Winter 12-15-2018

The splice is not right: splice-site-creating mutations in cancer genomes

Reyka Glencora Jayasinghe
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Li Ding, Chair
Sergej Djuranovic
John Edwards
Christopher Maher
James Skeath
Matthew Walter

The Splice is Not Right: Splice-site-creating Mutations in Cancer Genomes

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2018
St. Louis, Missouri

Table of Contents

List of Figures iv

List of Tables vi

Acknowledgments vii

Abstract of the Dissertation xii

Chapter 1: Introduction 1

1.1 Cancer Cell Fitness 1

1.2 Alternative Splicing and Disease 6

1.3 References 17

Chapter 2: Complex patterns of splice site mutations 26

2.1 Introduction 26

2.2 Results 30

SPLICEINATOR – AUTOMATIC PRIORITIZATION OF SPLICE MUTATION EVENTS 30

DIVERSE SPLICING PHENOTYPES ASSOCIATED WITH SPLICE SITE MUTATIONS 31

SPLICE SITE MUTATIONS LINKED TO COMPLEX SPLICING EVENTS 33

LACK OF MEASURABLE SPLICING DEFECT FOR MANY SPLICE SITE MUTATIONS 38

IMPACT OF SPLICE SITE MUTATION ON GENE EXPRESSION 41

NON-CANONICAL SPLICE SITE MUTATION DISCOVERY 46

2.3 Methods 47

3.4 Discussion 54

3.5 References 56

Chapter 3: Discovery of splice-site-creating mutations 63

3.1 Introduction 63

3.2 Results 65

SPLICE-SITE-CREATING MUTATION DISCOVERY 65

MOLECULAR AND BIOLOGICAL PATTERNS OF SPLICE SITE-CREATING MUTATIONS 68

EXPRESSIVITY AND PENETRANCE OF SPLICE-SITE-CREATING MUTATIONS 72

SPLICE SITE-CREATING MUTATIONS ACROSS GENES AND CANCER TYPES 74

NEO-ANTIGENS INTRODUCED BY SPLICE SITE-CREATING MUTATIONS 80

3.3 Methods 84

3.4 Discussion 98

3.5 References 100

Chapter 4: Rare germline and somatic splice-site-creating mutations disrupt exon definition in cancer genomes *109*

