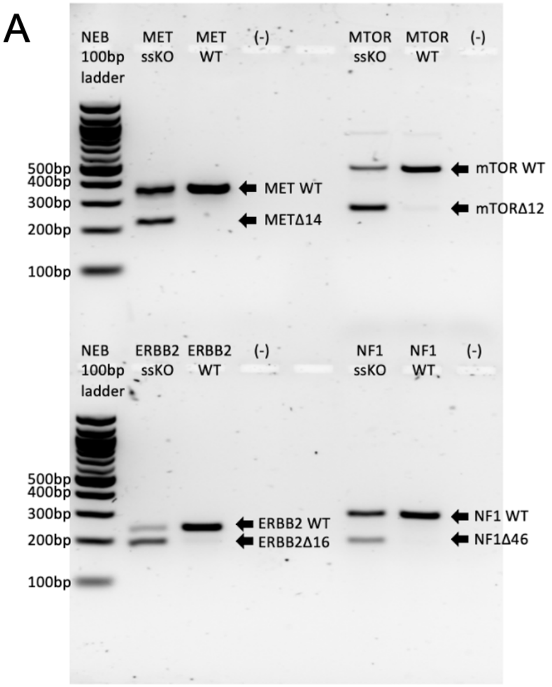
To assess the candidate’s oncogenic potential from another angle, I used an erlotinib-rescue assay developed for the PC9 lung adenocarcinoma line, which assays for activators of the Ras-MAPK pathway. PC9 cells harbor an activating mutation in EGFR, and thus are sensitive to EGFR inhibitors like erlotinib (Sharifnia et al. 2014; Berger et al. 2016). Expression of strong oncogenes like mutant KRAS re-activate downstream signaling, rescuing these cells from otherwise succumbing to the inhibitor (Figure 2.9).

We contracted Synthego to generate ssCRISPR PC9s using transient splice site-targeting sgRNAs and Cas9, which I refer to as “ssKO” for **s**plice **s**ite **k**nock**o**ut. Like the AALE ssCRISPR lines, these cell lines are heterogeneous. Due to technical errors, Synthego did not generate a PC9 *BRAFΔ13* ssKO line. I validated the amount of exon skipping in the Synthego-generated PC9 cell lines using RT-PCR and exon spanning primers, as done previously (Figure 2.10A). Overall each ssKO cell line expressed higher collective exon skipping than the AALE heterogeneous lines, ranging from 34% *NF1Δ46* for NF1 ssKO up to 76% *ERBB2Δ16* for ERBB2 ssKO (Figure 2.10B). Using these ssKO lines in the erlotinib-rescue I found CRISPR-based expression of *METΔ14*, *ERBB2Δ16*,



| | parental | KRAS G12V |
|---|---|---|
| IC50 | 0.03138 | 0.3357 |

**Figure 2.9: KRAS G12V overexpression rescues PC9 cells after erlotinib treatment.**
Increasing concentrations of erlotinib decreased cell viability of the parental PC9 cell line in a dose-dependent manner. PC9 KRAS G12V cells retained high cell viability through increasing concentrations of erlotinib. Raw cell viability measurements normalized to the lowest erlotinib concentration. IC50 represents the concentration of erlotinib required to inhibit viability by 50%.



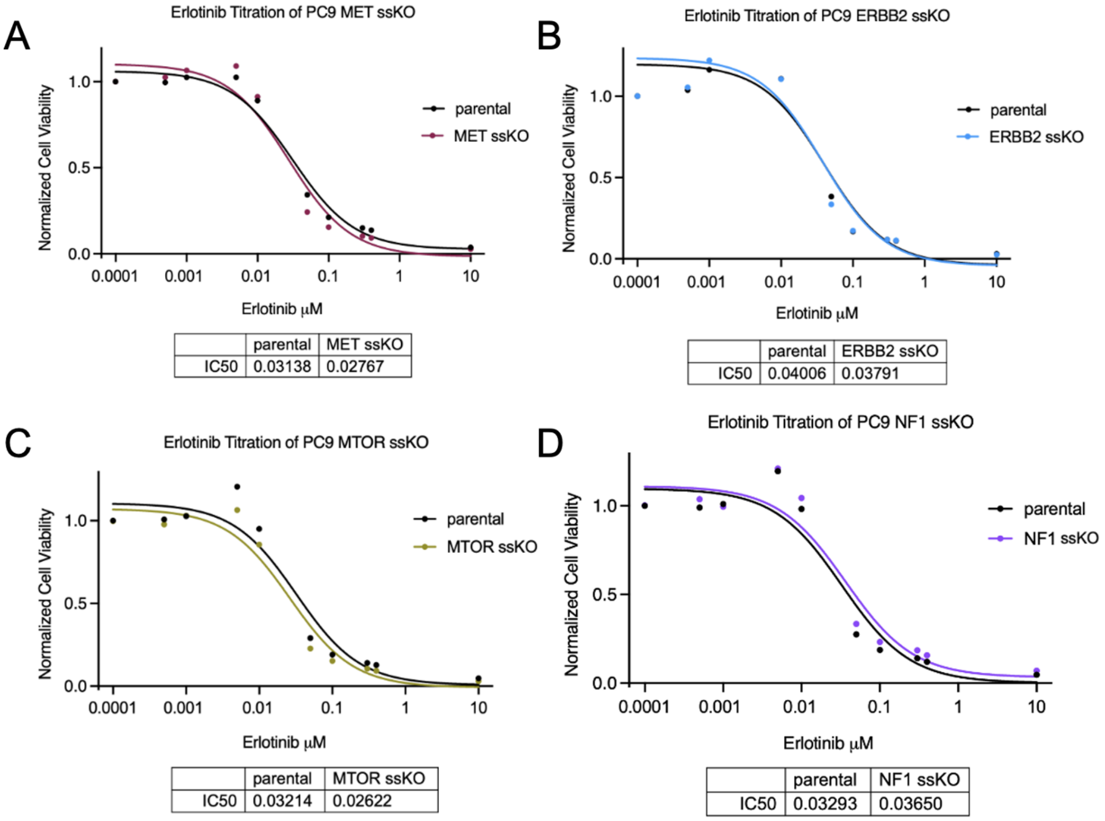
B

| | Percent exon skipped |
|---|---|
| MET ssKO | 42% |
| MTOR ssKO | 69% |
| ERBB2 ssKO | 76% |
| NF1 ssKO | 34% |

**Figure 2.10: RNA validation of exon skipping in PC9 ssKO cells.**

A) RNA validation using RT-PCR and exon spanning primers for each candidate exon. “WT” represents the wild type isoform including the candidate exon, and “Δ” exon delinates the skipped isoform. B) Percent candidate exon skipped from A calculated using Image Studio Lite.

*MTORΔ12* and *NF1Δ46* did not rescue the PC9 ssKO cells after erlotinib treatment (Figure 2.11). Because CRISPR-based expression of our positive controls, *METΔ14* and *ERBB2Δ16,* was not sufficient to rescue these cells, I conclude that this assay requires higher expression of oncogenes in order to activate the Ras-MAPK pathway in lieu of EGFR inhibition.



| | parental | MET ssKO |
|---|---|---|
| IC50 | 0.03138 | 0.02767 |

| | parental | ERBB2 ssKO |
|---|---|---|
| IC50 | 0.04006 | 0.03791 |

| | parental | MTOR ssKO |
|---|---|---|
| IC50 | 0.03214 | 0.02622 |

| | parental | NF1 ssKO |
|---|---|---|
| IC50 | 0.03293 | 0.03650 |

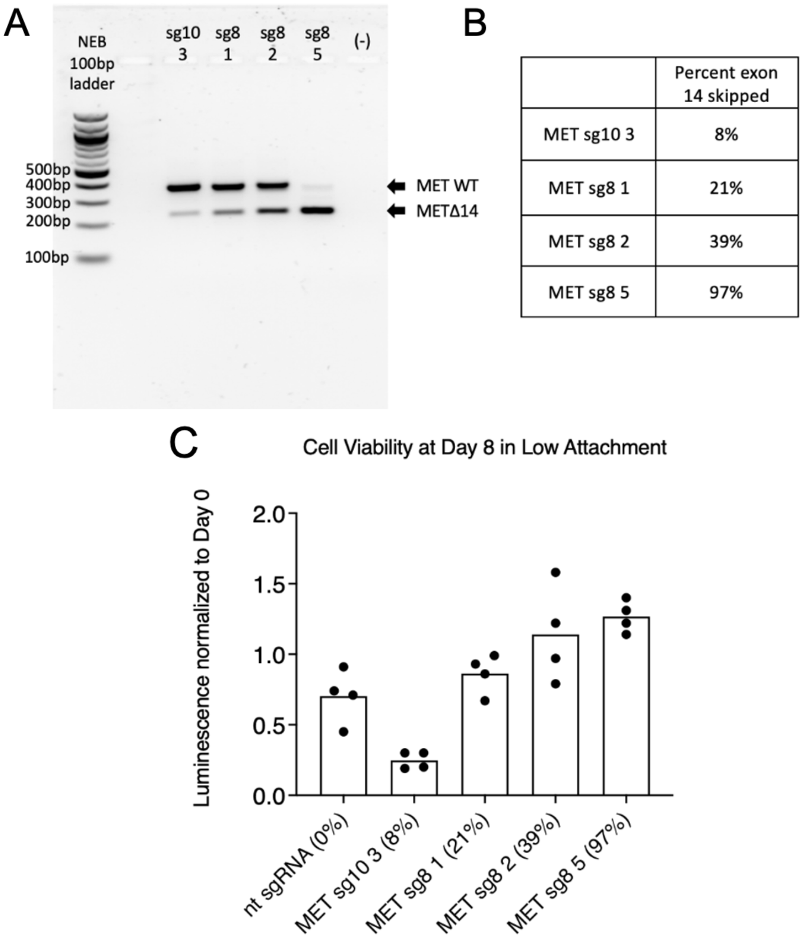
**Figure 2.11: CRISPR-based expression from PC9 ssKO cells unable to rescue cells after erlotinib treatment.**

Increasing concentrations of erlotinib decreased cell viability of both the PC9 parental line and PC9 ssKO A) MET B) ERBB2 C) MTOR D) NF1 in a dose-dependent manner. Raw cell viability measurements normalized to the lowest erlotinib concentration. IC50 represents the concentration of erlotinib required to inhibit viability by 50%.

The combined results from both the heterogeneous AALE ssCRISPR lines and PC9 ssKO lines suggest that the level of CRISPR-induced expression of the candidate exon skipping events is not sufficient for these assays. This suggests isolating clonal cell lines with near 100% candidate exon skipping is necessary to observe consistent phenotypic differences in these assays.

### *AALE ssCRISPR and Synthego-generated PC9 clones*

To alleviate variability and maximize the amount of isoform expressed, I isolated clonal cell lines from our heterogeneous cell lines. Beginning with the positive controls to serve as a benchmark for transformative phenotype, I isolated clonal *METΔ14* from the heterogeneous line (Fig 2.8A) using FACS single cell sorting followed by clonal expansion. I measured the expression of *METΔ14* in each cell line using RT-PCR (Figure 2.12A), with *METΔ14* quantified in figure 2.12B. Overall, I isolated a range of clonal cell lines with CRISPR-based expression of 8% *METΔ14* to 97% *METΔ14*. If success in low attachment growth is correlated with the amount of candidate exon skipping, I would anticipate increased cell viability through increasing amounts of *METΔ14* expression. I performed a GILA cell viability assay on these clonal AALE *METΔ14* cells (Figure 2.11C). Overall, across these AALE *METΔ14* clones there is an increase in cell viability as the amount of *METΔ14* expression increases. However, the cell viability from the nt sgRNA negative control is similar to METsg8 1, expressing 21% *METΔ14.* Additionally, METsg8 5 expressing 97% *METΔ14* produced similar



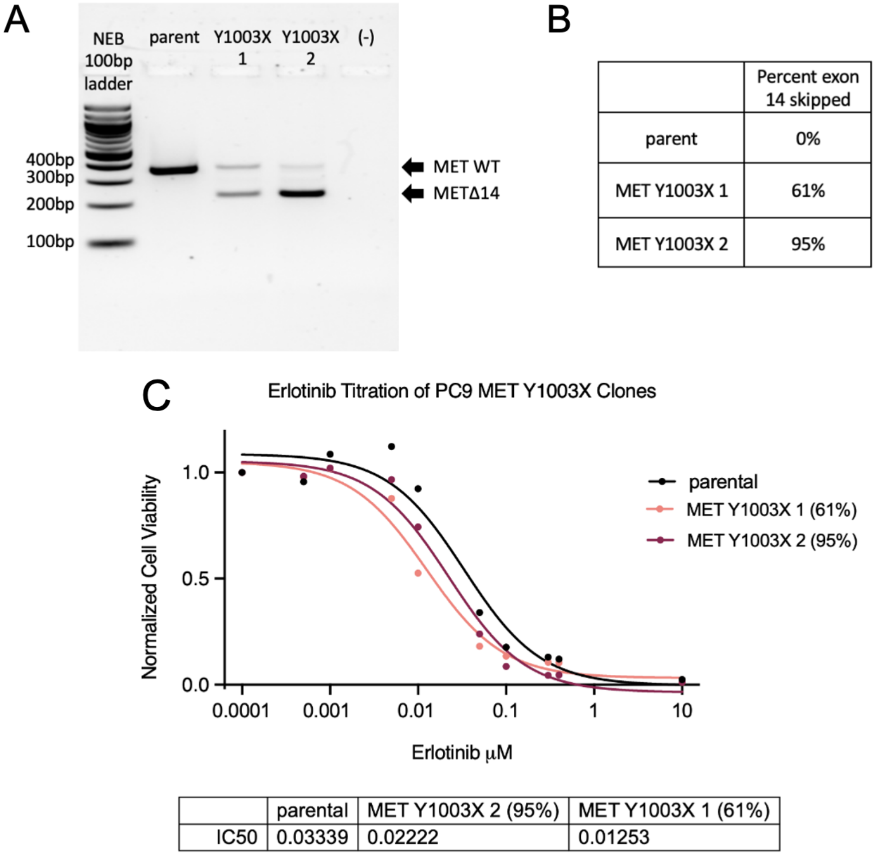
| | Percent exon 14 skipped |
|---|---|
| MET sg10 3 | 8% |
| MET sg8 1 | 21% |
| MET sg8 2 | 39% |
| MET sg8 5 | 97% |

**Figure 2.12: Increasing levels of *METΔ14* expression from AALE clones leads to a subtle increase in low attachment growth.**

A) four *METΔ14* clones (METsg10 3, METsg8 1, METsg8 2, and METsg8 3) with increased amounts of exon 14 skipping created with ssCRISPR. B) Quantification of percent skipped isoform from RT-PCR using Image Studio Lite. C) Quantification of cell viability from the AALE ssCRISPR *METΔ14* clones.

cell viability to a clone expressing 39% *METΔ14* (METsg8 2), whereas if success in low attachment was directly correlated with *METΔ14* expression I would expect much higher cell viability from METsg8 5. This highlights the background cell viability from our negative control, and suggests *METΔ14* may not be the sole reason these cells survived in low attachment.

I also obtained clonal PC9 MET Y1003X cell lines from Synthego, which express *METΔ14*. This mutation at position Y1003 is predicted to disrupt several splicing enhancer sequences within exon 14, forcing its exclusion from the mature mRNA (Cancer Genome Atlas Research Network ...). I have two MET Y1003X clones with varying amounts of exon 14 skipping (Figure 2.13A), ranging from 61% *METΔ14* (Y1003X 1) to 95% *METΔ14* (Y1003X 2) (Figure 2.13B). An erlotinib-rescue assay with these clones revealed that despite near 100% *METΔ14* from MET Y1003X 2, the CRISPR-based expression of *METΔ14* was not sufficient to rescue these cells (Figure 2.13C). I conclude that CRISPR-based



| | Percent exon 14 skipped |
|---|---|
| parent | 0% |
| MET Y1003X 1 | 61% |
| MET Y1003X 2 | 95% |

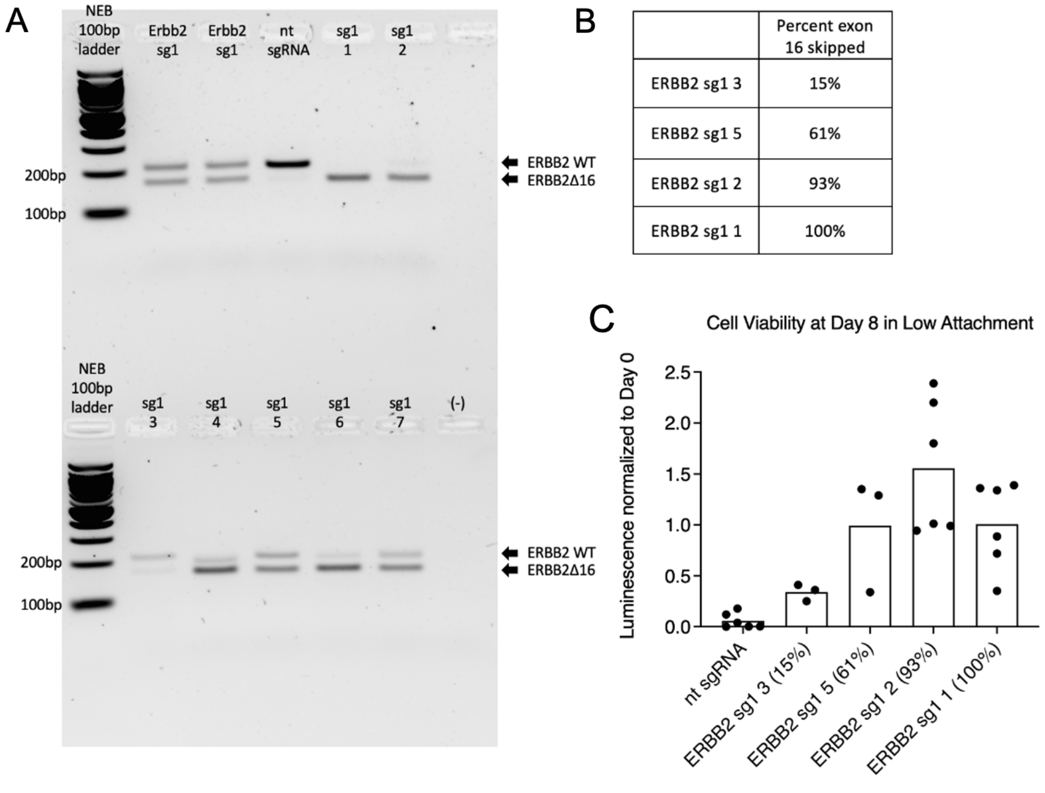
| | parental | MET Y1003X 2 (95%) | MET Y1003X 1 (61%) |
|---|---|---|---|
| IC50 | 0.03339 | 0.02222 | 0.01253 |

**Figure 2.13: CRISPR-based expression of clonal *METΔ14* is not sufficient to rescue PC9 MET Y1003X cells after erlotinib treatment.**

A) Two MET Y1003X clones with varying amounts of exon 14 skipping created with Synthego. B) Quantification of *METΔ14* from RT-PCR. C) Erlotinib-rescue assay using MET Y1003X clones.

expression of *METΔ14* is not sufficient for the PC9 erlotinib rescue assay, and future assays would require overexpression of *METΔ14.*

Next I isolated clonal AALE *ERBB2Δ16* cells from the heterogenous pool in figure 2.8B. RT-PCR and primers spanning exon 16 of ERBB2 identified ERBB2 clones with variable *ERBB2Δ16* expression (Figure 2.14A). I selected clones ranging from 15% to 100% *ERBB2Δ16* expression for further analysis (Figure 2.14B). A GILA cell viability assay with these clones revealed a correlation between *ERBB2Δ16* expression and proliferation in low attachment



| | Percent exon 16 skipped |
|---|---|
| ERBB2 sg1 3 | 15% |
| ERBB2 sg1 5 | 61% |
| ERBB2 sg1 2 | 93% |
| ERBB2 sg1 1 | 100% |

**Figure 2.14: Variable levels of ERBB2Δ16 from ssCRISPR based expression in AALE clones leads to subtle differences in low attachment growth.**

A) RT-PCR of *ERBB2Δ16* clones with varying amounts of exon 16 skipping created with ssCRISPR. B) Quantification of percent skipped isoform from RT-PCR for 4 clones with varying amounts of exon skipping. C) Quantification of cell viability from the ssCRISPR *ERBB2Δ16* clones.

as expected (Figure 2.14C). However, I also identified high variability in cell viability in replicates of clones expressing 61% *ERBB2Δ16* (ERBB2sg1 5) and 100% *ERBB2Δ16* (ERBB2sg1 1) (Fig. 2.14C), with some replicates correlating with the nt sgRNA values. This high variability in cell viability mirrors the variability in replicates from our AALE heterogeneous lines (Fig. 2.8F), and indicates that *ERBB2Δ16* expression is not solely responsible for the ability of these replicates to survive in low attachment.

I repeated the AALE clonal pipeline of isolation and analysis for all Ras-MAPK candidates. RT-PCR with exon spanning primers identified several clones with varied candidate exon skipping for *MTORΔ12* (Figure S2.1), *BRAFΔ13* (Figure S2.2)*,* and *NF1Δ46* (Figure S2.3). Selecting clones that range from the lowest to highest candidate exon skipping, I performed a GILA cell viability assay (Figure 2.15). Overall, the amount of candidate isoform expression did not correlate with success in low attachment. As an example, MTORsg1 14 (10 % *MTORΔ12)* retained the highest cell viability out of the *MTORΔ12* clones (Figure 2.15A). In the *BRAFΔ13* clones, the nt sgRNA had the highest cell viability, highlighting the high background of the negative controls in this assay (Figure 2.15B). Two *NF1Δ46* clones expressed similar levels of *NF1Δ46*, NF1sg3 13 (71% *NF1Δ46*) and NF1sg3 18 (73% *NF1Δ46*), however they

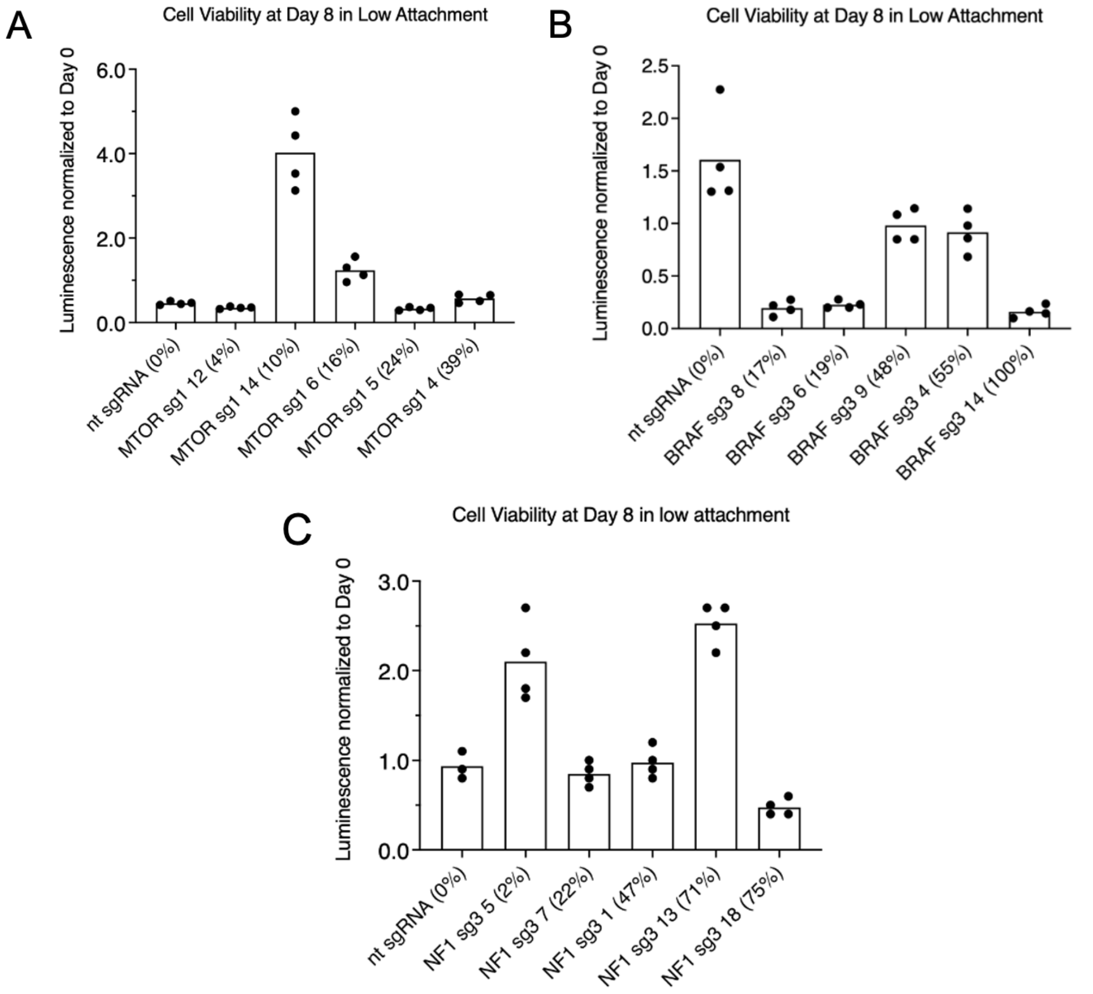


**Figure 2.15: RAS Candidate AALE clones' ability to survive in low attachment not correlated with amount of isoform present.**

GILA Cell Viability assays of AALE clones for A) MTORΔ12 ) BRAFΔ13 and C) NF1Δ46. Amount of candidate isoform on increases from 0% (nt sgRNA) to clone with maximum amount of candidate exon skipping.

displayed opposite viability phenotypes in the assay (Figure 2.15C). Based on these results, it is unlikely that the Ras-MAPK candidates contribute to oncogenic transformation. However, due to the high background of our negative controls masking the transformative phenotype of the positive controls (Figure 2.12C), our results on the Ras-MAPK candidates are inconclusive. I suggest that CRISPR-based expression of these aberrant splicing events combined with the GILA and erlotinib-rescue assay are not appropriate methods to measure the oncogenic potential of these splicing variants.

### *Discussion*

Overall, CRISPR-based expression of isoforms was not sufficient to observe a consistent and obvious phenotypic effect in these assays, even with the positive controls *METΔ14* and *ERBB2Δ16*. As touched on previously, ssCRISPR itself as a tool has variable success. One source of variability from CRISPR is based on intrinsic factors of the sgRNA sequence (Concordet and Haeussler 2018), which consequently cause different sgRNAs targeting the same exon to have variable success in forcing exon skipping. I observed this effect in the screen using 5' and 3' splice site-targeting sgRNAs of *METΔ14,* where the 3' sgRNAs moderately enriched in low attachment while the 5' sgRNAs did not (Fig. 2.4D). Interestingly, while the 5' sgRNAs failed to enrich in the screen, I used one of these 5' sgRNA in the single target analysis to produce clonal 97% *METΔ14* (Figure 2.12 - METsg8 5). This further highlights the variability of ssCRISPR, as the same sgRNA resulted in different phenotypes in low attachment. This sgRNA variability is further exemplified by Figure 2.7 revealing 2/3 sgRNAs targeting *NF1Δ46* did not produce exon skipping. This evidence suggests that several sgRNAs in the ssCRISPR screen were likely non-functional, and therefore exist as nt sgRNAs with no bearing on transformation. In a published screen using splice site targeting guide RNAs to force exon skipping, a pair of guide RNAs (pgRNA) were used to effectively cut out the splice site and surrounding regions providing more confidence in the ability to skip exons in a screen format (Thomas et al. 2020). These pgRNAs also accounted for variable success in the ability of CRISPR to edit a location, as there is the opportunity for the other guide RNA within the pair to edit the second locus surrounding the splice site. For future screens focused on the functions of skipped exons, I recommend using pgRNAs to limit the on-targeting variability of a single guide RNA approach.

For the single target studies of *METΔ14, ERBB2Δ16, MTORΔ12, NF1Δ46,* and *BRAFΔ13,* I accounted for the variability in CRISPR editing efficiency, as I individually assessed the ability of these sgRNAs to force exon skipping events. In the example referenced in Figure 2.7 showing most sgRNAs failed to produce *NF1Δ46*, with the single target pipeline I chose the single functioning sgRNA, NF1 sg3, for future experiments. However, while I could control for the binary functionality of Cas9 to force exon skipping, I could not control for the indels generated at splice sites that would dictate the impact on exon skipping. I observed the consequences of this by assessing AALE heterogeneous cell lines in low attachment, and found a striking difference in replicate performance in each ssCRISPR lines (Fig. 2.8F). This variability was specific to cell lines expressing sgRNAs that target the genome; the non-targeting sgRNA reflected a similar viability to the parental line. Given that ssCRISPR-based expression of oncogenes is not sufficient to measure oncogenicity in these assays, it is impossible to draw a conclusion on the oncogenic potential of the Ras-MAPK candidates. Assuming the Ras-MAPK candidates do not confer a proliferative advantage, the high viability in some of these replicates could be the result of the catalytic activity of Cas9 on the mammalian genome, as opposed to the genotype resulting from splice site edit.

This variability in the heterogenous lines inspired us to generate clonal cell lines, as I hypothesized this would ameliorate the fickle behavior of the heterogeneous cells in low attachment. While clonal selection successfully isolated cell lines with near 100% candidate exon skipping for both *METΔ14* and *ERBB2Δ16*, these clones resulted in positive yet inconsistent success in these assays. For the *METΔ14* clones, the nt sgRNA resulted in similar cell viability to METsg8 1 (21% *METΔ14*), highlighting the background high viability of the negative controls. Additionally, despite obvious differences in *METΔ14* expression, METsg8 5 (97% *METΔ14*) and METsg8 2 (39% *METΔ14*) resulted in similar cell viability, suggesting that METsg8 2 did not solely rely on *METΔ14* expression to survive in low attachment (Fig. 2.12C). For the *ERBB2Δ16* clones, I observed a general association of *ERBB2Δ16* expression and cell viability as expected. However, I found a striking inconsistency of replicate viability with clones expressing 61% *ERBB2Δ16* and 100% *ERBB2Δ16* (Fig. 2.14C), with some replicates correlating with the nt sgRNA values. This is reminiscent of the replicate variability within the heterogeneous cell lines (Fig. 2.8F). There was no correlation with the amount of Ras-MAPK candidate exon skipping and ability to survive in low attachment (Fig. 2.15). If I assume that *MTORΔ12*, *BRAFΔ13* and *NF1Δ46* are not functionally impactful skipping events, then Fig 2.15 is an illustration of the variation of the AALE parental line itself. In hindsight, it would have been ideal to clonally isolate the AALE parent line before performing ssCRISPR. In conclusion, our AALE model combined with the GILA assay was not an ideal method to study the subtle effects of ssCRISPR-induced splicing events.

The variability that afflicted these assays would be ameliorated if the amount of isoform produced with CRISPR-based expression was sufficient. This is also exemplified by the PC9 studies, which revealed no ssCRISPR-expressed isoforms rescued PC9 cells after EGFR inhibition. As opposed to the AALE low attachment assay which is more adept at detecting small changes in viability, the PC9 assay requires a certain threshold of oncogene expression to see rescue. Because I did not see rescue after expression of 95% *METΔ14* from PC9 MET Y1003X clone 2 (Fig. 2.13C), I determined that CRISPR-based expression of this isoform was not sufficient for this assay. For the *METΔ14* example, there are two possible solutions to this problem. First, MET activation is dependent on its ligand, HGF, which was used in previous studies generating *METΔ14* with CRISPR to activate the receptor (Wang et al. 2022; Togashi et al. 2015; Lu et al. 2017). Because I could not preferentially activate MET in our ssCRISPR screen format, I avoided HGF for our experiments. I also avoided HGF in our single target studies to ensure the results between *METΔ14, ERBB2Δ16,* and the Ras-MAPK candidates were equal. Second, previous research demonstrated that *METΔ14* overexpression successfully rescued PC9s after EGFR inhibition (Suzawa et al. 2019). Additionally, previous research on *ERBB2Δ16* also used *ERBB2Δ16* overexpression to study this oncogene in both breast and lung cancer models (Smith et al. 2020; Turpin et al. 2016). In the published CRISPR screen using pgRNAs to force exon skipping, the readout was poison exon inclusion and subsequent pgRNA dropout (Thomas et al. 2020), suggesting a strong phenotype like LOF instead of subtle GOF is an ideal use for CRISPR-based expression. Furthermore, previous work with the GILA assay in a screen format measured success in low attachment from overexpressed oncogenes, suggesting that overexpression may be necessary to see enrichment in this assay  (Rotem et al. 2015)

Because overexpression experiments typically express the longest coding sequence, we were drawn to using CRISPR-based expression. We determined that the exon skipping events may not be impactful when expressed on the longest coding isoform, because about 50% of transcripts dominantly expressed in tissues are not the longest coding sequence (Gonzàlez-Porta et al. 2013). CRISPR-based expression capitalizes on the cell's inherent gene regulation to produce the exact isoforms in which these exons skipping events would be functionally impactful. However, given the variability and subtle changes in isoforms generated from using CRISPR-based expression in these assays, I conclude that overexpression is required to functionally characterize these exon skipping events. Due to the variability of this work, I could not confirm the tumorigenic potential of *MTORΔ12*, *BRAFΔ13* and *NF1Δ46*, although based on the AALE clonal analysis it seems unlikely. I believe I could achieve a binary answer if I repeated this screen overexpressing these variants instead. Based on previous research overexpressing *METΔ14* and *ERBB2Δ16* (Suzawa et al. 2019; Turpin et al. 2016; Smith et al. 2020) it appears overexpression of these isoforms is the best way to tease out their oncogenic potential.

## Chapter 2 Material and Methods

*Identification of candidate aberrant splicing events*

Whole exome (DNA) and matched RNA-seq data was used from a cohort of 495 LUAD patients from The Cancer Genome Atlas (TCGA). Splice site alterations in whole exome data were defined as mutations 3bp in the exon or 30bp in the adjacent exon around the splice site (33bp window). JuncBase (Brooks et al. 2011) was used to identify skipped exons in the RNA associated with splice site alterations due to its ability to detect novel splice variants. JuncBase outputs a percent spliced in (PSI) value, which corresponds to how often a single splicing event occurred relative to other mutually exclusive splicing events. An 'aberrant exon skipping event' in this analysis indicated the PSI value was 3 standard deviations below the mean. Using JuncBase and custom python scripts this analysis resulted in 635 candidate exon skipping events.

In addition to these splice site mutations, exon skipping events from splicing factor mutations (U2AF1 S45F and RMB10 loss of function (LOF)) and samples with no previously identified oncogenic driver were incorporated into this pool of candidate exons. For splicing factor mutations, alternative splicing variants quantified by JuncBase were compared between patients with the mutations against the patients without the mutations. For U2AF1 S34F, 11 patients containing this hotspot mutation were compared to 451 completely lacking any splicing factor mutations. For RMB10 LOF, 28 patients containing this mutation were compared to the 451 patients with no splicing factor alterations. These significantly different splicing events from this analysis were compared with the Wilcoxon rank sum test, corrected for false discovery rate (FDR <5%)

using the Benjamini–Hochberg method, and splicing events filtered using a δPSI > 10%. This resulted in 94 exon skipping events associated with U2AF1 S34F and 15 associated with RMB10 LOF. To identify exon skipping events in oncogene driver negative tumors, 106 samples with no identified driver were compared to the 389 patients with oncogenic drivers using the Wilcoxon rank sum test and Benjamini–Hochberg correction method. This resulted in 50 exon skipping events from driver negative tumors. In addition to these candidates identified from our analysis, we also obtained 200 candidate exon skipping events from Guardant Health via their proprietary parameters to bolster the candidate exons in our pool.

An sgRNA library was created to target these candidate exons in a pooled format. This was done using the CRISPOR guide RNA design tool to design sgRNAs most likely to disrupt exon splice sites while minimizing off target effects (Haeussler et al. 2016). Three sgRNAs were designed to target both the 5' and 3' splice site for each candidate exon. In cases where sgRNAs did not fit required criteria (no sgRNAs available in close proximity to splice site and high off target cutting prediction), less than three sgRNAs were chosen. Three different types of negative controls were used in this assay: sgRNAs targeting the first exon of each candidate gene, one sgRNA targeting the middle of the exon, and 150 non-targeting (nt) sgRNAs. The sgRNAs targeting the first exon were obtained from a previously designed CRISPR knockout screen targeting genes in the human genome. The sgRNA targeting the center of the exon was not predicted to result in skipping, and thus was used as a negative control in this assay. As for the nt sgRNAs, these were not anticipated to target the human genome, so they were anticipated to reflect background enrichment in the assay. Positive controls were sgRNAs targeting the splice sites of *MET* exon 14 and *ERBB2* exon 16, as these are previously identified events that drive lung cancer.

*Cell lines*

AALE cells and MET 1F10 clone were obtained from Eric Collison (UC San Francisco), and PC9 cells were obtained from Alice Berger (Fred Hutchinson Cancer Center). To get stable Cas9 expression in AALEs, cells were infected with a Cas9 expressing lentiviral vector (gift from Susan Carpenter, UC Santa Cruz). Cells were selected with Blasticidin for 2 weeks. AALE cells were cultured in SAGM Small Airway Epithelial Cell Growth Medium BulletKit (CC-3118, Lonza) using the ReagentPack Subculture Reagents (CC-5034, Lonza). PC9 cells were cultured in RPMI 1640 medium (11875093, ThermoFisher Scientific) with 10% FCS (26140079, ThermoFisher Scientific). Tells where grown under constant 37°C and 5% $CO_2$.

*Cloning sgRNA library*

The sgRNA library was cloned using the Agilent SureVector System.

*Validation of cloned ssCRISPR sgRNA library distribution*

Guides were prepared for sequencing by PCR amplification using Titanium Taq (639210, Takara):

i5_Forward_1:
5’-AATGATACGGCGACCACCGAGATCTACACAGCGCTAACACTCTTTCCCTAC
ACGACGCTCTTCCGATCTtcttgtggaaaggacgaaaca -3’

i7_Reverse_1:
5’-CAAGCAGAAGACGGCATACGAGATACCGCGGGTGACTGGAGTTCAGACG
TGTGCTCTTCCGATCTactttttcaagttgataacggactagc -3’

Using these primers which partially anneal to the lentivirally-integrated region surrounding the sgRNA, these amplicons contain all indexes and annealing sites for short read sequencing. These amplicons were run on a 1% agarose gel, and gel isolated using the NucleoSpin Gel and PCR Clean-up kit (740609.50, Macherey-Nagel). These isolated amplicons were spiked into a MiSeq run and sequenced at a coverage of 100 reads per sgRNA. The sgRNAs were quantified using a custom python script.

*Validation of Cas9 via CD44 knockdown*

Lentivirus containing an sgRNA targeting the first exon of CD44, as well as a negative control sgRNA which does not target the genome, was gifted to the project by Susan Carpenter. These sgRNAs were infected into AALE-Cas9 cells (stably expressing Cas9), and after a 24-hour infection and 24-hour recovery, 1ug/ml of puromycin was used to select the cells for 4 days. After selection, 1ml of these cells were collected (~250-500k cells) and cells were washed in FACS buffer (PBS + 2% FCS). After supernatant was removed via aspiration, the cell pellet was resuspended in 1ml FACS buffer and 10ul of suspension was removed for counting and volume adjusted to 100ul. Antibodies were diluted to 1:100 (MHCD4401 human CD44 FITC conjugate (Life Tech) or Ms IgG2b Isotype Control (PAS-3901F, ThermoFisher Scientific) by adding 1ul antibody in the 100ul cell dilution, then incubated on ice for 1 hour. After incubation, cells were washed with 1ml FACS buffer 2x and pelleted by centrifugation at 650xg for 3 minutes. Final volume was adjusted 500ul in FACS buffer in a FACS tube, then FACSed for GFP.

*Mixing Experiment*

AALE MET 1F10 cells and AALE parental lines were grown to 70% confluency, trypsinized, then mixed in the following proportions: 100% MET 1F10 and 0% parent line, 50% MET 1F10 and 50% parent line, 10% MET 1F10 and 90% parent line, and 0% MET 1F10 and 100% parent line. These mixed populations were plated in both a tissue-treated 6 well plate ("high attachment") and a matched low-attachment 6 well plate. Cells were grown undisturbed for 8 days, then RNA was isolated via TriReagent (T9424-200ML, Sigma Aldrich) and the Direct-zol RNA Miniprep Kit (R2050, Zymo Research). cDNA was generated using the High-capacity cDNA Reverse Transcription kit (4368814, Thermo Fisher Scientific), and a MET splice PCR was performed using HotStart ReadyMix PCR kit (Kapa Biosystems, KK2602) and primers spanning exon 14:

MET_splice_Forward: 5' – TGGGTTTTTCCTGTGGCTGA – 3'
MET_splice_Reverse: 5' – GGGCCCAATCACTACATGCT– 3'

Resulting amplicons were run on a 2% agarose gel and visualized with Licor D-DiGit scanner. Band intensities were quantified using Image Studio Lite, and Percent skipped was calculated as:

Percent skipped = (skipped exon 14 amplicon/ (skipped exon 14 amplicon + included exon 14 amplicon)) x 100

*<u>ssCRISPR screen in low attachment</u>*

Screen virus was obtained from the UCSC CRISPR Core. AALE cells were thawed and expanded in a 15 cm tissue culture plate (Corning - 430599). The screen virus was added at an MOI of 0.3 and allowed to transfect for 24 hours. Virus was removed, cells were washed with DPBS, and fresh media was added for cells to recover for 24 hours. AALE cells were selected with 1ug/mL puromycin for 4 days, then puromycin was removed and cells were allowed to recover in fresh media for 24 hours. This infection was expanded to roughly 100 million cells across multiple 15 cm dishes. After trypsinization, the cell suspension was counted to add 600,000 cells in each T75 flask - a value determined to be an optimal plating density for this cell line. Per each growth condition, cells were plated into 18 tissue treated T75 flasks (Corning - 430641U) or 18 Ultra Low Attachment T75 flasks (Corning - 3814). This was to ensure a representation of 1000x coverage of each guide in each growth condition. This plating was repeated for each time point (Day 3, Day 8 and Day 15). The remaining 20 million cells were washed with DBPS and pelleted for future use as the early Day 0 time point (store -20$^{O}$C). For each timepoint, cells were pelleted, washe, and stored at -20$^{O}$C.

Genomic DNA pellets were extracted using the DNEasy kit. sgRNAs were amplified using a nested PCR strategy. Round 1 PCR to amplifies guides from the genome. The entirety of the isolated genomic DNA was used in excess in each PCR reaction in a max volume of 100ul using Titanium Taq for 15 cycles and below primers:

Sure_vector_Forward: 5'- AGGGCCTATTTCCCATGATTCC -3'
Sure_vector_Reverse: 5'- CACATGCATGGCGGTAATACG -3'

This resulting PCR reaction was used as input for the Round 2 PCR to add the sequences necessary for next generation sequencing. The entirety of the round 1 PCR reaction was used in excess in each Round 2 PCR reaction for a max volume of 100ul using Titanium Taq and 15 cycles. The above primers (i5_Forward_1 and i7_Reverse_1) were used, in addition to below primers to index all seven screen timepoints:

i5_Forward_2:

5’-AATGATACGGCGACCACCGAGATCTACACGATATCGACACTCTTTCCCTAC
ACGACGCTCTTCCGATCTtcttgtggaaaggacgaaaca -3’

i7_Reverse_2:
5’-CAAGCAGAAGACGGCATACGAGATGTTATAAGTGACTGGAGTTCAGACGT
GTGCTCTTCCGATCTactttttcaagttgataacggactagc -3’

i5_Forward_3:
5’-AATGATACGGCGACCACCGAGATCTACACCGCAGACACACTCTTTCCCTA
CACGACGCTCTTCCGATCTtcttgtggaaaggacgaaaca -3’

i7_Reverse_3:
5’-CAAGCAGAAGACGGCATACGAGATCAAGTCCGTGACTGGAGTTCAGACG
TGTGCTCTTCCGATCTactttttcaagttgataacggactagc -3’

Different combinations of these primers were used to uniquely dual index each timepoint. Resulting amplicons were run on a 1% agarose gel, and gel isolated using the NucleoSpin Gel and PCR Clean-up kit (740609.50, Macherey-Nagel). Isolated amplicon concentrations were quantified via Qubit 3.0 Fluorometer and validated for purity using the Agilent TapeStation 4150 system. Amplicons were sequence validated via Sanger sequencing to ensure all indexes and primer binding sites were correct. All samples were pooled and sequenced using the UC Davis sequencing core at a coverage of 1000 reads per sgRNA.

*ssCRISPR screen in analysis*

sgRNAs were quantified using a custom python script. For the 150 non-targeting sgRNAs, raw sgRNAs counts were normalized to the library median for comparison across samples. To determine sgRNA enrichment over time, these median-normalized values at each time point were normalized to the median-normalized sgRNA counts at Day 0:

non-targeting sgRNA value = (raw sgRNA count at time point/median sgRNA count at time point)/(raw sgRNA count at day 0/median sgRNA count at day 0)

For the positive control sgRNAs, raw sgRNAs counts were normalized to the **non-targeting median** at that time point. This was to determine enrichment in comparison to negative control. To determine sgRNA enrichment over time, these normalized values at each time point were normalized to the normalized sgRNA counts at Day 0 (see below). These values were plotted in Prism.

positive control sgRNA value = (raw sgRNA count at time point/non-targeting sgRNA median at time point)/(raw sgRNA count at day 0/non-targeting sgRNA median at day 0)

*ssCRISPR single target cell line generation*

The UCSC CRISPR core both cloned the sgRNAs and generated lentivirus. Splice site targeting sgRNA sequences are below:

METsg8: 5'- TACCGAGCTACTTTTCCAGA -3'

METsg10: 5'- TACCTTCTGGAAAAGTAGCT -3'

ERBB2sg1: 5'- TGTGTGGACCTGGATGACAA -3'

MTORsg1: 5'- TGGGATAACAGATCCTGGTA -3'

BRAFsg3: 5'- ATTGCACGACAGACTGCAC -3'

NF1sg3: 5'- AAAAATTCTGTTTTCCTAAA -3'

200ul of sgRNA lentivirus was added to a 24 well plate of AALEs at a confluency of 40-50%. Cells were incubated with virus for 24 hours. After transfection, the virus was removed, cells were washed with DPBS, and fresh media was added for cells to recover for 24 hours. AALE were selected with 1ug/mL puromycin for 4 days, then puromycin was removed and cells were allowed to recover in fresh media for 24 hours.