4.1 Introduction 109

4.2 Results 112

DATASET DESCRIPTION 112

GERMLINE AND SOMATIC SPLICE-SITE-CREATING MUTATION DISCOVERY 114

COMPARATIVE SEQUENCE CONTEXTS AND CHARACTERISTICS 115

TRANSLATIONAL AND TRANSCRIPTIONAL IMPLICATIONS OF SCMS 120

LANDSCAPE OF SOMATIC AND GERMLINE SCMS GENOME-WIDE 124

HYPERMUTATOR PHENOTYPE OF RAD54L RGSCM 126

4.4 Methods 129

4.5 Discussion 138

4.6 References 143

Chapter 5: Conclusion and Future Directions *148*

CANCER SPECIFIC EXON DEFINITION 148

5.1 References 168

Appendix: Additional Projects *173*

5.1 Complex Insertions and Deletions in Cancer Genomes 173

5.2 Analysis of Somatic Complex Insertions and Deletions in Pediatric Hematopoietic Malignancies 174

5.3 Children’s Discovery Institute: Utilizing CharGer to determine novel variants in undiagnosed pediatric cases 206

5.4 Discovery of Novel Fusions in Cancer Genomes 208

5.5 Evaluating Novel Germline Variants in Cancer Genomes 209

5.6 Effect of Blood Somatic Mutations in PPM1D on TP53 and Cell Cycle Arrest 210

5.7 Discovery of Truncation Mutations Leading to Protein Alterations 216

5.8 Viral integration 233

5.9 DROSHA Mutation Analysis 234

List of Figures

Figure 2.1: …… 29
Figure 2.2 …… 33
Figure 2.3: …… 36
Figure 2.4 …… 39
Figure 2.5: …… 41
Figure 2.6 …… 43
Figure 3.1: …… 66
Figure 3.2 …… 69
Figure 3.3: …… 73
Figure 3.4 …… 75
Figure S1 …… 77
Figure 3.5: …… 78
Figure 3.6: …… 82
Figure 3.7 …… 83
Figure S2: …… 89
Figure S3: …… 94
Figure S4: …… 95
Figure S5: …… 96
Figure S6: …… 97
Figure 4.1: …… 113
Figure 4.2: …… 117
Figure 4.3: …… 122
Figure 4.4: …… 125
Figure 4.5: …… 127
Figure S7: …… 140
Figure S8: …… 141
Figure S9: …… 142
Figure A.1: …… 210
Figure A.2: …… 212
Figure A.3: …… 221
Figure A.4: …… 224
Figure A.5: …… 225
Figure A.6: …… 227
Figure A.7: …… 228
Figure A.8: …… 230
Figure A.9: …… 234
Figure A.10: …… 235
Figure A.11: …… 235

Figure A.12: ………………………………………………...…..………………………...…236

List of Tables

Table 3.1 ……………………………….. ………………..…………………………89
Table 3.2 ……………………………………………………..……………………….....90
Table 5.1 ………………………………………………………………………..……...…152
Table 5.2 …………………………………………………………….…………...…158
Table 5.3 ……………………………………………………………….…………...…164
Table A.1 ………………………………………………………………………...…175
Table A.2 ………………………………………………………………………...…212
Table A.3 ………………………………………………………………………...…225
Table A.4 ………………………………………………………………………...…236

Acknowledgments

Li Ding, thank you for being incredibly supportive and fun to work with. You taught me how to be a strong leader and supported all of my endeavors, even when you didn’t necessarily agree. You helped me fall in love with research and I hope to continue pushing the boundaries of this field with your continued support, even after I leave.

My thesis committee has been a great source of support and inspiration throughout my time in graduate school. Although we did not meet as often as I would have liked, our meetings were always scientifically stimulating and immensely helpful in developing my project.

Ding Lab members, past and present including: Steven Foltz, Michael Wendl, Song Cao, Matthew Wyczalkowski, Sohini Sengupta, Yize Li, Adam Scott, Clara Oh, Yanyan Zhao, Alla Karpova, Qingsong Gao, Preet Lal, R. Jay Mashl, Sunantha Sethuraman, Matthew Bailey, Dan Cui, Kuan-Lin Huang, Wen-Wei Liang, Ruiyang Liu, Liang-Bo Wang, Yige Wu, Chris Yoon, Terrence Tsou, Wen-Wei Liao, Venkata Yellapantula, Kai Ye, Jie Ning, Beifang Niu, Jiayin Wang, Mingchao Xie, Fernanda Martins Rodrigues, Yige Wu, Lijun Yao, Dawn King, Mo Huang, Charles Lu, Amila Weerasinghe, Erik Storrs and Hua Sun. It has been an incredible experience working with you all each day and I will cherish all the ridiculous memories we have made together. Sukas for life.

Thanks to collaborators including Feng Chen, Kimberly Johnson, Eduardo Eyras, Lihua Yu, Matthew Walter and Michael McLellan. None of these projects could have been completed without your guidance and expertise.

Thanks to my past mentors including Javier Arsuaga and Morris Head. Javier, thanks for believing in me and inviting me to join your lab after doing a small research project in your class, it changed the whole trajectory of my career and I am so happy I got to start exploring research and science in your lab. Morris, thank you for teaching me to be a leader. Your confidence in me and continued support at the campus academic resource program helped me tackle leadership tasks in graduate school that I couldn't have done without the experience gained through the opportunities you provided to me.

All the awesome people in DBBS have enriched and supplemented my graduate school experience immensely.

Jim Skeath, I have never met anyone who was so motivated and supportive of people who weren't their own blood. But over the last 5 years I've have grown to appreciate and emulate the person you are and I thank you so much for your support of the WashU community. I wouldn't have made it this far if it wasn't for your confidence in me as a scientist, and I thank you from the bottom of my heart for everything that you have done and for making St. Louis feel like another home.

Cherilynn Shadding, thank you for always keeping me in mind when you are developing new curriculum or applying to new grants. I have learned a lot from the opportunities you have provided to me via teaching or being directly involved with the community.

Rochelle Smith, John Russell, Melanie Relich and Patrice Foy. I have enjoyed spending a ridiculous amount of time in the DBBS offices getting to know each one of you. I always felt comfortable because of the type of work environment you all promoted, and your continued support of students will always be appreciated.