*ssCRISPR single target RNA Validation*

RNA was isolated via TriReagent (Sigma-Aldrich, T9424) and the Zymo Direct Zol RNA Miniprep kit (Zymo Research Corporation, R2050). Subsequent cDNA prep was performed using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, 4368814), and resulting cDNA was used as a PCR template using the HotStart ReadyMix PCR Kit (Kapa Biosystems, KK6202). Exon spanning primers were ordered from IDT:

MET_splice_F: 5'- TGGGTTTTTCCTGTGGCTGA -3'
MET_splice_R: 5'- GGGCCCAATCACTACATGCT -3'

ERBB2_splice_F: 5'- GATGAGGATCCCAAAGACCA -3'
ERBb2_splice_R: 5'- CGGTGTGAAACCTGACCTCT -3'

MTOR_splice_F: 5'- ATCACTCTTGCCCTCCGAAC-3'
MTOR_splice_R: 5'- GGCCAGGTGTGCATCAAAG-3'

BRAF_splice_F: 5'-CCAGCTTGTATCACCATCTCC -3'
BRAF_splice_R: 5'-CTGTTCAAACTGATGGGACCC -3'

NF1_splice_F: 5'- CTCTTGTTGTCTTTGGGTGTATTAG-3'
NF1_splice_R: 5'- GGAGACTATCTAAAGTATGCAGGTT-3'

Resulting amplicons were run on a 1.2-2% agarose gel until wild type and exon-skipped isoforms were resolved. Relative isoform abundance ratios were quantified using Image Studio Lite.

*AALE - GILA Cell TIter Glo assay*

AALE ssCRISPR lines (both heterogeneous and clonal lines) were grown to ~80%. Plates were trypsinized and 2,500 cells in a volume of 100ul were seeded per well into a 96 well low attachment plate (Corning, 3474). Four replicates were used per cell line. Additionally seed 2,500 cells per well (100ul) using 4 replicates of each cell line in a white, flat bottom 96 well assay plate (Thomas Scientific, 290-8027-W1F). It's important that there is one well of space in between each replicate, as to avoid fluorescence bleed over during measurement. Low attachment plate left in a 37°C incubator for 8 days.

For the cells in the white plate, the 'Day 0' measurement, we assessed cell viability in order to normalize Day 8 to the initial plating concentration. First, Cell Titer Glo was allowed to come to room temperature (Promega, G7572). Then, 1:1 ratio of Cell Titer Glo to the cells was added using a multichannel pipette, so the total volume was 200ul. Plates shaken 2 minutes by hand to lyse the cells, then mixed with single channel pipette by pipetting 4 -5 times. Plate left undisturbed and sheltered from light for 20 minutes, then emitted fluorescence was measured using the Varioskan LUX microplate reader (Thermo Fisher, VL0000D0). This provided raw values of fluorescence proportional to the amount of ATP produced, and served as a normalization value for Day 0 per cell line.

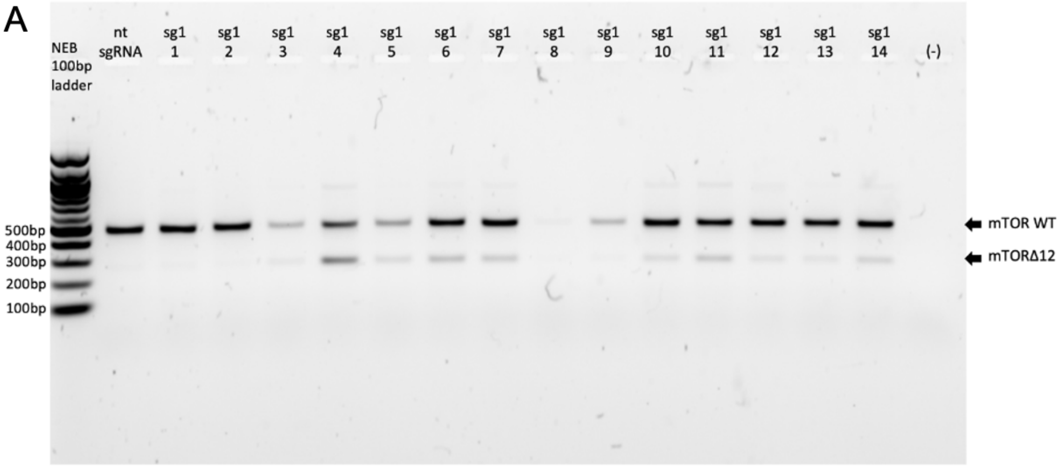
After 8 days, both Cell Titer Glo and the GILA plate were allowed to come to room temperature, and lysis protocol repeated as above. Then the lysed cells were transferred to the white assay plate, and during this transfer cells were mixed via pipetting 4-5 times. As done with Day 0, all replicates were spaced in a checker-board pattern leaving a one blank well in between each replicate. Like with Day 0, cells incubated for 20 minutes undisturbed and covered from light, then fluorescence measurement was taken. These raw values were normalized to the average of the four Day 0 replicates to compare cell viability between cell lines.

*PC9 - EGFR inhibitor titration assay*

PC9s were grown to 70% confluency, then cells were trypsinized and added at 9,000 (100µl) cells per well in 48/96 wells of a 96 well tissue treated plate

(Corning, 3598). This number of wells accounts for four replicates of each inhibitor concentration. Cells were allowed to adhere to the plate in a 37°C incubator for 24 hours. The next day, varying concentrations of erlotinib (Selleck Chemicals, S1023) were diluted in PC9 growth medium: 0μM (1:1000 DMSO), 0.0001μM, 0.0005μM, 0.001μM, 0.005μM, 0.01μM, 0.05μM, 0.1μM, 0.3μM, 0.4μM, 1μM, and 10μM. 200μl of the diluted inhibitor or DMSO was added to their corresponding wells, and cells were incubated for 4 days at 37degC. After four days cell viability was measured using Cell Titer Glo as done previously. Briefly, a multichannel pipette was used to add a 1:1 ratio of Cell Titer Glo reagent to cells, plate was shaken for 2 minutes to lyse cells, and the plate was allowed to incubate at room temperature undisturbed and protected from light. Then the fluorescence was measured using the Varioskan to get raw cell viability values for each replicate. Before importing this data into Prism, each replicate was normalized to the lowest possible inhibitor value, 0.0001μM. Once in prism plotting values in an XY format, log-transform inhibitor concentrations then perform a nonlinear regression using log(inhibitor) vs response (three parameters) to visualize data.

## Chapter 2 Supplementary Figures

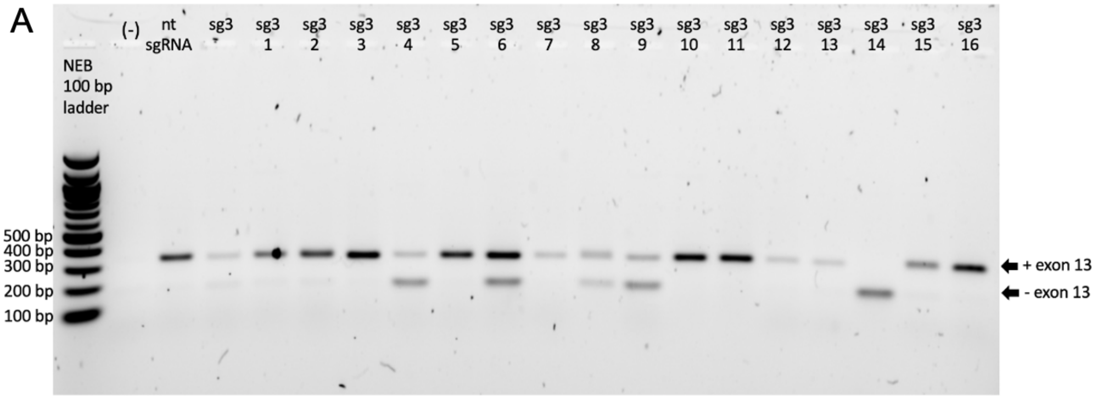


B

| | Percent exon 12 skipped |
|---|---|
| MTOR sg1 12 | 4% |
| MTOR sg1 14 | 10% |
| MTOR sg1 6 | 16% |
| MTOR sg1 5 | 24% |
| MTOR sg1 4 | 39% |

**Figure 2.S1: RNA validation of *MTORΔ12* in AALE MTOR clones.**

A) RT-PCR and exon-spanning primers were used to validate and quantify the isoform ratios of MTORΔ12 in the AALE MTOR clones. B) MTORsg1 clones 12, 14, 6, 5 and 4 were selected for further analysis.

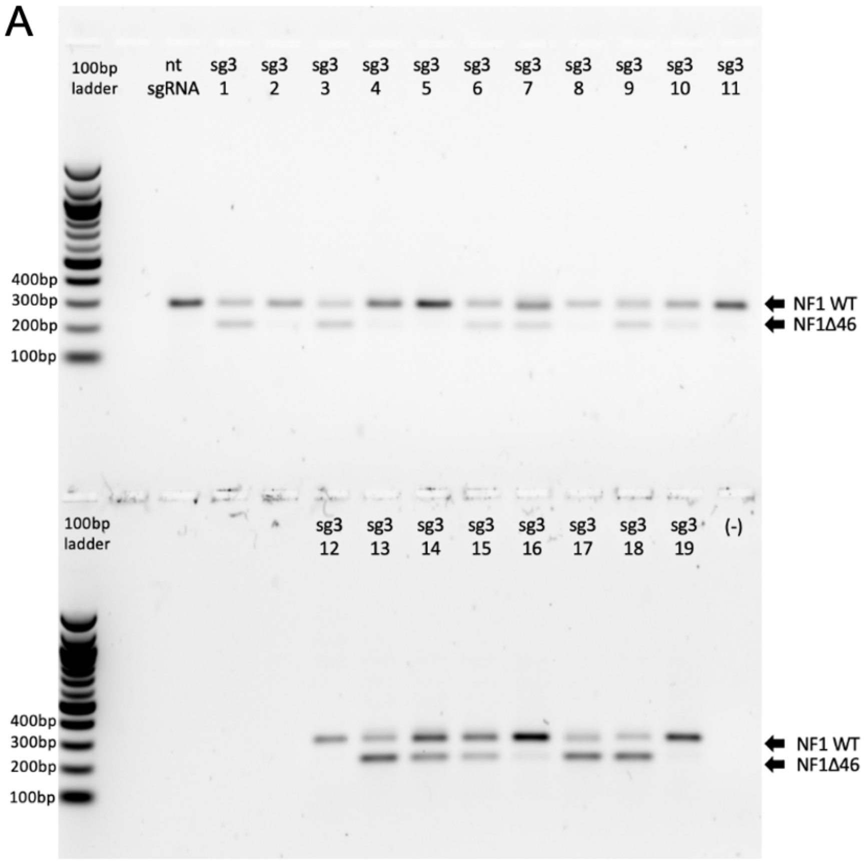


B

| | Percent exon 13 skipped |
|---|---|
| BRAF sg3 8 | 17% |
| BRAF sg3 6 | 19% |
| BRAF sg3 9 | 48% |
| BRAF sg3 4 | 55% |
| BRAF sg3 14 | 100% |

**Figure 2.S2: RNA validation of *BRAFΔ13* in AALE BRAF clones.**

A) RT-PCR and exon-spanning primers were used to validate and quantify the isoform ratios of BRAFΔ13 in the AALE BRAF clones. B) BRAFsg3 clones 8, 6, 9, 4 and 14 were selected for further analysis.



B

| | Percent exon 46 skipped |
|---|---|
| NF1 sg3 5 | 2% |
| NF1 sg3 7 | 22% |
| NF1 sg3 1 | 47% |
| NF1 sg3 13 | 71% |
| NF1 sg3 18 | 75% |

**Figure 2.S3: RNA validation of *NF1Δ46* in AALE NF1 clones.**

A) RT-PCR and exon-spanning primers were used to validate and quantify the isoform ratios of NF1Δ46 in the AALE NF1 clones. B) NF1sg3 clones 5, 7, 1, 13 and 18 were selected for further analysis.

## References

Berger, Alice H., Angela N. Brooks, Xiaoyun Wu, Yashaswi Shrestha, Candace Chouinard, Federica Piccioni, Mukta Bagul, et al. 2016. “High-Throughput Phenotyping of Lung Cancer Somatic Mutations.” Cancer Cell 30 (2): 214–28.

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Profiling of Lung Adenocarcinoma.” Nature 511 (7511): 543–50.

Concordet, Jean-Paul, and Maximilian Haeussler. 2018. “CRISPOR: Intuitive Guide Selection for CRISPR/Cas9 Genome Editing Experiments and Screens.” Nucleic Acids Research 46 (W1): W242–45.

Gonzàlez-Porta, Mar, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma. 2013. “Transcriptome Analysis of Human Tissues and Cell Lines Reveals One Dominant Transcript per Gene.” Genome Biology 14 (7): R70.

Haeussler, Maximilian, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, et al. 2016. “Evaluation of off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR.” Genome Biology 17 (1): 148.

Herbst, Roy S., Daniel Morgensztern, and Chris Boshoff. 2018. “The Biology and Management of Non-Small Cell Lung Cancer.” Nature 553 (7689): 446–54.

Imielinski, Marcin, Alice H. Berger, Peter S. Hammerman, Bryan Hernandez, Trevor J. Pugh, Eran Hodis, Jeonghee Cho, et al. 2012. “Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing.” Cell 150 (6): 1107–20.

Izar, Benjamin, and Asaf Rotem. 2016. “GILA, a Replacement for the Soft-Agar Assay That Permits High-Throughput Drug and Genetic Screens for Cellular Transformation.” Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel... [et Al.] 116 (October): 28.8.1–28.8.12.

Kim, Eddo, Alon Magen, and Gil Ast. 2007. “Different Levels of Alternative Splicing among Eukaryotes.” Nucleic Acids Research 35 (1): 125–31.

Kim, Pora, Mengyuan Yang, Ke Yiya, Weiling Zhao, and Xiaobo Zhou. 2020. “ExonSkipDB: Functional Annotation of Exon Skipping Event in Human.” Nucleic Acids Research 48 (D1): D896–907.

Kong-Beltran, Monica, Somasekar Seshagiri, Jiping Zha, Wenjing Zhu, Kaumudi Bhawe, Nerissa Mendoza, Thomas Holcomb, et al. 2006. “Somatic Mutations Lead to an Oncogenic Deletion of Met in Lung Cancer.” Cancer Research 66 (1): 283–89.

Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” Nature 499 (7457): 214–18.

Lu, Xinyuan, Nir Peled, John Greer, Wei Wu, Peter Choi, Alice H. Berger, Sergio Wong, et al. 2017. “MET Exon 14 Mutation Encodes an Actionable Therapeutic Target in Lung Adenocarcinoma.” Cancer Research 77 (16): 4498–4505.

Lundberg, Ante S., Scott H. Randell, Sheila A. Stewart, Brian Elenbaas, Kimberly A. Hartwell, Mary W. Brooks, Mark D. Fleming, et al. 2002. “Immortalization and Transformation of Primary Human Airway Epithelial Cells by Gene Transfer.” Oncogene 21 (29): 4577–86.

Mathieu, Luckson N., Erin Larkins, Oladimeji Akinboro, Pourab Roy, Anup K. Amatya, Mallorie H. Fiero, Pallavi S. Mishra-Kalyani, et al. 2022. “FDA Approval Summary: Capmatinib and Tepotinib for the Treatment of Metastatic NSCLC Harboring MET Exon 14 Skipping Mutations or Alterations.” Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 28 (2): 249–54.

Rotem, Asaf, Andreas Janzer, Benjamin Izar, Zhe Ji, John G. Doench, Levi A. Garraway, and Kevin Struhl. 2015. “Alternative to the Soft-Agar Assay That Permits High-Throughput Drug and Genetic Screens for Cellular Transformation.” Proceedings of the National Academy of Sciences of the United States of America 112 (18): 5708–13.

Sharifnia, Tanaz, Victor Rusu, Federica Piccioni, Mukta Bagul, Marcin Imielinski, Andrew D. Cherniack, Chandra Sekhar Pedamallu, et al. 2014. “Genetic Modifiers of EGFR Dependence in Non-Small Cell Lung Cancer.” Proceedings of the National Academy of Sciences 111 (52): 18661–66.

Smith, Harvey W., Lei Yang, Chen Ling, Arlan Walsh, Victor D. Martinez, Jonathan Boucher, Dongmei Zuo, et al. 2020. “An ErbB2 Splice Variant Lacking Exon 16 Drives Lung Carcinoma.” Proceedings of the National Academy of Sciences 117 (33): 20139–48.

Suzawa, Ken, Michael Offin, Adam J. Schoenfeld, Andrew J. Plodkowski, Igor Odintsov, Daniel Lu, William W. Lockwood, et al. 2019. “Acquired MET Exon 14 Alteration Drives Secondary Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor in EGFR-Mutated Lung Cancer.” JCO Precision Oncology 3 (May). https://doi.org/10.1200/PO.19.00011.

Thomas, James D., Jacob T. Polaski, Qing Feng, Emma J. De Neef, Emma R. Hoppe, Maria V. McSharry, Joseph Pangallo, et al. 2020. “RNA Isoform Screens Uncover the Essentiality and Tumor-Suppressor Activity of Ultraconserved Poison Exons.” Nature Genetics 52 (1): 84–94.

Togashi, Yosuke, Hiroshi Mizuuchi, Shuta Tomida, Masato Terashima, Hidetoshi Hayashi, Kazuto Nishio, and Tetsuya Mitsudomi. 2015. “MET Gene Exon 14 Deletion Created Using the CRISPR/Cas9 System Enhances Cellular Growth and Sensitivity to a MET Inhibitor.” Lung Cancer  90 (3): 590–97.

Tomasetti, Cristian, Luigi Marchionni, Martin A. Nowak, Giovanni Parmigiani, and Bert Vogelstein. 2015. “Only Three Driver Gene Mutations Are Required for the Development of Lung and Colorectal Cancers.” Proceedings of the National Academy of Sciences of the United States of America 112 (1): 118–23.

Tomasetti, Cristian, Bert Vogelstein, and Giovanni Parmigiani. 2013. “Half or More of the Somatic Mutations in Cancers of Self-Renewing Tissues Originate prior to Tumor Initiation.” Proceedings of the National Academy of Sciences of the United States of America 110 (6): 1999–2004.

Turpin, J., C. Ling, E. J. Crosby, Z. C. Hartman, A. M. Simond, L. A. Chodosh, J. P. Rennhack, et al. 2016. “The ErbB2ΔEx16 Splice Variant Is a Major Oncogenic Driver in Breast Cancer That Promotes a pro-Metastatic Tumor Microenvironment.” Oncogene 35 (47): 6053–64.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr, and Kenneth W. Kinzler. 2013. “Cancer Genome Landscapes.” Science 339 (6127): 1546–58.

Wang, Feng, Yang Liu, Wanglong Qiu, Elaine Shum, Monica Feng, Dejian Zhao, Deyou Zheng, Alain Borczuk, Haiying Cheng, and Balazs Halmos. 2022. “Functional Analysis of MET Exon 14 Skipping Alteration in Cancer Invasion and Metastatic Dissemination.” Cancer Research 82 (7): 1365–79.

# Chapter 3: Overexpression of *METΔ14* better models *METΔ14*-driven lung adenocarcinomas.

## Abstract:

The application of CRISPR/Cas9 to mammalian cells expanded the possibilities for genetic editing in human research. Commonly, CRISPR/Cas9 is used to create indels (point mutations, insertions or deletions) at specific regions in the genome. This tool was used to study *METΔ14,* an oncogene driving lung adenocarcinoma (LUAD) by creating indels at the splice sites of exon 14 and forcing its exclusion from the mRNA. This CRISPR-based expression of *METΔ14* has recently competed with overexpression experiments, which flood the cell with the mutant isoform. CRISPR-based expression is attractive to researchers with the idea that it could accurately reflect the tumor transcriptome through expression of *METΔ14* from its endogenous promoter. However, I found that *METΔ14* is specifically overexpressed in LUAD primary samples. Furthermore, I found that overexpression of *METΔ14* consistently outperformed CRISPR-based expression of *METΔ14* in a growth in low attachment (GILA) assay. This work suggests that overexpression of *METΔ14* is the most biologically relevant way to replicate the biology of *METΔ14* driven cancers, and that the expression of aberrantly spliced oncogenes should be considered before using CRISPR-based expression or overexpression models.

## Introduction

Oncogenic drivers are generated through a variety of different mechanisms, but a recently appreciated class of oncogenic drivers are created through aberrant mRNA splicing (Goldstein et al. 2016; Kong-Beltran et al. 2006; Kwong and Hung 1998; Siegel et al. 1999). Often mutations or deletions in the vicinity of splice sites in the DNA lead to exon-skipping events where canonical exons are excluded from the mature mRNA (Kong-Beltran et al. 2006; Lu et al. 2017; Smith et al. 2020). This change can have huge implications on protein function, such as removing critical regulatory regions (Kong-Beltran et al. 2006), or stabilizing dimerization to promote proliferative signaling (Siegel et al. 1999). As aberrantly spliced oncogenes attract the interest of researchers, different ways of studying these splicing events have arisen. Commonly, overexpression is used to overwhelm the model system with the aberrant isoform (Kwong and Hung 1998; Siegel et al. 1999; Smith et al. 2020). More recently, CRISPR/Cas9 has been used to target to splice sites to force the production of aberrant exon skipping events (Lu et al. 2017; Togashi et al. 2015; Wang et al. 2022). This method allows expression of the aberrantly spliced oncogene from the endogenous promoter. While the goal of both methods is to study these oncogenic splicing events in a laboratory setting, the question remains which method is a more accurate model to study cancer.

*METΔ14* is a well-characterized aberrantly spliced oncogene that drives the initiation and maintenance of LUAD (Kong-Beltran et al. 2006). *MET* is a receptor tyrosine kinase in the Ras-MAPK pathway whose unchecked signaling leads to cancer development. With *METΔ14*, skipped exon 14 encodes a key regulatory region that targets MET for lysosomal degradation (Kong-Beltran et al. 2006). Without this regulatory region, the half-life of *METΔ14* is prolonged in the plasma membrane (Lu et al. 2017). *METΔ14* has been studied using both overexpression vectors (Suzawa et al. 2019) and CRISPR-based methods (Lu et al. 2017; Togashi et al. 2015; Wang et al. 2022). However, these studies did not consider the level of *METΔ14* expression in LUAD samples. Because of the well-characterized nature of this aberrant splicing event, it is an ideal paradigm system to use for the comparison of aberrantly spliced variants generated by overexpression versus CRISPR-based expression.

In this chapter I show that LUAD primary samples expressing *METΔ14* exhibit an allele-specific overexpression of the mutant allele. The phenomenon of allele-biased, splice variant overexpression may be more widespread within tumor samples. Indeed, our finding corroborates previous studies in primary samples of LUAD revealing the predominant expression of *METΔ14* despite heterozygous backgrounds for exon 14 deletions (Kong-Beltran et al. 2006; Onozato et al. 2009). As *METΔ14* tends to be overexpressed in primary samples, applying this to functional assays I show that overexpressing *METΔ14* leads to more consistent results and limits false negatives in a growth in low attachment (GILA) assay (Rotem et al. 2015) as opposed to CRISPR-based expression. Furthermore, overexpression of *METΔ14* activates signaling independent of ligand (Suzawa et al. 2019), facilitating research focused on aberrantly spliced receptors. Our results suggest that the expression of aberrantly spliced oncogenes within their cancer type should also be considered when choosing an overexpression or CRISPR-based expression model.

## Results

### *METΔ14 is overexpressed in an allele specific manner in LUAD primary samples*

Two separate studies investigating *METΔ14* revealed that although both primary samples and cell lines were heterozygous for exon 14 deletions, predominantly, the *METΔ14* isoform was produced (Kong-Beltran et al. 2006; Onozato et al. 2009). Furthermore, Lu *et al.* identified somatic mutations at exon 14 of *MET* in a larger cohort of LUAD samples from The Cancer Genome Atlas (TCGA) and found that the majority of these samples overexpress *MET* and predominantly express the mutant isoform (Lu et al. 2017). These three studies suggest an allele-specific expression bias of *METΔ14*. Allele-specific expression occurs when one allele exhibits a higher level of expression compared to the other and is implicated in cancer (PCAWG Transcriptome Core Group et al....; Mayba et al. 2014; Ongen et al. 2014; Castel et al. 2018; Bielski et al. 2018). Notably, this bias in allele expression can directly influence the expression of cancer driver genes through specific expression of the oncogenic allele (Mayba et al. 2014; Bielski et al. 2018). I hypothesized that *METΔ14* undergoes allele-specific expression which may be critical to its transformative abilities. The *Lu et al.* study did not further investigate the zygosity of the LUAD samples nor any associated copy number changes, which may contribute to allelic imbalance. Additionally, the matched DNA and RNA sequencing data associated with these samples permit tracking of allele usage. Therefore, I further investigated the underlying mechanism driving the predominant expression of *METΔ14* in these LUAD samples.

While typically allele-expression bias is due to copy number alteration (PCAWG Transcriptome Core Group et al....), I found no sample with a *METΔ14* mutation and an associated copy number amplification of the *MET* gene (Figure 3.1). This is consistent with previous findings revealing that *METΔ14* and *MET* copy number alterations are mutually exclusive (Onozato et al. 2009; Baldacci et al. 2020; Guo et al. 2019). However, the majority of *METΔ14* samples also have co-occurring *MET* mRNA overexpression (Figure 3.1). This indicates these *METΔ14* samples employ a different mechanism to drive *MET* overexpression independent of MET copy number alterations.

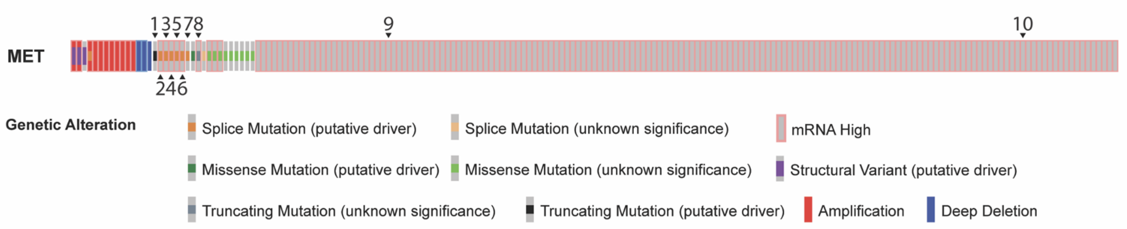


**Figure 3.1: Primary Samples with *METΔ14* alterations co-occur with *MET* overexpression.**

Oncoprint adapted from cBioPortal reveals genetic mutations in *MET* per LUAD TCGA primary sample. Arrows correspond to TCGA IDs in C from *Lu et al 2017* - described in Table 3.1. Note not all splice mutations in *MET* are captured using the analysis from cBioPortal.

To determine if *METΔ14* mutation zygosity could explain *METΔ14* allele-specific expression, I compared whole exome and matched RNA-seq data from these TCGA samples (Figure 3.2 and Table 3.1). Additionally, I confirmed the strong bias of *METΔ14* mutant allele expression from the quantification of the percentage of *METΔ14* isoform usage in the LUAD tumor samples

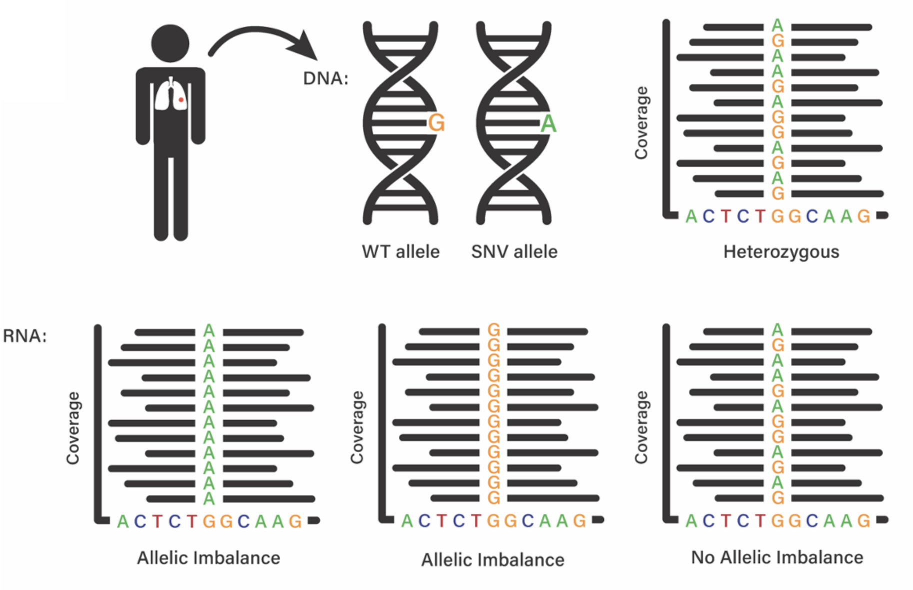
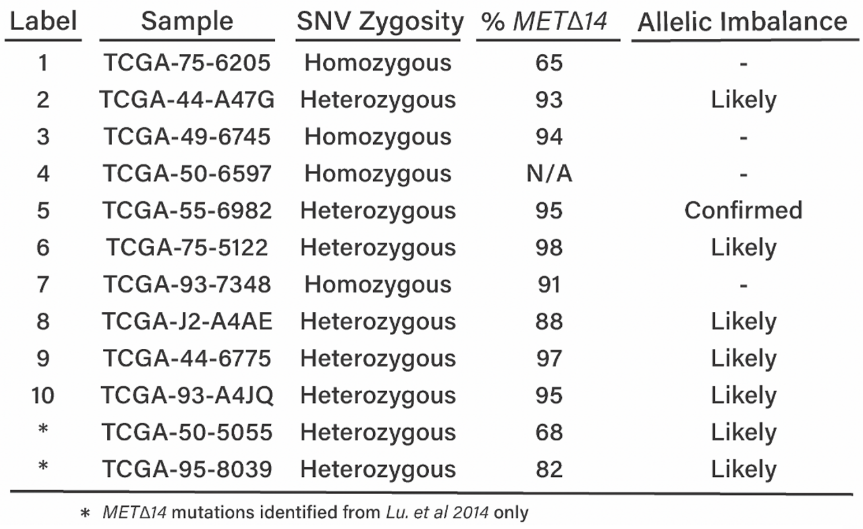


**Figure 3.2: Allelic imbalance identified using germline SNV ratios.**

Diagram describing how allelic imbalance was identified in TCGA DNA and matched RNA sequencing data. At the DNA level, lung adenocarcinoma tumors heterozygous for SNVs within the *MET* gene were identified and chosen for further analysis. These SNV locations were identified in matched RNA-seq data, and ratios of SNV to WT nucleotide will determine if there is an allele-based expression of *METΔ14*.

**Table 3.1**: TCGA sample, SNV zygosity from DNA-seq data, allelic imbalance from RNA-seq data, and percent *METΔ14* expression calculated with JuncBase results from *Soulette* et al. 2023



| Label | Sample | SNV Zygosity | % *METΔ14* | Allelic Imbalance |
|---|---|---|---|---|
| 1 | TCGA-75-6205 | Homozygous | 65 | - |
| 2 | TCGA-44-A47G | Heterozygous | 93 | Likely |
| 3 | TCGA-49-6745 | Homozygous | 94 | - |
| 4 | TCGA-50-6597 | Homozygous | N/A | - |
| 5 | TCGA-55-6982 | Heterozygous | 95 | Confirmed |
| 6 | TCGA-75-5122 | Heterozygous | 98 | Likely |
| 7 | TCGA-93-7348 | Homozygous | 91 | - |
| 8 | TCGA-J2-A4AE | Heterozygous | 88 | Likely |
| 9 | TCGA-44-6775 | Heterozygous | 97 | Likely |
| 10 | TCGA-93-A4JQ | Heterozygous | 95 | Likely |
| * | TCGA-50-5055 | Heterozygous | 68 | Likely |
| * | TCGA-95-8039 | Heterozygous | 82 | Likely |

* *METΔ14* mutations identified from *Lu. et al 2014* only

(Soulette et al. 2023) (Figure 1C). I used germline single nucleotide variants (SNVs) to determine both zygosity and to track which allele was used to express *MET.* I manually scanned the entire length of the *MET* gene for evidence of SNVs. For heterozygous samples, I identified the ratio of RNA expression at the SNVs to determine the allelic imbalance in these samples. (Figure 3.2). To confirm this allele-specific expression of *METΔ14* is unique to cancer, it is necessary to analyze both the tumor and matched normal samples; however, I identified only one *METΔ14* sample that was heterozygous and had a matched normal sample (Figure 3.3A). As predicted, while the matched normal sample equally expressed both the wild type and non-reference SNVs, there is a clear allele-specific expression in the tumor. While the remaining heterozygous samples lacked matched normal RNA-seq data, all samples exhibited an allelic imbalance in the tumor (Figure 3.3B-H). This suggests the allele-specific expression of *METΔ14* is a widespread phenomenon among these LUAD tumors. Furthermore, given that 8/12 of *METΔ14* samples also overexpress total *MET* compared to normal samples (Figure 3.1), this suggests that *METΔ14* is overexpressed in an allele-specific manner in these samples.

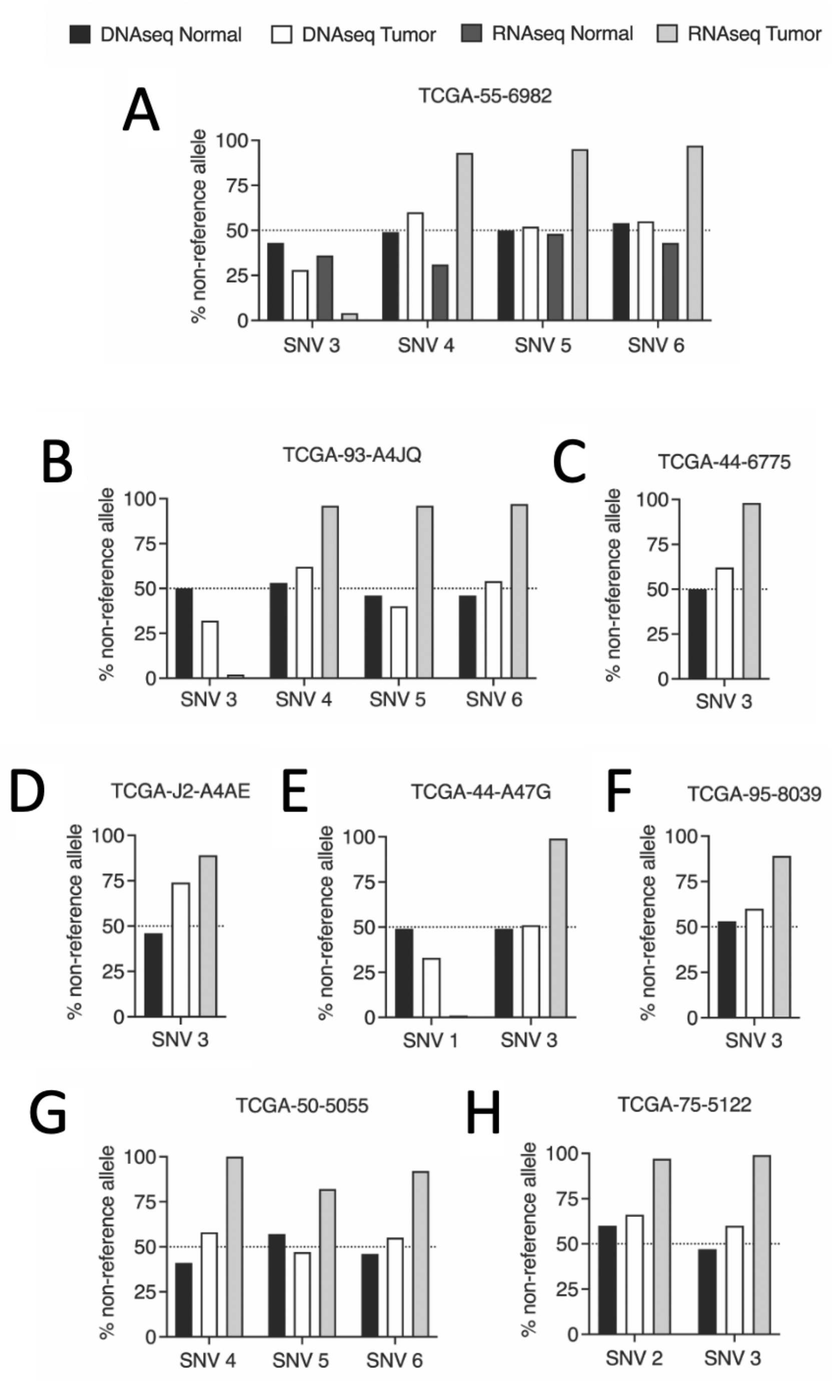


**Figure 3.3: Allelic imbalance in TCGA LUAD tumors.**

Graphs representing the percentage of non-reference allele for germline SNVs in the heterozygous TCGA primary samples from Table 3.1.

*AALE METΔ14 Overexpression leads to consistent survival in GILA assay*

This allele-specific overexpression of *METΔ14* in LUAD primary samples suggests that overexpression of *METΔ14* is required for transformation. I hypothesize introducing the splice site mutation alone by CRISPR is not sufficient to drive oncogenesis because of likely low levels of *METΔ14* expression. To directly compare these two methods, I lentivirally introduced *METΔ14* via an overexpression plasmid into immortalized tracheobronchial epithelial cells (AALE) (Lundberg et al. 2002). I also generated an AALE line with 97% CRISPR-expressed *METΔ14* as described in chapter 2. Briefly, I lentivirally introduced an sgRNA targeting the splice sites of exon 14 into AALE cells constitutively expressing Cas9. Then, I clonally isolated AALE cells expressing 97% *METΔ14.* Using RT-PCR and primers spanning exon 14, I confirmed *METΔ14* in both the overexpressed line (over. *METΔ14*) and CRISPR-expressed line (METsg8 5). (Figure 3.4A). Because ssCRISPR edits the genome, very little wild type *MET* is produced in METsg8 5, insinuating CRISPR-editing occurred on both alleles.

I used a western blot to quantify MET protein in these AALE cells (Figure 3.4B). These findings demonstrated that only overexpression of *METΔ14* led to activated MET, despite both overexpressed *METΔ14* and METsg8 5 producing *METΔ14* mRNA. Additionally, the observed down-shift in molecular weight with total MET in the overexpressed *METΔ14* and METsg8 5 lines provided strong evidence for the presence of the METΔ14 protein, confirming successful translation of this truncated oncogenic protein.

I used a GILA cell viability assay to determine the dose of *METΔ14* required to support the growth of AALE cells in low attachment (Figure 3.4C). While overexpression of *METΔ14* did not result in cell viability equivalent to KRAS G12V overexpression, it did enable AALE cells to consistently survive in low attachment. In contrast, CRISPR-based expression of *METΔ14* resulted in cell viability measurements similar to those of the parental negative control. These findings demonstrate the pivotal role of *METΔ14* overexpression in transformation, and demonstrate that CRISPR-based expression of *METΔ14* alone does not give cells a proliferative advantage in low attachment.

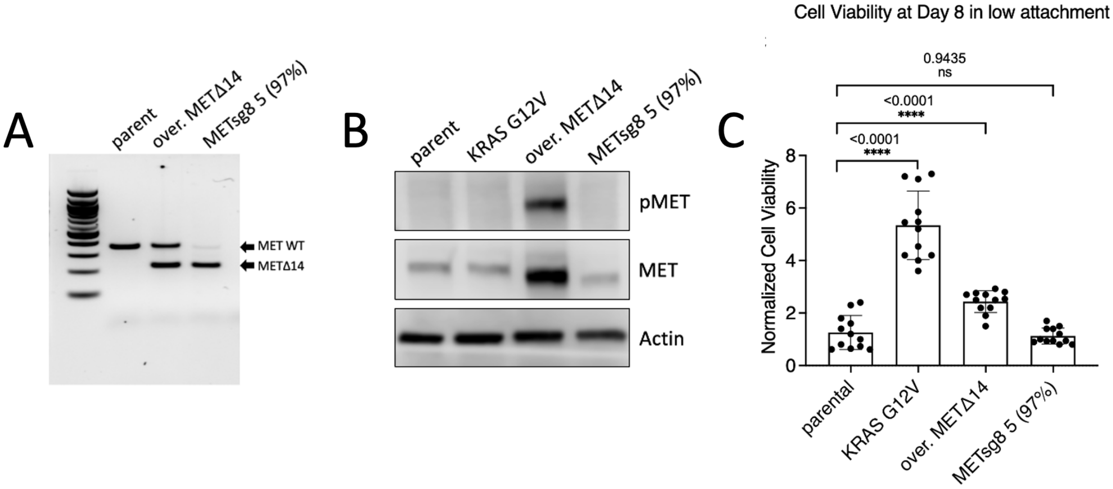


**Figure 3.4: Overexpression of *METΔ14* permits AALE cells to survive in low attachment.**

A) RT-PCR of parental AALE cell, overexpressed METΔ14 (over. METΔ14) and ssCRISPR-expressed METΔ14 (MET sg8 5 clone with 97% skipping). B) Western blot of over. METΔ14, MET sg8 5, and an AALE line with KRAS G12V overexpression. C) Low attachment growth assay of 3 biological replicates of over. METΔ14, MET sg8 5 and KRAS G12V compared to the negative control in three experiments. Comparisons performed with a Mann Whitney test. Error bars represent standard deviation of the mean.

*Low PC9 METΔ14 Overexpression does not rescue cells.*

Next, I used the PC9 osimertinib-rescue assay to determine the amount of *METΔ14* required to activate the Ras-MAPK pathway in lieu of EGFR. This lung adenocarcinoma line is dependent on an activating mutation in EGFR, however expression of strong oncogenes can "rescue" PC9 cells despite EGFR inhibition with inhibitors like osimertinib (Sharifnia et al. 2014; Berger et al. 2016). In chapter 2 of this dissertation I confirmed that CRISPR-based expression of near complete *METΔ14* isoform did not rescue PC9 cells after EGFR inhibition (Figure 2.13C). Therefore, I hypothesized that overexpression of *METΔ14* is required for the PC9 osimertinib-rescue assay.

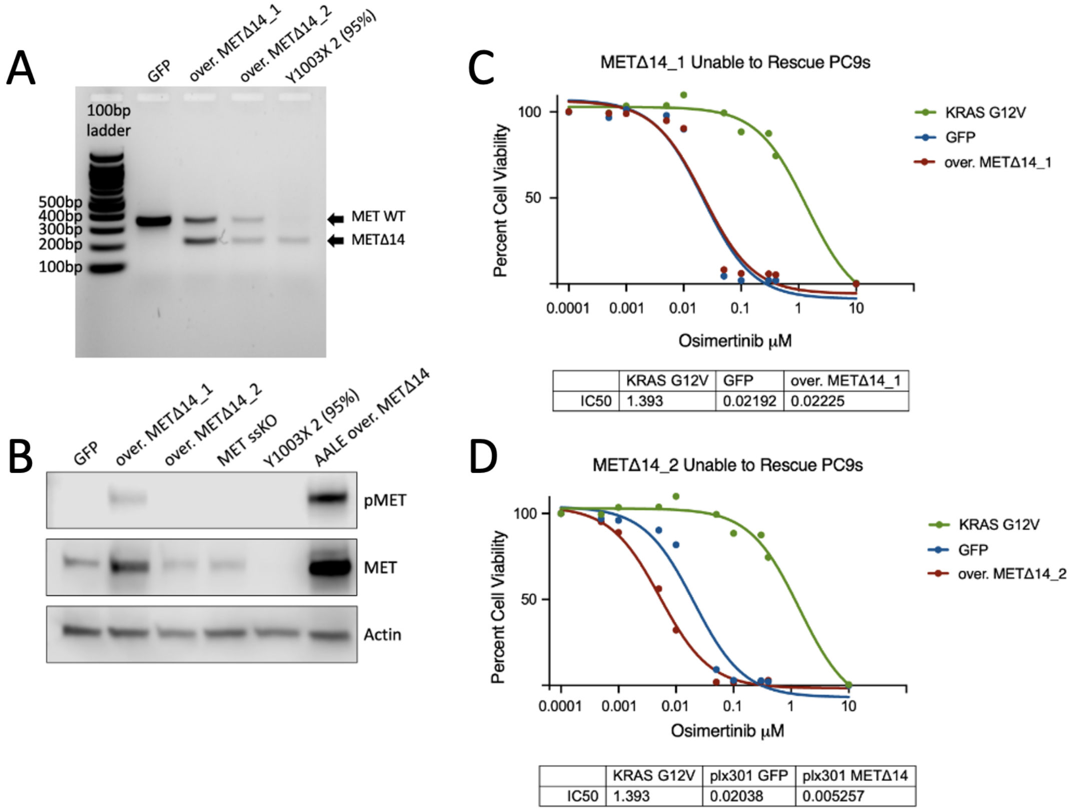
I created two PC9 cell lines containing lentivirally-integrated overexpression plasmids for *METΔ14* (over. METΔ14_1 and over. METΔ14_2). I validated the presence of METΔ14 at the RNA level in these cell lines (Figure 3.5A), and quantified MET protein using a western blot (Figure 3.5B). Additionally, I compared this RNA and protein expression with CRISPR-expressed *METΔ14* PC9 cells generated as described in Chapter 2. Briefly, both lines were generated by Synthego using CRISPR targeting splice regions. MET Y1003X 2 is a clonal cell line with 95% *METΔ14* expression, and MET ssKO is a heterogeneous cell line producing 42% *METΔ14*. While I identified *METΔ14* mRNA in both overexpression lines, METΔ14_2 produced similar levels of *METΔ14* as MET Y1003X 2. This suggests that the METΔ14_2 overexpression cell line expresses low levels of *METΔ14*. This low expression in METΔ14_2 is corroborated by the western data, revealing that METΔ14_2 produced very little total MET protein, similar to the GFP negative control.