Oncology administrative staff including Becky, Denise, Nicole, Helen and Bill. I have had so much fun chatting and drinking with you all over the years. I'm bummed we didn't meet earlier but so glad we could share the last few years together. Future Dr. Abby, I thank you *so* much for your continued distractions and helping me out in the lab when I was lost.

Thanks to the Young Scientist Program crew. Boahemaa Adu-Oppong and Laura Blanton thank you for making me a confident leader and being incredible inspirations.

Thanks to friends in St. Louis and San Francisco that kept me sane including Bhavna Murali, Michael Ross, Renee Sears, Miriam Foltz, Ian Schillbeeckx, Luis Sandoval, Omar Delannoy-Bruno, Natalie Omattage, Janaki Guruge and Michelle Beltran.

Thanks to Jason Solnit and the rest of the Solnit’s, Greenfield’s and Wilbee’s for always opening up your homes to me and your continued encouragement throughout undergrad and graduate school.

Thanks to my dearest family Esrom Jayasinghe, Glen Jayasinghe, Jeremy Jayasinghe, Sheromie Vittachi, Sonali Vittachi, and Erica Vittachi. The last 6 years have been difficult for all of us for many reasons, but knowing we all had each other made it all the more bearable. I love you all very deeply and thanks for supporting me in chasing my dreams .

Thanks to SLRJ: Sohini Sengupta, Laura Arthur, and Jeanette Gehrig. I am so happy to have been on this journey with you all. I think we had way too much fun in graduate school and we have all grown into some pretty dope Dr’s. I look forward to the next memories we make together but will always look back to the fun times we had together in graduate school.

Reyka Jayasinghe

Washington University

December 2018

Dedicated to my mother and grandfather, for always pulling me back to science even when you weren’t around.

Abstract of the Dissertation

The Splice is Not Right: Splice-site-creating Mutations in Cancer Genomes

by

Reyka Jayasinghe

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2018

Professor Li Ding, Chair

Accurate interpretation of cancer mutations in individual tumors is a prerequisite for precision medicine. Large-scale sequencing studies, such as The Cancer Genome Atlas (TCGA) project, have worked to address the functional consequences of genomic mutations, with the larger goal of determining the underlying mechanisms of cancer initiation and progression. Many studies have focused on characterizing non-synonymous somatic mutations that alter amino acid sequence, as well as splice disrupting mutations at splice donors and acceptors. Current annotation methods typically classify mutations as disruptors of splicing if they fall on the consensus intronic dinucleotide splice donor, GT, the splice acceptor, AG. Splice site mutations as a group have been presumed to be invariably deleterious because of their disruption of the conserved sequences that are used to identify exon-intron boundaries. While this classification method has been useful, increasing evidence suggests that splice site mutations can lead to transcriptional

changes beyond disruption and that many exonic mutations that act primarily through alternative splicing are still being overlooked in cancer genomics. My thesis work focuses on developing tools to systematically classify and functionally validate splice site and splice creating mutations using RNA-Seq data, to more accurately understand the functional consequences of mutations on alternative splicing by integrating DNA and RNA-Sequencing data.

First we developed SpliceInator, a semi-automated tool to systematically detect splicing phenotypes using mutation and gene expression data. We interrogated 1,146 conserved splice site mutations across 19 cancer types revealing a wide range of complex splicing phenotypes and emphasize the importance of analyzing patient specific RNA-Sequencing. We further explored beyond the splice site by interogating all mutations in a splicing context using MiSplice for the first large-scale discovery of splice-creating mutations (SCMs) across 8,656 TCGA tumors. We reported 1,964 originally mis-annotated mutations having clear evidence of creating novel splice junctions. Mutations in a subset of genes including *PARP1, BRCA1,* and *BAP1*, were experimentally validated for splice-creating function using a mini-gene splicing assay. Notably, we found neoantigens induced by SCMs are likely several folds more immunogenic compared to missense mutations, exemplified by the recurrent GATA3 SCM. Our work highlights importance of integrating DNA and RNA data for understanding functional and clinical implications of mutations in human diseases. Finally, to further capture the full landscape of SCMs, we explored both somatic and germline mutations for splice-site-creating function using MiSplice. Altogether, we have gathered a set of 2,888 SCMs enabling us to effectively compare the landscape of rare and germline SCMs. This compendium of SCMs has also started to elucidate novel genomic properties of mutations located at the donor and acceptor splice site and SCM containing exons including an overall decrease in the size of the novel exon post mutation, mimicking a natural evolutionary selective pressure but exploited in the cancer genome to maintain proper alternative splicing. To date, this is the first analysis comparing rare germline SCMs and somatic SCMs revealing their comparable dysregulation to the splicing code in cancer. Together my thesis work revealed that splice-site-creating mutants play a much larger role than previously appreciated in contributing to cancer and further expands our understanding of the genetic basis by which mutations can alter the mRNA landscape by dysregulating alternative splicing. More broadly, my work calls for a deeper analysis of seemingly "silent" mutations in any disease as such mutations may alter gene function via alternative splicing and integrating RNA and DNA-Seq can allow for accurate evaluation of mutations in a splicing context.