Additionally, there is no activated MET (pMET) in METΔ14_2, confirming this overexpression line failed to overexpress *METΔ14*. Interestingly, there was no MET protein produced in Y1003X 2, which is likely due to degradation. However, the absence of pMET in MET ssKO demonstrates that CRISPR-based expression of *METΔ14* is not sufficient for receptor activation. While METΔ14_1 produced pMET, this level was not comparable to pMET produced from AALE over. METΔ14*,* which provided AALE cells with a proliferative advantage in the GILA assay (Figure 3.4C). This indicates that overexpression of *METΔ14* in PC9s was not as successful as AALEs.

Using these cells in an osimertinib-rescue assay, I found that *METΔ14* expression from neither METΔ14_1 or METΔ14_2 rescued PC9s after EGFR inhibition (Figure 3.5C-D). Given the low pMET produced by these cells, these results are not surprising. Because of the low amount of pMET quantified by the western, I concluded that *METΔ14* was never overexpressed in these cell lines. However, similar to the CRISPR-expressed cell lines from chapter 2, a higher threshold of *METΔ14* expression is required to rescue PC9 cells than produced by these cell lines.

### *Discussion*

Overall, this work revealed that *METΔ14* is overexpressed in an allele-specific manner in LUAD primary samples, and this overexpression contributes to this oncogene's transformative capabilities. In a study of cancer genomes examining the positive selection of oncogenic driver mutant alleles through ploidy changes, alleles expressing activating mutations in *MET* were found to experience strong positive selection (Bielski et al. 2018). This study



| | KRAS G12V | GFP | over. METΔ14_1 |
|---|---|---|---|
| IC50 | 1.393 | 0.02192 | 0.02225 |

| | KRAS G12V | plx301 GFP | plx301 METΔ14 |
|---|---|---|---|
| IC50 | 1.393 | 0.02038 | 0.005257 |

**Figure 3.5: Overexpression of *METΔ14* in PC9 cells not successful.**

A) RT-PCR using primers spanning *MET* exon 14 of two cell lines of overexpressed *METΔ14* (over. METΔ14_1 and over. METΔ14_2) and MET Y1003X (95% CRISPR-expressed *METΔ14*). B) Western blot comparing activated MET (pMET) in overexpressed *METΔ14* (METΔ14_1 and METΔ14_2) and CRISPR-expressed *METΔ14* PC9 cells (MET Y1003X and MET ssKO). Osimertinib titration using overexpressed (over.) C) METΔ14_1 (C) and D) METΔ14_2.

examined all *MET* oncogenic driver mutations at DNA-level copy number selection, which does not capture RNA-level changes in expression. However, the work presented in this chapter uses both DNA and RNA sequencing data to show that ploidy changes are not required for allele-specific expression of *METΔ14* (Figure 3.1). Given the consistency of *METΔ14* overexpression in the majority of samples examined, our data suggest the overexpression of the oncogenic allele is required for *METΔ14*-driven cancer progression. Further studies are necessary to understand the molecular mechanism of the allele-specific overexpression which could involve *cis*-acting genetic or epigenetic factors, or allele-specific transcript stability.

This allele-based overexpression of *METΔ14* in LUAD primary samples suggests that overexpressing *METΔ14* in functional assays is the most accurate method to reflect its expression in a tumor cell. Building upon previous work in this dissertation revealing that CRISPR-based expression of *METΔ14* is not sufficient for proliferation in the AALE GILA assay, I found that overexpression of *METΔ14* is critical to provide AALE cells with a proliferative advantage in low attachment (Figure 3.4C). This activation occurs independent of HGF, the ligand activating MET, revealing the oncogenic potential of *METΔ14* overexpression alone. Furthermore, because LUAD tumors are associated with high HGF expression (Lu et al. 2017), it's possible that both *METΔ14* overexpression and high HGF expression may be required for oncogenic progression in *METΔ14*-driven cancers. This is in direct comparison to CRISPR-based expression of *METΔ14*, which resulted in cell viability measurements similar to the negative control. These experiments provide further evidence that *MET* exon 14 skipping, alone, is not sufficient for oncogene activation. They also support our model whereby additional increased dosage of the *METΔ14* allele is necessary for oncogene activation; consistent with the characterization in primary samples.

While the AALE experiments demonstrated the dosage of *METΔ14* contributes to the transformative ability of this oncogene, I was unable to determine if *METΔ14* overexpression in PC9s can rescue cells after EGFR inhibition. Unlike the AALE GILA assay which can detect subtle differences in proliferation in low attachment, the PC9 osimertinib-rescue assay requires a higher threshold of oncogene activation to ultimately rescue cells in the presence of the inhibitor. As opposed to the strong pMET activation in the AALE *METΔ14* overexpression cell line, the western blot revealed both PC9 *METΔ14* overexpression cell lines resulted in no activation or low activation of pMET (Figure 3.5B, METΔ14_2 and METΔ14_1 respectively). Because the PC9 osimertinib-rescue assay requires a higher threshold of oncogene activation, the resulting low amount of pMET activation from the PC9 METΔ14_1 cell line was not sufficient to rescue PC9s after osimertinib treatment (Figure 3.5C). Because of the low pMET levels in these cell lines, I conclude that the PC9 results from this chapter were due to unsuccessful integration of *METΔ14* plasmids. However, this further supports our hypothesis that high levels of *METΔ14* are required to observe an oncogenic phenotype. This is further supported by a study which found that *METΔ14* overexpression can rescue PC9 cells after osimertinib inhibition, supporting a role of *METΔ14* arising from a secondary mutation in a LUAD case initially driven by mutant EGFR (Suzawa et al. 2019). This study insinuates that *METΔ14* overexpression, when introduced successfully, can rescue PC9s after EGFR inhibition and therefore can be used in these assays.

I believe that two factors drive *METΔ14* as an oncogene in LUAD: 1) *METΔ14* allele-specific overexpression, and 2) the abundance of MET ligand, HGF (Lu et al. 2017). These factors have implications for how *METΔ14* is studied. A body of work aims to decouple the *METΔ14* mutation from overexpression due to *MET* amplification by using CRISPR-based methods (Lu et al. 2017; Togashi et al. 2015; Wang et al. 2022). However this CRISPR-based expression does not lead to overexpression of the *METΔ14* mRNA. Thus, I suggest future studies on *METΔ14* will more accurately recapitulate tumor cells if *METΔ14* is overexpressed and assays performed with and without the presence of ligand. Expanding this work beyond *METΔ14*, before choosing CRISPR-based expression or overexpression as a model system I suggest examining how these splice variants are expressed in primary samples. Furthermore, when using CRISPR-based expression I suggest coupling this method with subsequent functional assays which require less oncogenic activation to limit potential false negative results.

## Chapter 3 Material and Methods

### *Data Acquisition*

For each TCGA sample from the Lung Adenocarcinoma TCGA PanCancer Atlas cohort labeled in Figure 1C, previously-aligned whole exome (DNA-seq) and RNA-seq samples were securely downloaded from the NCI Genomic Data Commons. The OncoPrint was generated in cBioPortal (Cerami et al. 2012; Gao et al. 2013)

### *SNV Characterization*

All TCGA samples were associated with files: tumor and matched normal DNA-seq data, as well as tumor RNA-seq data. Two files had matched normal RNA-seq data, which could be used to confirm cancer-specific expression. To visualize allele-specific expression within this data, all files per sample were imported into Interactive Genomics Viewer (IGV) (Robinson et al. 2011). As SNVs provide a track record of allele abundance, we scanned the entire MET gene and determined relative percentages of SNVs at those loci between the DNA and RNA-seq data, which can be calculated in IGV. For the DNA-seq, a near 1:1 ratio of SNV to wild type allele indicates heterozygosity. For the matched RNA-seq data, a proportion of SNV to wild type allele close to 0% or 100% of heterozygous SNV loci indicates allelic imbalance.

### *Cell lines*

AALE cells and PC9 cells were obtained and cultured as described in Chapter 2. Clonal METsg8 5 was generated as described in Chapter 2. The MET Y1003X clone and MET ssKO cell line were generated by Synthego as described in Chapter 2.

*Lentiviral Plasmids*

The overexpression METΔ14 virus (plx317 METΔ14) and overexpression KRAS G12V virus (plx301 KRAS G12V) was a gift from Alice Berger.

*Lentiviral Generation*

The Lentivirus was created using the UC San Francisco Viracore facility.

*AALE and PC9 overexpression lines*

Cells were grown to 70% confluency, trypsinized, then plated into a 24 well plate so cells were ~50% confluent the following day. 200μl of lentivirus was mixed with 200μl cell media + 8μg/mL polybrene (Millipore Sigma, TR1003). Cells were incubated 24 hours in the lentivirus/polybrene media, then cells were left to recover in normal media for 24 hours. Cells were selected with 1μg/ml of puromycin (Millipore Sigma, P8833) for 4-7 days. Successful lentiviral integration was measured at the RNA level using RT-PCR.

*RT-PCR METΔ14 mRNA Validation*

RNA was isolated via TriReagent (Sigma-Aldrich, T9424) and the Zymo Direct Zol RNA Miniprep kit (Zymo Research Corporation, R2050). Subsequent cDNA prep was performed using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, 4368814), and resulting cDNA was used as a PCR template using the HotStart ReadyMix PCR Kit (Kapa Biosystems, KK6202). Exon spanning primers were ordered from IDT:

MET_splice_F: 5'- TGGGTTTTTCCTGTGGCTGA -3'

MET_splice_R: 5'- GGGCCCAATCACTACATGCT -3'

and run using these cycling conditions: initial denaturation of 95°C 3min, 30 cycles of 98°C 20sec, 61°C 15sec and 72°C 30sec, a final extension of 72°C

2min. Resulting amplicons were run on a 1.2-2% agarose gel until wild type and exon-skipped isoforms were resolved. Relative isoform abundance ratios were quantified using Image Studio Lite.

### *Western Blot*

One tablet of Protease Inhibitor cocktail (Roche, 04693124001) was added to a 10mL aliquot of RIPA lysis buffer (Thermo Fisher Scientific, 89900) and dissolved completely. Cells were grown to 80% confluency in 10cm tissue treated dishes (Santa Cruz Biotechnology, Inc., sc-200286). Plates washed 2x with ice cold DPBS (Life Technologies Corp., 14190144). 1mL ice cold RIPA with Protease Inhibitor was added to the cells and scraped into 2mL tube. Tubes were incubated on ice and vortexed periodically. Lysed cells were pelleted at max speed in a chilled centrifuge for 10min, then supernatant aliquoted in volumes of 200µl. Lysates were stored at -80°C.

Lysate was sonicated at the maximum setting for increments of 30 seconds 2x. Lysate was left to recover on ice for 1 minute between sonications. The protein content in the sonicated samples was quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, 23225).

30µg protein was added to 11.7µl MLB (1:10 dilution of 2-mercaptoethanol (Bio-Rad, 1610710) and 4x Laemmli sample buffer (Bio-Rad, 1610747)), and volume was brought to 46.7ul total using RIPA buffer in a 2mL tube. Samples were denatured at 95°C for 5 minutes. Samples and Precision Plus protein standards (Bio-Rad, 1610374) were loaded into a 4-15% Mini-PROTEAN TGX Precast Gel (Bio-Rad, 4561083EDU). Gel was run at 70V using 1x TGX Buffer (Bio-Rad, 1610772) until adequate separation of ladder was obtained.

The Trans-Blot Turbo RTA Transfer PVDF kit (Bio-Rad, 1704272) protocol with 1x Bio-Rad Transfer Buffer (Bio-Rad, 10026938) was used with the High Molecular Weight setting. After the transfer the membrane was blocked with 4% BSA (Millipore Sigma, A3059) for 1 hour. Then 1:2000 primary antibody (pMET (Cell Signaling Technology, 3077), total MET (Cell Signaling Technology, 8198) was incubated at 4°C overnight.

The next day, the membrane was washed 3x 10min with 1x TBST (1x TBS with 1mL Tween-20 (Fisher Scientific, BP337)). Blots were incubated with 1:1000 HRP-conjugated secondary antibody (Li-Cor, 92601000) in 4% BSA for 1 hour. Blots were washed 3x 5min in 1x TBST. Luminol pen (Li-Cor, 926-91000) was used to mark ladder, then blots incubated with ECL (WesternSure PREMIUM Chemiluminescent Substrate (Li-Cor, 926-95000) for 5 minutes. Blots were visualized with C-Digit Imager (Li-Cor) according to equipment instructions.

To blot for actin, blots were washed with 1x TBST 3x 10min, then incubated in 1:1000 actin-HRP antibody (Cell Signaling Technology, 7074S) in 4% BSA. Blots were washed 3x 5min in 1x TBST, ladder marked with Luminol pen, incubated with ECL for 5 minutes, and imaged as above.

*AALE - GILA Cell TIter Glo assay*

As performed in Chapter 2.

*PC9 - EGFR inhibitor titration assay*

As performed in Chapter 2, but with osimertinib (Selleck Chemicals, S7297-5MG) instead of erlotinib.

## References

Baldacci, Simon, Martin Figeac, Martine Antoine, Clotilde Descarpentries, Zoulika Kherrouche, Philippe Jamme, Marie-Christine Copin, et al. 2020. “High MET Overexpression Does Not Predict the Presence of MET Exon 14 Splice Mutations in NSCLC: Results From the IFCT PREDICT.amm Study.” Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 15 (1): 120–24.

Berger, Alice H., Angela N. Brooks, Xiaoyun Wu, Yashaswi Shrestha, Candace Chouinard, Federica Piccioni, Mukta Bagul, et al. 2016. “High-Throughput Phenotyping of Lung Cancer Somatic Mutations.” Cancer Cell 30 (2): 214–28.

Bielski, Craig M., Mark T. A. Donoghue, Mayur Gadiya, Aphrothiti J. Hanrahan, Helen H. Won, Matthew T. Chang, Philip Jonsson, et al. 2018. “Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer.” Cancer Cell 34 (5): 852–62.e4.

Castel, Stephane E., Alejandra Cervera, Pejman Mohammadi, François Aguet, Ferran Reverter, Aaron Wolman, Roderic Guigo, Ivan Iossifov, Ana Vasileva, and Tuuli Lappalainen. 2018. “Modified Penetrance of Coding Variants by Cis-Regulatory Variation Contributes to Disease Risk.” Nature Genetics 50 (9): 1327–34.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” Cancer Discovery 2 (5): 401–4.

Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, et al. 2013. “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal.” Science Signaling 6 (269): l1.

Goldstein, Leonard D., James Lee, Florian Gnad, Christiaan Klijn, Annalisa Schaub, Jens Reeder, Anneleen Daemen, et al. 2016. “Recurrent Loss of NFE2L2 Exon 2 Is a Mechanism for Nrf2 Pathway Activation in Human Cancers.” Cell Reports 16 (10): 2605–17.

Guo, Robin, Lynne D. Berry, Dara L. Aisner, Jamie Sheren, Theresa Boyle, Paul A. Bunn, Bruce E. Johnson, et al. 2019. “MET IHC Is a Poor Screen for MET Amplification or MET Exon 14 Mutations in Lung Adenocarcinomas: Data from a Tri-Institutional Cohort of the Lung Cancer Mutation Consortium.” Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 14 (9): 1666–71.

Kong-Beltran, Monica, Somasekar Seshagiri, Jiping Zha, Wenjing Zhu, Kaumudi Bhawe, Nerissa Mendoza, Thomas Holcomb, et al. 2006. “Somatic Mutations

Lead to an Oncogenic Deletion of Met in Lung Cancer.” Cancer Research 66 (1): 283–89.

Kwong, K. Y., and M. C. Hung. 1998. “A Novel Splice Variant of HER2 with Increased Transformation Activity.” Molecular Carcinogenesis 23 (2): 62–68.

Lu, Xinyuan, Nir Peled, John Greer, Wei Wu, Peter Choi, Alice H. Berger, Sergio Wong, et al. 2017. “MET Exon 14 Mutation Encodes an Actionable Therapeutic Target in Lung Adenocarcinoma.” Cancer Research 77 (16): 4498–4505.

Lundberg, Ante S., Scott H. Randell, Sheila A. Stewart, Brian Elenbaas, Kimberly A. Hartwell, Mary W. Brooks, Mark D. Fleming, et al. 2002. “Immortalization and Transformation of Primary Human Airway Epithelial Cells by Gene Transfer.” Oncogene 21 (29): 4577–86.

Mayba, Oleg, Houston N. Gilbert, Jinfeng Liu, Peter M. Haverty, Suchit Jhunjhunwala, Zhaoshi Jiang, Colin Watanabe, and Zemin Zhang. 2014. “MBASED: Allele-Specific Expression Detection in Cancer Tissues and Cell Lines.” Genome Biology 15 (8): 405.

Ongen, Halit, Claus L. Andersen, Jesper B. Bramsen, Bodil Oster, Mads H. Rasmussen, Pedro G. Ferreira, Juan Sandoval, et al. 2014. “Putative Cis-Regulatory Drivers in Colorectal Cancer.” Nature 512 (7512): 87–90.

Onozato, Ryoichi, Takayuki Kosaka, Hiroyuki Kuwano, Yoshitaka Sekido, Yasushi Yatabe, and Tetsuya Mitsudomi. 2009. “Activation of MET by Gene Amplification or by Splice Mutations Deleting the Juxtamembrane Domain in Primary Resected Lung Cancers.” Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 4 (1): 5–11.

PCAWG Transcriptome Core Group, Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğlu, Nuno A. Fonseca, Yao He, André Kahles, et al. 2020. “Genomic Basis for RNA Alterations in Cancer.” Nature 578 (7793): 129–36.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” Nature Biotechnology 29 (1): 24–26.

Rotem, Asaf, Andreas Janzer, Benjamin Izar, Zhe Ji, John G. Doench, Levi A. Garraway, and Kevin Struhl. 2015. “Alternative to the Soft-Agar Assay That Permits High-Throughput Drug and Genetic Screens for Cellular Transformation.” Proceedings of the National Academy of Sciences of the United States of America 112 (18): 5708–13.

Sharifnia, Tanaz, Victor Rusu, Federica Piccioni, Mukta Bagul, Marcin Imielinski, Andrew D. Cherniack, Chandra Sekhar Pedamallu, et al. 2014. “Genetic Modifiers of EGFR Dependence in Non-Small Cell Lung Cancer.” Proceedings of the National Academy of Sciences 111 (52): 18661–66.

Siegel, P. M., E. D. Ryan, R. D. Cardiff, and W. J. Muller. 1999. “Elevated Expression of Activated Forms of Neu/ErbB-2 and ErbB-3 Are Involved in the Induction of Mammary Tumors in Transgenic Mice: Implications for Human Breast Cancer.” The EMBO Journal 18 (8): 2149–64.

Smith, Harvey W., Lei Yang, Chen Ling, Arlan Walsh, Victor D. Martinez, Jonathan Boucher, Dongmei Zuo, et al. 2020. “An ErbB2 Splice Variant Lacking Exon 16 Drives Lung Carcinoma.” Proceedings of the National Academy of Sciences 117 (33): 20139–48.

Soulette, Cameron M., Eva Hrabeta-Robinson, Carlos Arevalo, Colette Felton, Alison D. Tang, Maximillian G. Marin, and Angela N. Brooks. 2023. “Full-Length Transcript Alterations in Human Bronchial Epithelial Cells with U2AF1 S34F Mutations.” Life Science Alliance 6 (10). https://doi.org/10.26508/lsa.202000641.

Suzawa, Ken, Michael Offin, Adam J. Schoenfeld, Andrew J. Plodkowski, Igor Odintsov, Daniel Lu, William W. Lockwood, et al. 2019. “Acquired MET Exon 14 Alteration Drives Secondary Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor in EGFR-Mutated Lung Cancer.” JCO Precision Oncology 3 (May). https://doi.org/10.1200/PO.19.00011.

Togashi, Yosuke, Hiroshi Mizuuchi, Shuta Tomida, Masato Terashima, Hidetoshi Hayashi, Kazuto Nishio, and Tetsuya Mitsudomi. 2015. “MET Gene Exon 14 Deletion Created Using the CRISPR/Cas9 System Enhances Cellular Growth and Sensitivity to a MET Inhibitor.” Lung Cancer  90 (3): 590–97.

Wang, Feng, Yang Liu, Wanglong Qiu, Elaine Shum, Monica Feng, Dejian Zhao, Deyou Zheng, Alain Borczuk, Haiying Cheng, and Balazs Halmos. 2022. “Functional Analysis of MET Exon 14 Skipping Alteration in Cancer Invasion and Metastatic Dissemination.” Cancer Research 82 (7): 1365–79.

# Chapter 4: Interrogation of Isoform-level Aberrant Splicing in mutant *SF3B1* in collaboration with Dr. Esther Obeng.

## Abstract:

Mutations in the spliceosomal component SF3B1 is the most common aberration in Myelodysplastic Syndromes (MDS). The consequential dysregulated splicing changes work in insidious ways to contribute to overall disease, disrupting erythroid maturation and leading to MDS-associated anemia. The current treatments for MDS-associated anemia aren't curative, and the greatest barrier in developing better therapies is the lack of understanding of the molecular basis of erythroid maturation. While previous studies have used short read sequencing to tease apart splicing changes created by *SF3B1* mutations, this lacks the isoform-context of the splicing events which can limit our ability to design treatment. Therefore, to identify isoform-level splicing events related to MDS-associated anemia, I used Nanopore long read sequencing on K562 cells expressing mutant *SF3B1*. This analysis identified isoforms with cryptic 3' splice site selection anticipated from *SF3B1* mutations, and an unexpectedly high number of exon skipping events in unannotated isoforms. Additionally, I discovered isoforms with aberrant splice changes in genes commonly associated with MDS, and a novel isoform of *TNNI3* which otherwise would not have been identified with short reads. Overall, this work expands the context of misspliced genes from *SF3B1* mutations to identify variants for future characterization in MDS-associated anemia.

## Introduction

Myelodysplastic syndromes (MDS) are a group of blood disorders characterized by abnormal blood cell development and function in the bone marrow. One of the most common features of early MDS is erythroid dysplasia, or the abnormal development of erythroid cells, the precursors to red blood cells. In fact, 82% of MDS patients present with anemia at baseline, and the majority indicate this has a substantial impact on their health and quality of life (Sekeres et al. 2011; Oliva et al. 2012). However, only about 30% of patients treated with erythropoiesis-stimulating agents (ESAs) achieve an erythroid response, and this response tends to be transient (Fenaux et al. 2018). As current treatments aren't curative, this forces a reliance on blood transfusions, with consequences of fluid and iron overload and risk of alloimmunization (Cassanello et al. 2022). A critical barrier in developing more curative therapies for MDS-associated anemia is the lack of understanding of the molecular basis for aberrant erythroid maturation.

*SF3B1* is the most commonly mutated gene in MDS present in about 28% of all MDS cases (Garcia-Manero et al. 2020), and has been recently proposed to be classified as its own distinct MDS subtype (Malcovati et al. 2020). *SF3B1* mutations are often associated with ring sideroblasts in MDS (MDS-RS) occurring in 90% of patients and leads to abnormal red blood cell precursors with aberrant iron accumulation around the nucleus (Malcovati et al. 2020). The most common hotspot mutation is K700E, followed by K666N (Kanagal-Shamanna et al. 2021; Dalton et al. 2020). While both present in MDS-RS, they are associated with different mutational profiles and disease outcomes (Kanagal-Shamanna et al. 2021; Dalton et al. 2020).

SF3B1 is a component of the U2 snRNP and is critical in 3’ splice site selection via branchpoint recognition (Gozani et al. 1996; Krämer 1996). Therefore, it is not surprising that many studies surveying the splicing consequences of mutant *SF3B1* identify predominant aberrant 3’ splice site selection typically 10-30bp upstream of the canonical splice site (Darman et al. 2015; Alsafadi et al. 2016; Wang et al. 2016; DeBoever et al. 2015; Obeng et al. 2016), through the selection of a non-canonical adenosine-rich branch point sequence (Darman et al. 2015; Alsafadi et al. 2016). The vast majority of these cryptic splice site choices introduces early stop codons, leading to ~30-50% of transcripts degraded through Nonsense-Mediated Decay (NMD) and overall downregulation of canonical transcripts (Darman et al. 2015; Alsafadi et al. 2016; Obeng et al. 2016). These studies identified event-level splicing events using short read data, however, this prevents researchers from identifying the isoforms in which these splicing events occur and are impactful (Steijger et al. 2013). Long read sequencing offers an attractive solution by capturing the entire transcript sequence to provide isoform-level context (Bolisetty et al. 2015; Sharon et al. 2013). Nanopore long-read sequencing determines nucleic acid sequences by converting changes in current to bases by either DNA or RNA passing through protein-based nanopores embedded in a membrane (Deamer et al. 2016). Previous work performing long read sequencing on chronic lymphocytic leukemia samples with *SF3B1* mutations identified a deregulation of intron retention events, highlighting how long reads can capture information missed with short read data (Tang et al. 2020).

Previously, our collaborators used a conditional knockin mouse model of *SF3B1*$^{+/K700E}$ to show that expression of mutant *SF3B1* led to a progressive macrocytic anemia, a block in terminal erythroid maturation, erythroid dysplasia, and aberrant RNA splicing in upstream cryptic 3’ splice site selection (Obeng et al. 2016). Based on these findings, they hypothesize that *SF3B1* point mutations lead to the missplicing of genes involved in mitochondrial iron handling and terminal erythroid maturation. In this chapter, I perform long read sequencing on an erythroid isogenic K562 line expressing either *SF3B1* K700E or K666N in order to identify transcript-level aberrant splicing events in genes guiding erythroid development. Based on previous literature, I anticipate these event-level splicing events to be predominantly aberrant 3’ splice site selection and intron retention events. To bolster the raw accuracy expected from Nanopore cDNA sequencing, I used the Rolling Circle Amplification of Concatemeric Consensus (R2C2) Nanopore cDNA sequencing method which increases base accuracy to 94% (Volden et al. 2018; Jain et al. 2017), and determined differential isoform level splicing events with FLAIR (Full-Length Alternative Isoform analysis of RNA) (Tang et al. 2020). This provides confidence that aberrant isoforms created through mutant *SF3B1* are created through aberrant splicing, and not an artifact of sequencing. This is with the goal of shedding light on isoforms that would otherwise be missed by short reads.

## Results

*Validations of R2C2 library*

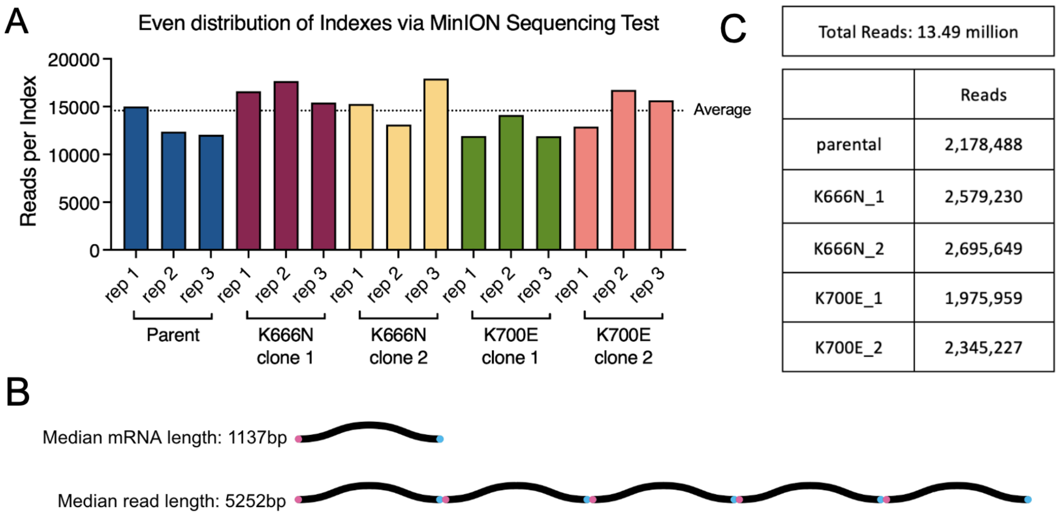
To identify splicing changes in genes related to red blood cell maturation instigated by mutant *SF3B1*, our collaborators used CRISPR/Cas9 knockin and a homology repair template to generate isogenic, heterozygous K652 cell lines that express either the *SF3B1* K700E or K666N mutation. This resulted in two clones heterozygous for each mutation; a total of four cell lines referred to as K700E clone 1, K700E clone 2, K666N clone 1, and K666N clone 2. Additionally, our collaborators used the parental K562 line as a negative control. They grew these cell lines in triplicate, and isolated RNA for future library preparation and long read sequencing.

As splicing factor mutations result in complex splicing changes which creates transcripts dramatically different from annotated references, this project required a library preparation method that would ameliorate the low coverage from Nanopore sequencing. This led us to choose R2C2, which ultimately generates long, concatenated cDNA strands providing incredible accuracy through sequencing these repeats (Volden et al. 2018). Briefly, this method indexes each replicate, circularizes the cDNA, then performs rolling circle amplification which results in long concatenated cDNA molecules. Then I sequence these long DNA strands with Nanopore, and process the basecalled reads using C3POa, the compatible R2C2 software. As I anticipated predominantly alternative 3' splice site selection and decreased intron retention events in SF3B1 mutants, this method ensured any peculiar splicing patterns are accurate representations of isoform expression.

As library preparation can lead to technical variation between samples, I validated the library's index distribution and read length using a pilot MinION flow cell. During library preparation I assigned each R2C2 sample an index; a unique sequence used to distinguish samples from a sample pool. After C3POa demultiplexed the basecalled data, I determined that the pooled library contained an even ratio of each index (Figure 4.1A). This confirms that these samples are pooled in equal proportions, and will be evenly represented upon deeper sequencing. I used a custom Python script to calculate the average mRNA and concatenated read length in this library, which was 1137bp and 5252bp, respectively (Figure 4.1B). This confirms the quality of the library because the average mRNA length is within the normative range in humans (between 1000-3000 bp) (Lopes et al. 2021) and the concatenated reads provide on average about 5x coverage of mRNA.

In order to detect rare genes that may be involved in red blood cell maturation, the goal was to sequence at a coverage of 3-6 million reads per sample. The highest yield flow cell offered by Oxford Nanopore is the PromethION, which results in an average of 8-12 million reads per R2C2 library input. Therefore, I sequenced the library pool across three PromethION flow cells, which in theory should generate 24-36 million reads, or 4.8-7.2 million reads per sample. Using my MinION-validated library, I sequenced across three PromethION flow cells for 72 hours. Unfortunately, after 24 hours the majority of pores died and were unable to accept more cDNA molecules, greatly decreasing the throughput of this experiment. After basecalling and C3POa processing, the final yield was 13.5 million reads with close to 2 million reads per sample (Figure

4.1C). While this final read yield was less than expected, previous studies found key insights from low long read depth of splicing factor mutations (Tang et al. 2020), so I proceeded forward.



Total Reads: 13.49 million

| | Reads |
|---|---|
| parental | 2,178,488 |
| K666N_1 | 2,579,230 |
| K666N_2 | 2,695,649 |
| K700E_1 | 1,975,959 |
| K700E_2 | 2,345,227 |

**Figure 4.1: R2C2 library validated to have even distribution of indexes, even coverage, and expected lengths of transcripts.**

A) Even distribution of indexes across all replicates by sequencing on a MinION.
B) Median length of mRNAs sequenced and read length of the concentric circles.
C) Even read depth across all samples after PromethION sequencing.

*FLAIR quantifies isoform-level splicing and expression events in SF3B1 mutants*

I used FLAIR to identify high-confident isoforms and perform differential expression and splicing event analysis from R2C2 long-read sequencing data. FLAIR works by 1) generating a reference transcriptome containing all isoforms in the dataset, and 2) mapping each demultiplexed sample to these references in order to quantify isoforms in each sample. FLAIR generates this reference transcriptome using the consensus calls from C3POa which contain all isoforms from every sample sequenced. To generate the reference, FLAIR maps these consensus calls to the genome with a spliced aligner, then corrects splice junctions using matched short read data from our collaborators. This short read correction is essential in this analysis, as this will identify unannotated splice sites to improve the confidence of splice junction boundaries (Tang et al. 2020). Finally, FLAIR collapses and filters isoforms into one reference transcriptome. FLAIR quantifies splicing events by mapping demultiplexed samples to this reference transcriptome, resulting in a counts file quantifying isoforms and genes in each sample. As another quality control step, I used these counts files to perform a principal component analysis (PCA), a method to dimensionally reduce data while retaining all information in the data set to visualize relationships of all replicates (Figure 4.2). This PCA plot reveals all replicates cluster with each other as anticipated, with both K700E clones and K666N clone 1 distinctly clustering away from the negative control. Because the K666N clone 2 replicates clustered closely with the parental line, I chose to omit this sample from future differential expression and splicing analyses.

*Differential expression and splicing analysis identified with FLAIR*

The FLAIR counts files are used as input for both differential expression and differential splicing analysis by comparing either the K700E samples (both clone 1 and 2) or the K666N samples (clone 1) to the parental line. The FLAIR differential expression function, `diffExp`, determines differential gene expression (DGE) and differential isoform expression (DIE), both calculated using DESeq2, as well as differential isoform usage (DIU), calculated with DRIMseq.

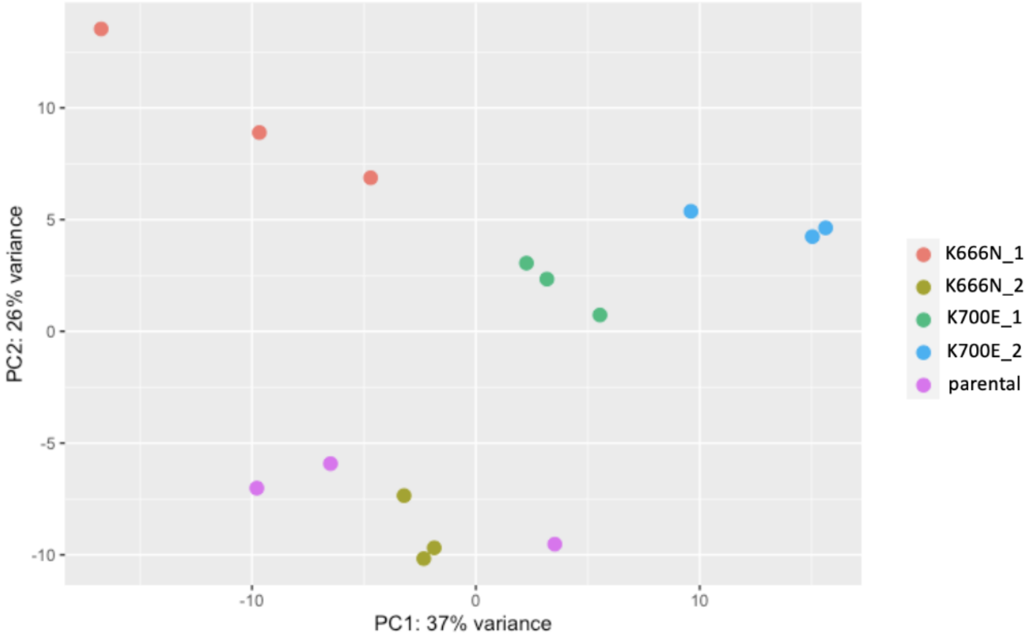


**Figure 4.2: Sample clustering using PCA reveals relationships between samples.**

PCA performed on all genes. Sample key is to the right. Due to close clustering of K666N_2 with the parental line, these samples will be excluded from further analysis.

This analysis not only determines the number of genes and isoforms affected by different expression and usage, but also generates plots to visualize expression differences in genes (Figure S4.1 A and B) and isoforms (Figure S4.1 C and D). These plots reveal similar expression patterns of the most differentially expressed genes and isoforms between both SF3B1 mutants, K700E and K666N.

The FLAIR differential splicing function, `diffSplice`, determines alternative 3' splice site usage (A3), alternative 5' splice site usage (A5), exon skipping (ES) and intron retention (IR) events in the SF3B1 mutants. While key information on its own, it is of keen interest to understand the relationship between these differential splicing and expression events. Therefore, using a custom python script I determined the overlap of differential expression, isoform usage, and splicing events of isoforms in both SF3B1 mutants (Figure 4.3 and events enumerated in Table 4.1). This plot filtered expression changes FC > 1.5, and splicing events with δPSI > 10%.

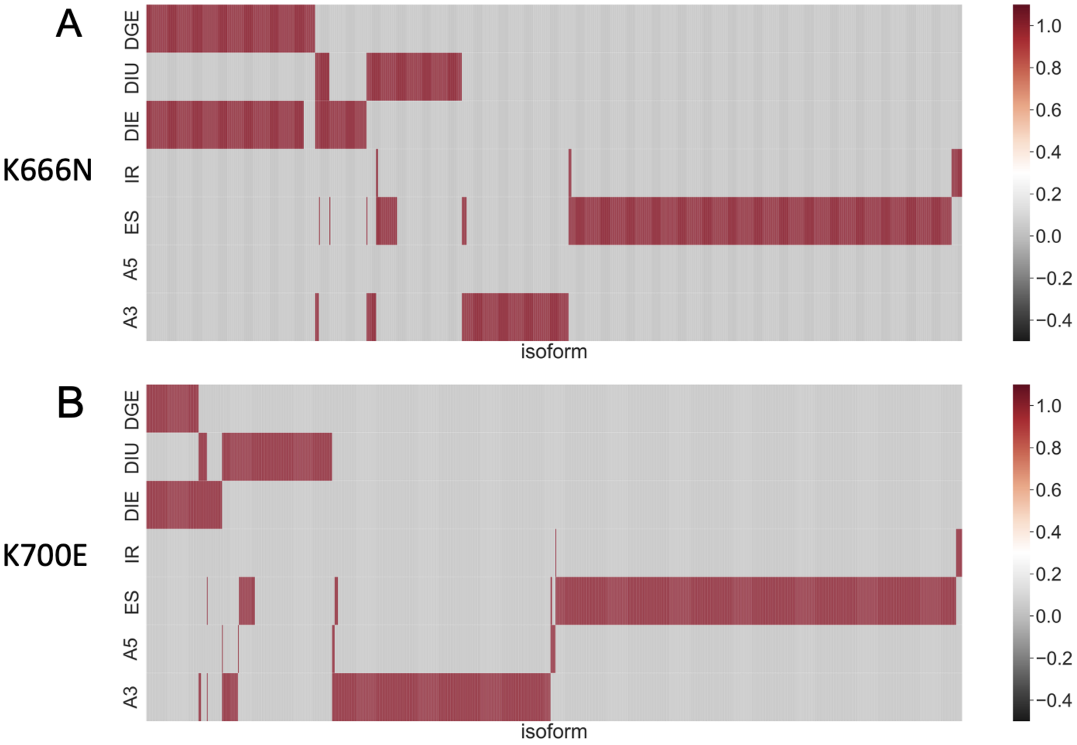
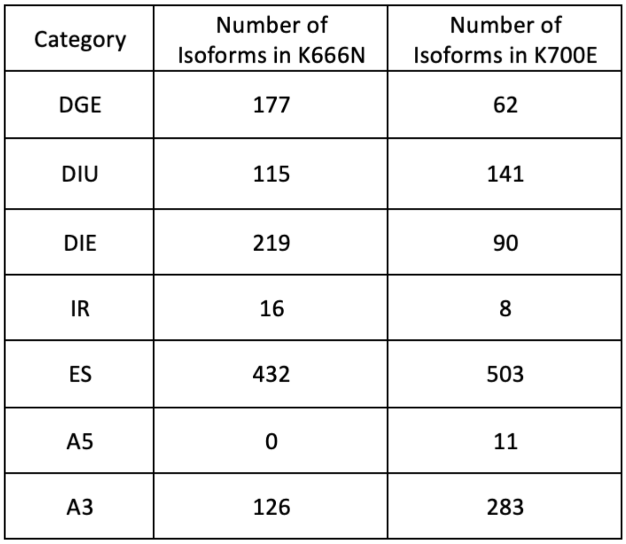


**Figure 4.3: Mutant SF3B1-associated splicing alterations.**

A) in K666N samples and B) K700E samples. Each block represents an isoform, red indicates a significant change and gray indicates no change. DGE = differential gene expression. DIU = differential isoform usage. DIE = differential isoform expression. IR = intron retention. ES = exon skipping. A5 = alternative 5' splice site selection, A5 = alternative 3' splice site selection.

**Table 4.1: Enumerated mutant SF3B1-associated splicing alterations.**
Categories are as described in figure 4.3



| Category | Number of Isoforms in K666N | Number of Isoforms in K700E |
|---|---|---|
| DGE | 177 | 62 |
| DIU | 115 | 141 |
| DIE | 219 | 90 |
| IR | 16 | 8 |
| ES | 432 | 503 |
| A5 | 0 | 11 |
| A3 | 126 | 283 |

Overall, these data reveal that, while DGE overlapped well with DIE, neither expression of genes or isoforms tended to overlap with DIU. It is possible DIU overlaps with other expression events that are missed due to filtering. However, as DRIMseq provides insight into isoform abundance independent of expression changes, DIU provides a list of isoforms altered by mechanisms like splicing. DIU also does not overlap with the majority of splicing events. Aside from filtering, this could insinuate that these isoforms are created through splicing independent mechanisms, like alternative promoter or termination sites, or more complex splicing events that FLAIR is unable to detect.

Expanding on differential expression and isoform usage, this analysis allows us to visualize the distribution of splicing events with isoform usage > 10%. As anticipated, these *SF3B1* mutants induced a high number of A3 events, and a low number of A5 events. As I anticipated a decrease in IR events based on long read primary sample sequencing with *SF3B1* mutations (Tang et al. 2020), I was surprised by the lack of differential IR events identified in these cell lines. Additionally, this analysis revealed a large number of unanticipated ES events, which were corroborated by short read data from our collaborators (data not shown). Interestingly, the vast majority of differential expression and isoform usage events occurred in annotated isoforms (Figure S4.2 A and C) while the majority of the ES and A3 events occurred in unannotated isoforms (Figure S4.2 B and D). Overall, these findings reveal global, isoform-level splicing changes in a model system for erythroid differentiation. Future work from our collaborators will characterize the productivity and functional consequences of these isoforms.

*Impactful isoforms identified in Long Read Data implicated in MDS*

Long reads are critical for isoform-specific context of these splicing events, and in this data I identified several isoforms otherwise missed by short read data. I identified aberrant 3’ splice site choice in *TMEM14C* and *DYNLL1*, two variants implicated in MDS (Dolatshad et al. 2016 and Figure 4.4). These events were incorrectly identified from the matched short read data. Aberrant 3’ splice site selection introduces an addition of 14 base pairs in the 5′UTR region of both these genes (Dolatshad et al. 2016). For *TMEM14C,* previous work demonstrated that this longer UTR led to a decrease in translational efficiency and an overall 40% reduction of endogenous TMEM14C protein levels (Clough et al. 2022). Interestingly, 5’UTR extension in *DYNLL1* had the opposite effect, correlating with *DYNLL1* overexpression (Tam et al. bioRxiv). *TMEM14C* plays an important role in the terminal steps of the heme synthesis pathway, and decreased expression is associated with aberrant iron accumulation in the

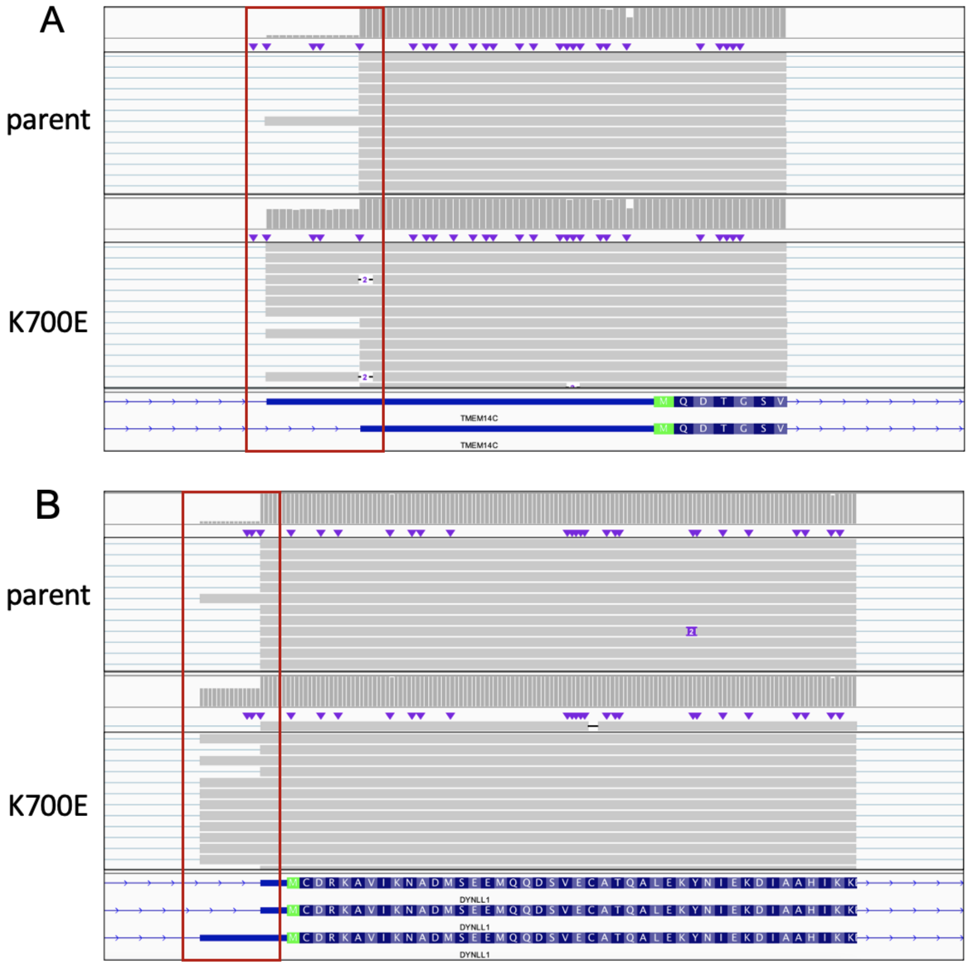


**Figure 4.4: Alternative 3' splice site choice the 5' UTRs MDS-implicated genes *TMEM14C* and *DYNLL1*.**

Alternative 3' splice site choice in the 5'UTRs of A) *TMEM14C* and B) *DYNLL1*. Alternative splice site selection boxed in red. Visualized in Interactive Genomics Viewer (IGV).

mitochondria of erythroid cells (Yien et al. 2014; Clough et al. 2022). *DYNLL1,* on the other hand, has a well-established role in DNA repair pathways, and its overexpression is associated with genomic instability (Tam et al. bioRxiv). While *DYNLL1* does not play a direct role in erythroid maturation, this gene is often implicated in MDS (Dolatshad et al. 2016; Tam et al. bioRxiv). Furthermore, *SF3B1* mutations are known to dysregulate DNA damage response (Wang et al. 2016). While the 5'UTR difference in *TMEM14C* distinguishes both of itis isoforms, making this isoform identifiable by short read data, the long reads of *DYNLL1* identified complex splicing patterns in the 5'UTR in addition to the cryptic 3' splice site choice. This overall leads to at least four isoforms with different 5'UTR lengths, likely leading to different expression efficiency in each isoform (Fig 4.5). Without long reads, these *DYNLL1* isoforms would otherwise be missed with short read data, suggesting our collaborators will have the ability to identify novel dysregulated isoforms impactful in erythroid maturation from these data.