Chapter 1: Introduction

1.1 Cancer Cell Fitness

Introduction: Cancer, the second leading cause of death in the US, will cause an estimated 609,640 Americans deaths in 2018. The scientific community has come to the conclusion that cancer is a “disease of the genome” but deciphering the origins of cancer is still an evolving field of research. Over time, cells acquire random mutations, many of which have no effect, some with deleterious effects, and a few that confer a selective advantage that allows the cell to grow faster than neighboring cells (Stratton et al., 2009). After a cell acquires a set of genomic alterations that allow it to grow freely, resist cell death, escape normal signaling, and acquire the properties of immortality, metastasis and angiogenesis, the cancer cell can form a malignant mass that is harmful to the host (Hanahan and Weinberg, 2011).

Type of Mutations: Mutations can be germline or somatic in nature. Germline mutations are inherited from parents or acquired *de novo*. Somatic mutations are acquired throughout an organism’s lifetime in individual cells due to genetic and environmental factors, such as chemicals and radiation. Most of the damage in the DNA is repaired, but sometimes the alterations are fixed. Acquired mutations include point mutations (single nucleotide variants / SNVs), insertions, deletions, larger copy number aberrations (CNA), or large scale structural variation (SV). Mutations in genic and regulatory regions can affect gene function in several ways by causing loss or gain of function, altering transcript splicing, and increasing or decreasing gene expression level. Epigenetic changes on histones and DNA, (such as methylation etc.) have been shown to play a central role in turning genes on and off, and are known mechanisms of dysregulation in cancer (Chen et al., 2014; Maunakea et al., 2013; Rajagopal et al., 2014). In combination, such variations can disrupt normal gene function and alter cellular response to regulation giving the cell a selective advantage to proliferate autonomously.

Passenger and Driver Mutations: Genomic changes contributing to cancer can be placed into two categories: driver and passenger events. Driver mutations confer a growth advantage to the cell (under selective pressure) and are positively selected for during the development of the tumor. Novel bioinformatic analyses of large scale sequencing data have helped to infer pathogenic/driver and passenger mutations. Lu *et al.* analyzed allelic imbalance to identify potential initiating germline mutations in cancer (Lu et al.). In a study to identify rare germline truncation variants in 12 cancer types from The Cancer Genome Atlas (TCGA), variants with a higher variant allele fraction (VAF) in the tumor compared to the normal tissue were hypothesized to have undergone positive selection in the tumor. More recently, our lab expanded on this analysis by expanding to a larger cohort of 10,389 samples and expanded the landscape of known pathogenic drivers in a subset of cancer types (Huang et al.). Passenger mutations are those that are acquired in somatic cells but do not increase or inhibit cellular growth potential. These mutations can occur prior to, during, and after tumor initiating mutations but most are present prior to the accumulation of driver mutations (Stratton et al.). Since they are present prior to the driver event, when the driver mutation does confer a selective advantage to a cancer clone, you may observe passenger mutations “hitch-hicking” with the driving clone. Alternatively, initiation mutations are events that confer an advantage to the disease, but require cooperating mutations for developing disease phenotypes. Xie *et al.* identified initiation mutations in *DNMT3A*, *ASXL1*, *TET2* and other genes associated with myelodysplasic syndrome (MDS), myeloproliferative neoplasm (MPN), chronic lymphocytic leukemia (CLL) and acute myeloid leukemia (AML) (Xie et al., 2014). These mutations were identified in patients without overt hematological malignancies suggesting that they may initiate clonal expansion but alone are insufficient to result in the development of cancer. Since this publication, there have been several novel studies that have expanded on this analysis by evaluating the presence of initiation mutations in the general population, but at extremely low variant allele fraction. This finding has appropriately initiated an entire cancer genomics field and spurred the development of novel sequencing technologies to accurately classify low level variants relative to background noise (Wong et al.; Young et al.).