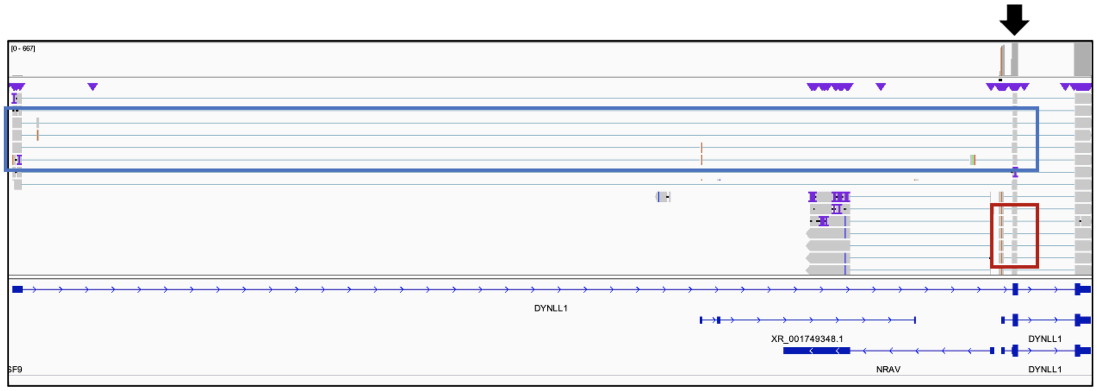


**Figure 4.5: Multiple isoforms impacted with *DYNLL1* alternative 3' splice site choice.**

Entire *DYNLL1* gene revealing isoform context of 3' splice site choice in the 5' UTR. Black arrow designates aberrant 3' splice site choice (Figure 4.4B for close up resolution). Blue box indicates splicing to distal 5' UTR exon. Red box indicates splicing to the proximal 5' UTR exon. Visualized in Interactive Genomics Viewer (IGV).

Another interesting case is a novel isoform of *TNNI3*, a cardiac-specific gene important in the regulation of muscle contraction. This is a part of the troponin complex which is composed of three proteins: Troponin I (TnI), encoded by *TNNI3* in cardiac muscle, Troponin T (TnT), and Troponin C (TnC) (Katrukha 2013). These three subunits work together with tropomyosin to promote muscle relaxation in the absence of calcium. In the long read data, I identified a shorter isoform of *TNNI3* utilized predominantly in the wild type, with the longer isoform used predominantly in the SF3B1 mutant (Figure 4.6A). To the best of my knowledge, this is a previously uncharacterized isoform of *TNNI3* generated through an alternative promoter upstream of exon 5. While the longer isoform utilizes an initiating methionine in exon 1, the shorter isoform is forced to use a methionine at position 154 in exon 7, with both isoforms terminating in exon 8 (Figure 4.6B). This shorter isoform lacks both the TnT and TnC binding regions, however retains the C-terminal domain which binds actin (Creso and Campbell 2021; Kühnisch et al. 2019) (Figure 4.6C).

The vast majority of point mutations known to cause cardiomyopathies, diseases that affect structure and function of the heart, occur in the C-terminal region of *TNNI3* (Chen et al. 2014; Mogensen et al. 2003; Kühnisch et al. 2019). This highlights this region's importance and suggests this truncated protein could retain another function. Because this protein contains the region required for actin-tropomyosin binding (Figure 4.6C), it is possible this protein is involved in the regulation of muscle contraction independent of calcium. Alternatively, there is a body of work revealing TnI retains non-canonical functions in the nucleus (Sahota et al. 2009; Lu et al. 2022). Although I estimate that this shorter *TNNI3* isoform generates a truncated protein, alternatively it is possible this shorter isoform is not translated and fulfills a regulatory role by decreasing the overall amount of functional TnI through NMD. This is supported by the observation of aberrant overexpression of *TNNI3* in both kidney and lung cancer, suggesting a potential undiscovered role in tumorigenesis and reason for wild type cells to decrease its expression (Chen et al. 2014; Zhao et al. 2021).

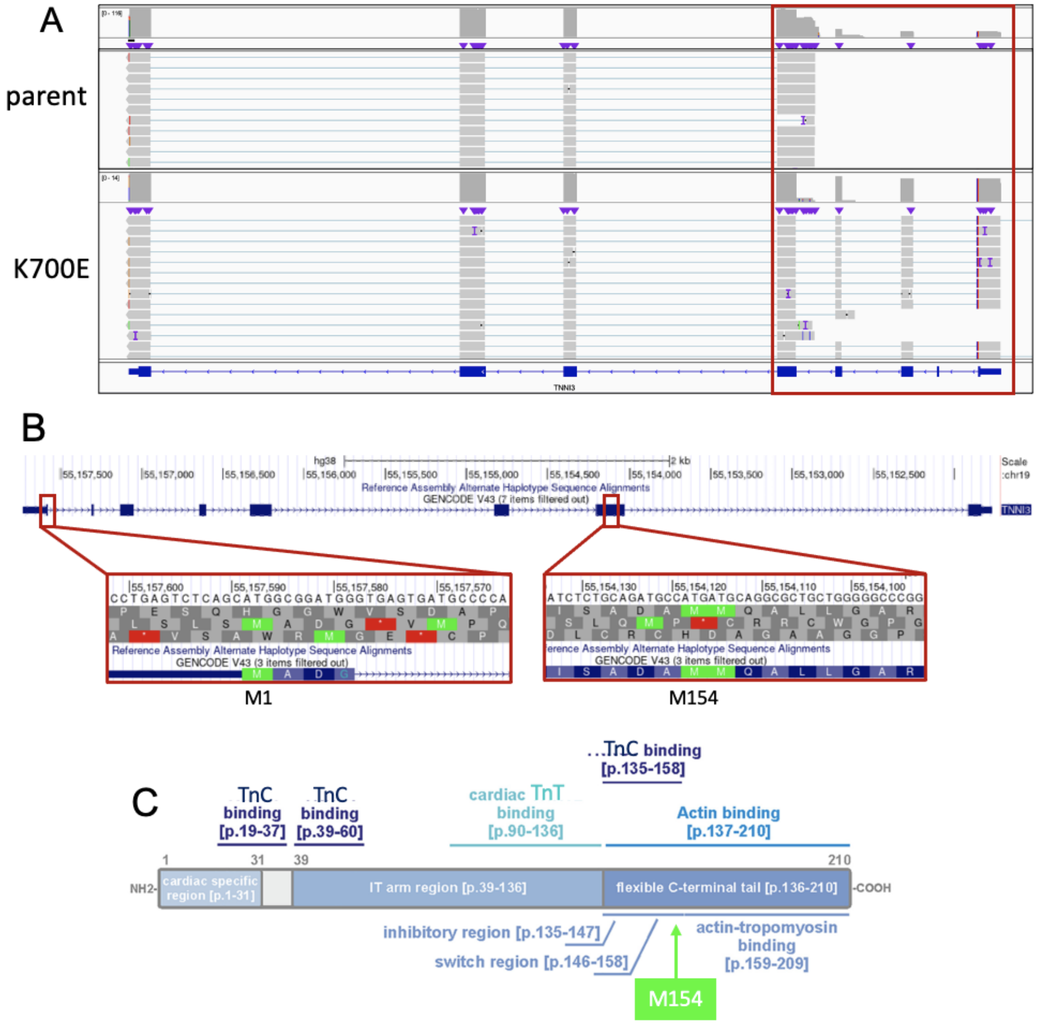

A
parent
K700E
TNNI3
B
hg38
2 kb
Scale
chr19
55,157,500
55,157,000
55,156,500
55,156,000
55,155,500
55,155,000
55,154,500
55,154,000
55,153,500
55,153,000
55,152,500
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V43 (7 items filtered out)
TNNI3
55,157,600
55,157,590
55,157,580
55,157,570
CCTGAGTCTCAGCATGGCGGATGGGTGAGTGATGCCCA
GENCODE V43 (3 items filtered out)
M1
55,154,130
55,154,120
55,154,110
55,154,100
ATCTCTGCAGATGCCATGATGCAGGCGCTGCTGGGGGCCGG
M154
C
TnC binding [p.19-37]
TnC binding [p.39-60]
cardiac TnT binding [p.90-136]
...TnC binding [p.135-158]
Actin binding [p.137-210]
1
31
39
210
NH2-
cardiac specific region [p.1-31]
IT arm region [p.39-136]
flexible C-terminal tail [p.136-210]
-COOH
inhibitory region [p.135-147]
switch region [p.146-158]
actin-tropomyosin binding [p.159-209]
M154

**Figure 4.6: Alternative promoter in *TNNI3* detected with long reads.**

A) Entire *TNNI3* isoform revealing alternative promoter use in the *SF3B1* mutant. Note this is in the reverse strand, so gene is transcribed right to left. B) Likely alternative translation initiation codon in the shorter *TNNI3* isoform. This gene in the reverse strand is flipped to appear on the forward strand specified on the UCSC genome browser. C) Regions on TNNI3 and alternative translation site at

M154 within the actin binding region of the protein. Image adapted from Kühnisch et al. 2019 © John Wiley & Sons, Inc.

Although to the best of my knowledge there is no direct link between MDS-induced anemia and *TNNI3*, cardiac complications are common in MDS (Gattermann 2018; Delea et al. 2009), suggesting this *TNNI3* isoform may be tangentially related. For all aberrant isoforms induced by mutant SF3B1, without the use of long reads in this data set, all these isoforms would be missed using short read data alone. This highlights the necessity of using long reads in sequencing studies related to aberrant splicing.

### *Discussion*

The goal of this collaboration was to detect isoform-level splicing events related to red blood cell maturation induced by mutant SF3B1. Going into this study, I anticipated an increase in aberrant 3' splice site choice and a decrease in intron retention events (Darman et al. 2015; Alsafadi et al. 2016; Wang et al. 2016; DeBoever et al. 2015; Tang et al. 2020). However, while I saw the expected increase in cryptic 3' splice site choice, I did not identify substantial alternative intron retention events (Fig. 4.3). While this could be due to how I size selected the R2C2 library before sequencing, it is also possible this is an artifact of the K562 cell line model used in this study, as *Tang et al.* performed long read sequencing on CLL primary samples with SF3B1 mutations and found a decrease in intron retention events. Surprisingly, I also identified a majority of exon skipping events, which was corroborated with our collaborator's findings in the matched short read data. While the vast majority of studies identify aberrant 3' splice site selection, substantial alternative exon skipping events are also implicated in *SF3B1* mutants (Kanagal-Shamanna et al. 2021). A known transcriptome-wide consequence of SF3B1 mutations is downregulation of genes through NMD, as cryptic 3’ splice site choice can alter the reading frame to introduce early termination codons (Darman et al. 2015; Alsafadi et al. 2016; Obeng et al. 2016). Aside from aberrant 3’ splice site selection, *SF3B1* mutations are also tied to intron retention and cryptic poison exon inclusion which overall decrease gene expression (Inoue et al. 2019; Lieu et al. 2022). In fact, poison exon inclusion and subsequence NMD is a recognized mechanism employed by cells to control gene expression (Weischenfeldt et al. 2012). Therefore, it is possible these exons are typically poison exons required for maintaining gene expression, which is dysregulated by mutant *SF3B1* altering their overall expression in the mutant.

While the most common consequence of mutant *SF3B1*-induced splicing changes degrade transcripts through NMD, occasionally these mutations alter expression through UTR modification, or generate functional proteins with non-canonical functions (Clough et al. 2022; Tam et al. ; Visconte et al. 2015; Bondu et al. 2019). In our dataset, I identified three genes modulated in this way that were missed in the matched short read data. I identified cryptic 3’ splice site selection in two genes often implicated in *SF3B1*-mutant MDS: *TMEM14C* and *DYNLL1*. The isoform level-context of *DYNLL1* revealed that this aberrant splice site choice affected multiple isoforms, resulting in a variety of 5’ UTR lengths (Fig. 4.5). As the +14bp 5’ UTR extension is known to cause overall *DYNLL1* overexpression, it is likely the different UTR lengths on these *DYNLL1* isoforms differentially affect expression as well (Tam et al. bioRxiv). Without long reads, it would be impossible to determine the different isoforms associated with the cryptic 3’ splice site.

In addition to the isoform context of these splicing events, long reads capture splicing independent mechanisms of isoform expression, such as the alternative promoter use I observed in TNNI3. This shorter isoform may either generate a truncated protein through an initiating methionine at position 154, or serve a regulatory role to decrease TNNI3 in cardiac tissue. It is possible alternative promoter use is more widespread with SF3B1 mutations, perhaps through to aberrant splicing of chromatin modulators themselves (Inoue et al. 2019), but due to our reliance on short read data researchers are unable to detect these isoforms. While *TNNI3* is not directly involved in erythroid maturation, cardiac complications are common in MDS (Gattermann 2018; Delea et al. 2009), drawing the connection between this gene and aberrant blood cell maturation. Due to this connection, its possible SF3B1-induced aberrant isoforms in cardiopathies and MDS work together in a subtle yet insidious way to exacerbate both conditions. With long read sequencing providing this greater context otherwise missed by short read sequencing, it is possible to ask these questions at the depth afforded by this larger picture.

## Chapter 4 Materials and Methods

*Generation of SF3B1 cell lines (do I need to put or say got from Esters lab)*

Our collaborators generated K652 heterozygous isogenic cell lines containing SF3B1 K700E or K666N using CRISPR homology directed repair to knock-in mutations into the genome.

*R2C2 library preparation and Nanopore sequencing*

RNA concentrations from all 15 samples were quantified via Qubit 3.0 Fluorometer and validated for purity using the Agilent TapeStation 4150 system. To synthesize cDNA from RNA, each unique oligo-dt index (below for sequences) was added in a concentration of 1µm index to 1mM dNTPs (Promega, U151B) to a PCR tube.

| | |
|---|---|
| Oligo_dT_Index1 | 5’-AAGCAGTGGTATCAACGCAGAGT CGCTCAGTTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index2 | 5’-AAGCAGTGGTATCAACGCAGAGT TATCTGACCT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index3 | 5’-AAGCAGTGGTATCAACGCAGAGT ATATGAGACG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index4 | 5’-AAGCAGTGGTATCAACGCAGAGT CTTATGGAAT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index5 | 5’-AAGCAGTGGTATCAACGCAGAGT TAATCTCGTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index6 | 5’-AAGCAGTGGTATCAACGCAGAGT GCGCGATGTT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index7 | 5’-AAGCAGTGGTATCAACGCAGAGT AGAGCACTAG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index8 | 5’-AAGCAGTGGTATCAACGCAGAGT TGCCTTGATC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index9 | 5’-AAGCAGTGGTATCAACGCAGAGT CTACTCAGTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index10 | 5’-AAGCAGTGGTATCAACGCAGAGT TCGTCTGACT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index11 | 5’-AAGCAGTGGTATCAACGCAGAGT GAACATACGG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index12 | 5’-AAGCAGTGGTATCAACGCAGAGT CCTATGACTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index13 | 5’-AAGCAGTGGTATCAACGCAGAGT accgtgtcag ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index14 | 5’-AAGCAGTGGTATCAACGCAGAGT agacgtcatc ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |
| Oligo_dT_Index15 | 5’-AAGCAGTGGTATCAACGCAGAGT tacaatcagg ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3’ |

200ng RNA was added to the oligo dt and dNTP tube. The RNA mix was incubated for 3 minutes at 37℃, then snap cooled on ice for 1 minute. The following cDNA synthesis reaction mix was added to each tube: 2µl 5x First-Strand Buffer (Takara, ST0062), 1µl 100mM DTT (Takara, ST0063), 0.3µl TSO primer at 10µM (see below for primer sequence), 1.45µl water, 0.25µl Superaesin (ThermoFisher Scientific, AM2694), and 1µl SMARTscribe reverse transcriptase (Takara, ST0065).

| | |
|---|---|
| TSO-Smart-seq2 oligo | 5'-AAGCAGTGGTATCAACGCAGAGTACATrGrGrG-3' |

Tubes were placed in the thermocycler under the following conditions: 42℃ for 1.5 hours, 70℃ for 5 minutes, and held at 4℃, resulting in first strand cDNA synthesis. To amplify this synthesized cDNA, 1µl 10µM ISPCR primer (sequence below), 12.5µl Kapa HiFi Hot Start (Kapa Biosystems, KK2602), 0.75µl RNAse A (ThermoFisher Scientific, EN0531), and 0.75µl Lambda exonuclease (New England BioLabs, M0262S) was added to each PCR tube.

| | |
|---|---|
| ISPCR primer | 5'-AAGCAGTGGTATCAACGCAGAGTAC-3' |

After gentle mixing, the thermocycler was run under the following conditions: 37℃ 30 minutes, 95℃ 3 minutes, 15 cycles of 98℃ for 20 seconds followed by 67℃ for 15 seconds then 72℃ 5 minutes, and finally 72℃ 10 minutes and held at 4℃.

The amplified cDNA was purified using a bead cleanup. Room Temperature AMPure XP beads (Beckman Coulter, A63880) were gently vortexed and used in a ratio of 0.8 bead: 1 sample in DNA loBind tubes (Eppendorf, 022431021). The cDNA samples were transferred to their corresponding DNA loBind tube, mixed with pipetting, and quickly spun down.

After incubation at room temperature for 10 minutes, tubes were placed on a magnetic rack for 5 minutes to separate beads from the fluid. Supernatant was removed and beads washed with 600μl freshly made 70% ethanol, added slowly as to not disturb the beads. After incubating the beads in 70% ethanol for one minute, ethanol was removed and the wash was repeated once more. After removing the second ethanol wash, beads were spun down and tubes placed back on the magnetic rack for 5 minutes. Using a P10 pipette, the remaining ethanol was aspirated while exercising caution to not disturb the beads. With the tube open on the rack, beads were allowed to dry for 2-3 minutes. 20μl water was used to elute the sample after a 10 minute incubation at 37℃. After incubation, tubes were placed back on the magnetic rack for 5 minutes, then the supernatant containing the pure cDNA was transferred to a new tube.

All 15 samples were pooled in equal proportions into a new tube for further sample preparation. First the cDNA was quantified with the Qubit 3.0 Fluorometer, and 200ng of each individual sample added into a new tube. To circularize the cDNA, the pooled cDNA was used in a 1:1 ratio with the UMI splint (protocol to generate splint below) - typically ranging from 100-200ng of each in a PCR tube with 10μl volume. If under 10μl, the volume was adjusted with water. To this 10μl cDNA pool:splint mix, 10μl 2x NEBuilder Assembly Mix (New England BioLabs, M5520AA) was added then these tubes were incubated in the thermocycler for 1 hour at 50℃. After incubation, 3μl exonuclease I (New England BioLabs, M0293S), 3μl exonuclease III (New England BioLabs, M0206S), 3μl Lambda exonuclease (New England BioLabs, M0262S), 5μl

NEBuffer 2 (New England BioLabs, B7002S), and 16µl water were added to each PCR tube. This mixture was left to incubate at 37℃ for 6 hours-overnight.

The circularization reaction was deactivated at 80℃ for 20 minutes, then transferred to a DNA loBind tube to bead purify the cDNA as done previously. During the final isolation step 52µl water was used to elute samples, and eluate was transferred into PCR tubes of 10µl each. To perform rolling circle amplification, 5µl Phi29 buffer (New England BioLabs, B0269S), 2.5µl 10mM dNTPs (Promega, U151B), 2.5µl exo-resistant random hexamer primers (ThermoFisher Scientific, SO181), 29µl water, and 1µl Phi29 enzyme (New England BioLabs, M0269L) was added to each PCR tube containing circularized cDNA. These reactions were incubated in a thermocycler at 30℃ overnight.

The next day, 4µl T7 endonuclease (New England BioLabs, M0302S) was added to each PCR tube and samples were pipetted slowly to mix the long cDNA. These tubes were incubated at 37℃ for 2 hours, and taken out occasionally to mix with finger vortexing. After incubation, a 26 gauge needle (Becton Dickinson, 305120) and syringe (Millipore Sigma, Z683531) were used to shear the DNA into a new 2mL tube. The Zymo cc-5 clean and concentrator kit (Zymo Research, D4004) purified cDNA using a ratio of 2:1 binding buffer to sample, using a maximum input of 5µg per column. The purified cDNA was eluted in 20µl, and all combined reactions were quantified using the Qubit.