Tumor Suppressors and Oncogenes: Cancer drivers can be categorized as tumor suppressors or oncogenes. In 1979, Oppermann and colleagues linked the neoplastic transformation of cells by avian sarcoma virus to a phosphoprotein with kinase activity encoded by *src*(Oppermann et al.). This report identified the first oncogene, a gene where an alteration to one allele could contribute to the malignant tumor phenotype and cellular transformation. In 1984, *RB1* was identified as the first tumor suppressor gene (TSG), a gene with suppressor or regulatory function(Murphree and Benedict). TSGs commonly require mutation or inhibition of both alleles in order for cellular transformation to occur. According to Knudson's "two hit hypothesis", patients carrying a germline mutation in a tumor suppressor are more likely to develop cancer since a second random somatic mutation affecting the wildtype allele in any cell is much more likely to occur than two random mutations in a non carrier(Knudson). The "two hit hypothesis" is a tale as old as time and is still commonly referred to in the literature as a method to identify putative cancer predisposition genes in large genomic datasets(Park et al.).

Large Scale Cancer Projects: With the first cancer genome sequenced in 2008, the amount of sequencing data in cancer genomics has increased dramatically, supplying the scientific community with a gold mine of data to process, analyze, and refine. Several large scale projects have taken the initiative to integrate genomic data and genome wide annotation tracks allowing multiple teams to work together and pour their efforts into determining key genes and pathways that contribute to cancer. The Cancer Genome Atlas (TCGA) is a project that collected expression, methylation, RNA and DNA sequence data for matched tumor and normal samples across 33 distinct cancer types. The International Cancer Genomics Consortium (ICGC) seeks to provide genomic, transcriptomic, and epigenomic data across 50 different tumor types. The Pediatric Cancer Genome Project (PCGP) is a collaboration between St. Jude Children's Research Hospital and Washington University School of Medicine in St. Louis whose goal is to determine genetic changes that give rise to childhood cancers. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) has expanded off the analyses and projects of the cancer genome atlas to expand the integration of omics data beyond the transcriptome and to the proteome and phosphoproteome. These projects are just a few of the many large-scale projects that are aimed at identifying the genetic origins that contribute to the development and progression of various forms of cancer.

Pan-Cancer Studies: With the accumulation of large-scale genomic data across tissues types due to the aforementioned large scale-cancer projects, some labs have dedicated their entire research focus on pan-cancer analyses. The goal of these pan-cancer analyses is to glean important and relevant signatures and patterns among many patients to determine and define cancer by molecular subtype in order to expand and properly design personal therapies across cancer types (Cancer Genome Atlas Research et al.; Cooper et al.; Hoadley et al.; Huether et al.; Kandoth et al.; Sveen et al.; Vogelstein et al.). Currently, small sample sizes for individual cancer types limits our statistical power to identify more infrequently mutated driver genes. To address this problem, TCGA lead the effort in emphasizing the importance (statistically and biologically) to compare samples across multiple cancer types to identify infrequently mutated sites that are potentially driving cancer development. Initial findings from pan-cancer studies included: the identification of 127 significantly mutated genes across 12 cancer types(Kandoth et al.), which has more recently been expanded to a list of 299 cancer related genes (Bailey et al.). The identification of rare germline pathogenic variants including an analysis across 12 cancer types and subsequently 33 cancer types(Huang et al.; Lu et al.). While the TCGA phase has wrapped up, the upcoming phases, which will expand into the clinic to guide personalized medicine, is going to be a very exciting new direction for the field and will bring novel genomic applications directly to the clinic. While these aforementioned pan-cancer analyses have elucidated many differentially mutated regions in the genome, it is important to remember all sequencing projects are limited by the mutational load of the tumor of interest.

1.2 Alternative Splicing and Disease

The potential for phenotypic variability within an individual species is encoded and dictated by the underlying genomic. Understanding how differences in gene architecture can alter a gene's function is essential to understanding many biological questions including disease diagnosis, progression, and treatment. To understand genomic variability, we will briefly discuss the genomic differences between species, evaluate how the genome has evolved in multicellular organisms, and discuss how mutations can lead to disease.