The resulting long cDNA strands were gel purified using a 1% SeaPlaque$^{TM}$ GTG$^{TM}$ Agarose gel (Lonza, 50111) in 1X TAE. A maximum of 3µg sample was loaded per well mixed with 6x loading dye (New England BioLabs, B7021S) and sybr gold (Invitrogen, S11494). The gel was run at 80V until the bands had the desired resolution. Gel bands between 5-10kb were cut from the gel and a maximum of 600mg of gel added per tube. Enough beta-agarose buffer was used to completely cover each gel slice, and incubated at 4℃ for 20 minutes. After buffer removal this step was repeated. The dry gel was melted at 65℃ for 10 minutes, then equilibrated to room temperature for 1 minute. Beta-agarase enzyme (New England BioLabs, M0392S) was used in a ratio of 3µl enzyme per 300mg melted gel, and placed at 42℃ overnight to completely digest the gel. The next day, the tubes were flash cooled on ice for 5 minutes, then spun at 15,000xg for 7 minutes. Up to 300µl liquid was bead purified per DNA IoBind tube, using a bead ratio of 0.7:1 sample and eluting in 50µl water. The Nanopore SQK-LSK110 library prep kit was used per the kit's instructions to prepare for both MinION and PromethION sequencing. The MinION test used a single flow cell run for 3 hours, and PromethION flow cells were used in triplicate and later combined.

### *Generate UMI Splint*

Across 6 PCR tubes the following reagents were mixed: 23µl water, 25µl Kapa Hifi Readymix (Kapa Biosystems, KK6202), and 1µl 100mM UMI_Splint_6_F and UMI_Splint_6_R (below).

UMI_Splint_6_F 5'-ACTCTGCGTTGATACCACTGCTT GAGTTTAGCACATGACTGGT NNNNNTATATNNNNN ACGTCTCTGAACTTTTACTCTGCTTATTTATCTAGTTATTTAGCATGCGTAGATGGAGCTGATTAC-3'

UMI_Splint_6_R 5'-ACTCTGCGTTGATACCACTGCTT CTAGGGAACGCTTATATTAG NNNNNATATANNNNN TAGCCACTATTCCAATCCTCCAGTTTAATCGACTAAGAGTTGTAACCGGCCTAAAACATCTAAA GTAATCAGCTCCATCTACGC-3'

Tubes were placed in the thermocycler at 95℃ for 3 minutes, 98℃ for 1 minute, 62℃ for 1 minute, 72℃ for 6 minutes, and held at 4℃. Spling was cleaned with the Zymo cc-5 kit (Zymo Research, D4004) before use.

*Computational analysis of R2C2 library*

Raw fast5 files from Nanopore sequencing were basecalled with guppy v2.3.5, using the high accuracy config file dna_r9.4.1_450bps_hac.cfg. The resulting fastq files were concatenated into one larger fastq file to be used as input in the R2C2 compatible software, C3POa 2.4.0. C3POa filters and demultiplexes the files, resulting in extremely accurate mRNA transcript sequences in the form of fasta files. C3POa outputs fasta files containing filtered reads for all samples ("Consensus" reads), as well as demultiplexed reads ("Sample" reads). The FLAIR pipeline maps isoforms against a reference transcriptome, which is generated from the Consensus reads. First, `FLAIR align` mapped the Consensus file to the hg38 genome using minimap2, then `FLAIR correct` corrected splice junctions using matched short read sequencing data collected from our collaborators. This matched short read data was aligned to the hg38 genome using STAR-2.7.10b. Finally, the isoforms are collapsed using `FLAIR collapse` into one reference transcriptome resulting in a map of unique isoforms uniquely associated with this data. All demultiplexed Sample files were mapped to this reference using `FLAIR quantify` in order to quantify individual isoforms in each sample. The output file from this function is used as input in all other FLAIR differential expression and splicing event analyses. `FLAIR diffExp` determined differential expression of isoforms and genes, as well as isoform usage, and `FLAIR diffSplice` identified alternative splicing events. Custom python scripts were used to parse and plot the output from FLAIR.

## Chapter 4 Supplementary Figures

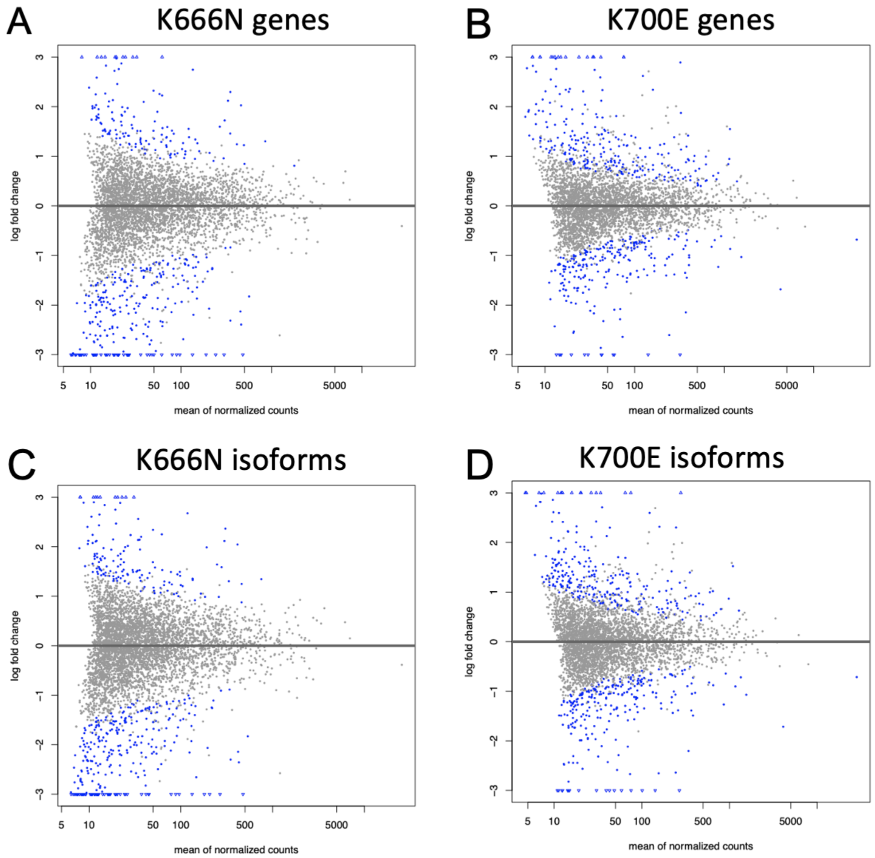


**Figure S4.1: Analysis with FLAIR reveals differentially expressed genes and isoforms in samples with SF3B1 mutations.**

Differentially expressed genes in the SF3B1 K666N samples (A) and SF3B1 K700E samples. Differentially expressed isoforms in the SF3B1 K666N samples and K700E samples.

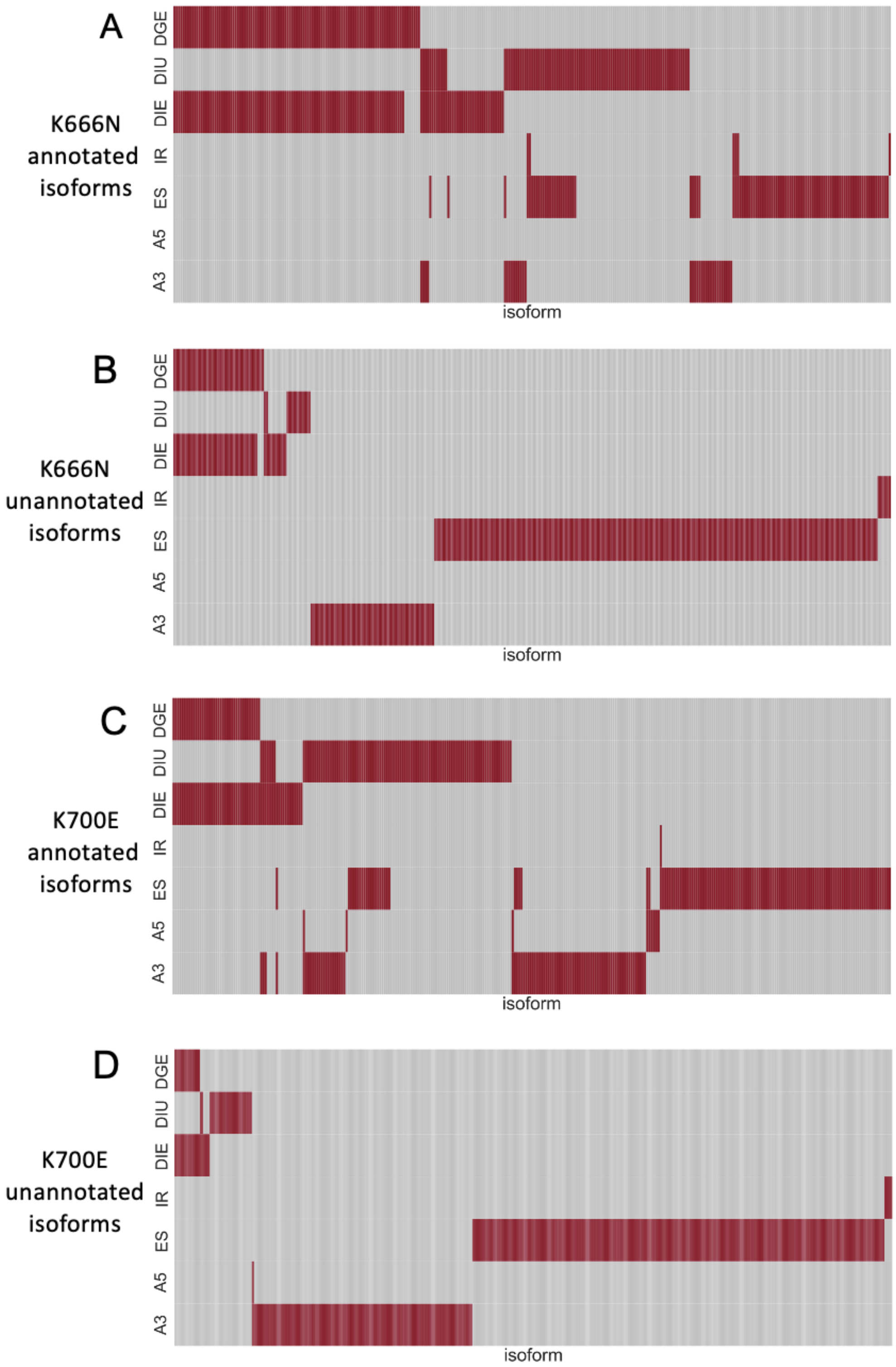

A
K666N
annotated
isoforms
DGE
DIU
DIE
IR
ES
A5
A3
isoform
B
K666N
unannotated
isoforms
DGE
DIU
DIE
IR
ES
A5
A3
isoform
C
K700E
annotated
isoforms
DGE
DIU
DIE
IR
ES
A5
A3
isoform
D
K700E
unannotated
isoforms
DGE
DIU
DIE
IR
ES
A5
A3
isoform

**Figure S4.2: Distribution of splicing alterations induced by mutant SF3B1.**

Distribution of splicing events in K666N annotated (A) and unannotated (B) isoforms, and K700E annotated (A) and unannounced (B) isoforms. A red box indicates a significant isoform for specific categories.

## References

Alsafadi, Samar, Alexandre Houy, Aude Battistella, Tatiana Popova, Michel Wassef, Emilie Henry, Franck Tirode, et al. 2016. “Cancer-Associated SF3B1 Mutations Affect Alternative Splicing by Promoting Alternative Branchpoint Usage.” Nature Communications 7 (1): 1–12.

Bolisetty, Mohan T., Gopinath Rajadinakaran, and Brenton R. Graveley. 2015. “Determining Exon Connectivity in Complex mRNAs by Nanopore Sequencing.” Genome Biology 16 (September): 204.

Bondu, Sabrina, Anne-Sophie Alary, Carine Lefèvre, Alexandre Houy, Grace Jung, Thibaud Lefebvre, David Rombaut, et al. 2019. “A Variant Erythroferrone Disrupts Iron Homeostasis in SF3B1-Mutated Myelodysplastic Syndrome.” Science Translational Medicine 11 (500). https://doi.org/10.1126/scitranslmed.aav5467.

Cassanello, Giulio, Raffaella Pasquale, Wilma Barcellini, and Bruno Fattizzo. 2022. “Novel Therapies for Unmet Clinical Needs in Myelodysplastic Syndromes.” Cancers 14 (19). https://doi.org/10.3390/cancers14194941.

Chen, Chao, Jia-Bao Liu, Zhi-Ping Bian, Jin-Dan Xu, Heng-Fang Wu, Chun-Rong Gu, Yi Shi, Ji-Nan Zhang, Xiang-Jian Chen, and Di Yang. 2014. “Cardiac Troponin I Is Abnormally Expressed in Non-Small Cell Lung Cancer Tissues and Human Cancer Cells.” International Journal of Clinical and Experimental Pathology 7 (4): 1314–24.

Chen, Yan, Shiwei Yang, Jun Li, Gannan Wang, Yuming Qin, Daowu Wang, and Kejiang Cao. 2014. “Pediatric Restrictive Cardiomyopathy due to a Heterozygous Mutation of the TNNI3 Gene.” Journal of Biomedical Research 28 (1): 59–63.

Clough, Courtnee A., Joseph Pangallo, Martina Sarchi, Janine O. Ilagan, Khrystyna North, Rochelle Bergantinos, Massiel C. Stolla, et al. 2022. “Coordinated Missplicing of TMEM14C and ABCB7 Causes Ring Sideroblast Formation in SF3B1-Mutant Myelodysplastic Syndrome.” Blood 139 (13): 2038–49.

Creso, Jenette G., and Stuart G. Campbell. 2021. “Potential Impacts of the Cardiac Troponin I Mobile Domain on Myofilament Activation and Relaxation.” Journal of Molecular and Cellular Cardiology 155 (June): 50–57.

Dalton, W. Brian, Eric Helmenstine, Lisa Pieterse, Bing Li, Christopher D. Gocke, Joshua Donaldson, Zhijian Xiao, et al. 2020. “The K666N Mutation in SF3B1 Is Associated with Increased Progression of MDS and Distinct RNA Splicing.” Blood Advances 4 (7): 1192–96.

Darman, Rachel B., Michael Seiler, Anant A. Agrawal, Kian H. Lim, Shouyong Peng, Daniel Aird, Suzanna L. Bailey, et al. 2015. “Cancer-Associated SF3B1

Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point." Cell Reports 13 (5): 1033–45.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. "Three Decades of Nanopore Sequencing." Nature Biotechnology 34 (5): 518–24.

DeBoever, Christopher, Emanuela M. Ghia, Peter J. Shepard, Laura Rassenti, Christian L. Barrett, Kristen Jepsen, Catriona H. M. Jamieson, Dennis Carson, Thomas J. Kipps, and Kelly A. Frazer. 2015. "Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3' Splice Site Selection in SF3B1-Mutated Cancers." PLoS Computational Biology 11 (3): e1004105.

Delea, Thomas E., May Hagiwara, and Pradyumna D. Phatak. 2009. "Retrospective Study of the Association between Transfusion Frequency and Potential Complications of Iron Overload in Patients with Myelodysplastic Syndrome and Other Acquired Hematopoietic Disorders." Current Medical Research and Opinion 25 (1): 139–47.

Dolatshad, H., A. Pellagatti, F. G. Liberante, M. Llorian, E. Repapi, V. Steeples, S. Roy, et al. 2016. "Cryptic Splicing Events in the Iron Transporter ABCB7 and Other Key Target Genes in SF3B1-Mutant Myelodysplastic Syndromes." Leukemia 30 (12): 2322–31.

Fenaux, Pierre, Valeria Santini, Maria Antonietta Aloe Spiriti, Aristoteles Giagounidis, Rudolf Schlag, Atanas Radinoff, Liana Gercheva-Kyuchukova, et al. 2018. "A Phase 3 Randomized, Placebo-Controlled Study Assessing the Efficacy and Safety of Epoetin-α in Anemic Patients with Low-Risk MDS." Leukemia 32 (12): 2648–58.

Garcia-Manero, Guillermo, Kelly S. Chien, and Guillermo Montalban-Bravo. 2020. "Myelodysplastic Syndromes: 2021 Update on Diagnosis, Risk Stratification and Management." American Journal of Hematology 95 (11): 1399–1420.

Gattermann, Norbert. 2018. "Iron Overload in Myelodysplastic Syndromes (MDS)." International Journal of Hematology 107 (1): 55–63.

Gozani, O., R. Feld, and R. Reed. 1996. "Evidence That Sequence-Independent Binding of Highly Conserved U2 snRNP Proteins Upstream of the Branch Site Is Required for Assembly of Spliceosomal Complex A." Genes & Development 10 (2): 233–43.

Inoue, Daichi, Guo-Liang Chew, Bo Liu, Brittany C. Michel, Joseph Pangallo, Andrew R. D'Avino, Tyler Hitchman, et al. 2019. "Spliceosomal Disruption of the Non-Canonical BAF Complex in Cancer." Nature 574 (7778): 432–36.

Jain, Miten, John R. Tyson, Matthew Loose, Camilla L. C. Ip, David A. Eccles, Justin O'Grady, Sunir Malla, et al. 2017. "MinION Analysis and Reference

Consortium: Phase 2 Data Release and Analysis of R9.0 Chemistry." F1000Research 6 (May): 760.

Kanagal-Shamanna, Rashmi, Guillermo Montalban-Bravo, Koji Sasaki, Faezeh Darbaniyan, Elias Jabbour, Carlos Bueso-Ramos, Yue Wei, et al. 2021. "Only SF3B1 Mutation Involving K700E Independently Predicts Overall Survival in Myelodysplastic Syndromes." Cancer 127 (19): 3552–65.

Katrukha, I. A. 2013. "Human Cardiac Troponin Complex. Structure and Functions." Biochemistry  78 (13): 1447–65.

Krämer, A. 1996. "The Structure and Function of Proteins Involved in Mammalian Pre-mRNA Splicing." Annual Review of Biochemistry 65: 367–409.

Kühnisch, Jirko, Christopher Herbst, Nadya Al-Wakeel-Marquard, Josephine Dartsch, Manuel Holtgrewe, Anwar Baban, Giulia Mearini, et al. 2019. "Targeted Panel Sequencing in Pediatric Primary Cardiomyopathy Supports a Critical Role of TNNI3." Clinical Genetics 96 (6): 549–59.

Lieu, Yen K., Zhaoqi Liu, Abdullah M. Ali, Xin Wei, Alex Penson, Jian Zhang, Xiuli An, et al. 2022. "SF3B1 Mutant-Induced Missplicing of MAP3K7 Causes Anemia in Myelodysplastic Syndromes." Proceedings of the National Academy of Sciences of the United States of America 119 (1). https://doi.org/10.1073/pnas.2111703119.

Lopes, Inês, Gulam Altab, Priyanka Raina, and João Pedro de Magalhães. 2021. "Gene Size Matters: An Analysis of Gene Length in the Human Genome." Frontiers in Genetics 12 (February): 559998.

Lu, Qian, Bo Pan, Haobo Bai, Weian Zhao, Lingjuan Liu, Gu Li, Ruimin Liu, et al. 2022. "Intranuclear Cardiac Troponin I Plays a Functional Role in Regulating Atp2a2 Expression in Cardiomyocytes." Genes & Diseases 9 (6): 1689–1700.

Malcovati, Luca, Kristen Stevenson, Elli Papaemmanuil, Donna Neuberg, Rafael Bejar, Jacqueline Boultwood, David T. Bowen, et al. 2020. "SF3B1-Mutant MDS as a Distinct Disease Subtype: A Proposal from the International Working Group for the Prognosis of MDS." Blood 136 (2): 157–70.

Mogensen, Jens, Toru Kubo, Mauricio Duque, William Uribe, Anthony Shaw, Ross Murphy, Juan R. Gimeno, Perry Elliott, and William J. McKenna. 2003. "Idiopathic Restrictive Cardiomyopathy Is Part of the Clinical Expression of Cardiac Troponin I Mutations." The Journal of Clinical Investigation.

Obeng, Esther A., Ryan J. Chappell, Michael Seiler, Michelle C. Chen, Dean R. Campagna, Paul J. Schmidt, Rebekka K. Schneider, et al. 2016. "Physiologic Expression of Sf3b1K700E Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation." Cancer Cell 30 (3): 404–17.

Oliva, Esther Natalie, Carlo Finelli, Valeria Santini, Antonella Poloni, Vincenzo Liso, Daniela Cilloni, Stefana Impera, et al. 2012. “Quality of Life and Physicians’ Perception in Myelodysplastic Syndromes.” American Journal of Blood Research 2 (2): 136–47.

Sahota, Virender Kumar, Benjamin Filip Grau, Alicia Mansilla, and Alberto Ferrús. 2009. “Troponin I and Tropomyosin Regulate Chromosomal Stability and Cell Polarity.” Journal of Cell Science 122 (Pt 15): 2623–31.

Sekeres, Mikkael A., Jaroslaw P. Maciejewski, Alan F. List, David P. Steensma, Andrew Artz, Arlene S. Swern, Paul Scribner, John Huber, and Richard Stone. 2011. “Perceptions of Disease State, Treatment Outcomes, and Prognosis among Patients with Myelodysplastic Syndromes: Results from an Internet-Based Survey.” The Oncologist 16 (6): 904–11.

Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” Nature Biotechnology 31 (11): 1009–14.

Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. 2013. “Assessment of Transcript Reconstruction Methods for RNA-Seq.” Nature Methods 10 (12): 1177–84.

Tam, Annie S., Shuhe Tsai, Emily Yun-Chia Chang, Veena Mathew, Alynn Shanks, T. Roderick Docking, Arun Kumar, Delphine G. Bernard, Aly Karsan, and Peter C. Stirling. n.d. “DYNLL1 Mis-Splicing Is Associated with Replicative Genome Instability in SF3B1 Mutant Cells.” https://doi.org/10.1101/2021.05.26.445839.

Tang, Alison D., Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. 2020. “Full-Length Transcript Characterization of SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns.” Nature Communications 11 (1): 1438.

Visconte, V., N. Avishai, R. Mahfouz, A. Tabarroki, J. Cowen, R. Sharghi-Moshtaghin, M. Hitomi, et al. 2015. “Distinct Iron Architecture in SF3B1-Mutant Myelodysplastic Syndrome Patients Is Linked to an SLC25A37 Splice Variant with a Retained Intron.” Leukemia 29 (1): 188–95.

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. 2018. “Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA.” Proceedings of the National Academy of Sciences 115 (39): 9726–31.

Wang, Lili, Angela N. Brooks, Jean Fan, Youzhong Wan, Rutendo Gambe, Shuqiang Li, Sarah Hergert, et al. 2016. “Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia.” Cancer Cell 30 (5): 750–63.

Yien, Yvette Y., Raymond F. Robledo, Iman J. Schultz, Naoko Takahashi-Makise, Babette Gwynn, Daniel E. Bauer, Abhishek Dass, et al. 2014. “TMEM14C Is Required for Erythroid Mitochondrial Heme Metabolism.” The Journal of Clinical Investigation 124 (10): 4294–4304.

Zhao, Weian, Lu Yang, Xiaoxiang Chen, and Wenqi Huang. 2021. “Cardiac-Specific Gene TNNI3 as a Potential Oncogene for Kidney Cancer and Its Involvement in Wnt Signaling Pathway.” Research Square. Research Square. https://doi.org/10.21203/rs.3.rs-754351/v1.