The Splicing Code: One of the characteristic differences between unicellular and multicellular eukaryotic organisms is the difference in genomes. The central dogma of biology postulates that DNA is transcribed to RNA which is then translated to proteins that perform most functions within the cell. This central dogma lead scientists to realize that certain segments of DNA were not transcribed into RNA or translated in the final protein structure. Instead, transcribed regions of the DNA (exons) were often interrupted by longer segments of DNA that were systematically removed from the RNA (introns) to generate the final sequence that is shuttled for protein translation. In 1978, Walter Gilbert hypothesized different combinations of exons could generate multiple mRNA isoforms from the same pre-mRNA molecule (Berget et al.; Chow et al.; Gilbert; Modrek and Lee, 2002). This hypothesis has come to be known as alternative splicing, the process of joining exons or coding sequences, and splicing out introns.

There is still active debate regarding where in our evolutionary history introns arose. The introns early (IE) hypothesis suggests that introns are ancient structures but were lost in intron-poor species such as prokaryotes (Irimia and Roy; Roy and Gilbert). In opposition, the introns late (IL) hypothesis postulates certain species like eurkaryotes gained introns due to insertion events after the divergence of prokaryotes and eukaryotes. Complicating this question more, intron-rich and intron-poor species are interspersed throughout the eukaryotic branch, suggesting massive intron gain or loss could lead to the diversity of genomic sequences present in eukaryotes.

The average human transcript contains approximately 8 introns. Current studies suggest that between 70-95% of human genes harbor multiple mRNA transcripts(Johnson et al., 2003; Matlin et al., 2005; Modrek and Lee, 2002). Alternative splicing expands the complexity and information content of the eukaryotic genome and allows for tissue, developmental, or temporally expressed isoforms which can perform alternate functions.

The splicing code is made up of cis-acting elements that help the splicing complex distinguish between non-coding and coding regions and facilitates the joining of exons and exclusion of introns (Wang and Cooper, 2007). The consensus intronic dinucleotide GT splice donor and AG splice acceptor flank spliced exons at the 3’ and 5’ ends respectively, and can be found in 99% of all introns. Intronic and exonic splicing enhancer (ISE and ESE) and suppressor (ESS and ISS) elements influence the splicing complex to target true splice sites and pseudo splice sites (Supek et al., 2014b) (Ast; Bonomi et al.; Faustino and Cooper; Keren et al.; Matlin et al.). The spliceosome is responsible for joining exons and splicing out introns to form the mature mRNA molecule. The spliceosome is made of five small nuclear ribonucleoproteins (snRNPs), U1, U2, U4, U5 and U6, along with a number of other proteins. The U1 snRNP interacts with the 5’ splice site (5’ ss), splicing factor 1 (SF1) identifies the branch site, and the U2 auxiliary factory (U2AF) binds the polypyrimidine tract and the 3’ splice site (3’ ss). Additional proteins that facilitate the splicing process include SR proteins (SR) which bind to ESEs and interact with U2AF. Mutations in splices sites and enhancer or suppressor sequences can ultimately contribute to a pathogenic phenotype by altering the binding affinity for spliceosomal or associated proteins.

The splicing process is dynamic and occurs co-transcriptionally. For this reason, genomic context alone cannot begin to describe the efficiency of the splicing code. But we still need to start somewhere.

RNA polymerase II transcribes DNA to create the nascent mRNA molecule. As the molecule is transcribed, the spliceosomal machinery is recruited to facilitate the removal of introns and the joining together of exons. RNA polymerase transcriptional dynamics can be influenced by underlying genetic architecture. For example nucleosomes are commonly found positioned on exons which have a high GC content. The median length of exons are coincidently very close in size to the length of genomic DNA that wraps around histones, suggesting an added selective pressure on maintaining an exon size of rougly 140 bp. Nucleosomes are made up of 8 core histone proteins which contain histone tails that have post-translational modifications (PTMs). PTMs can influence both RNA polymerase transcriptional dynamics and RNA processing by attracting necessary splicing factors to facilitate proper splicing. For example the H2A.Z histone 2A variant is known to promote splicing of non-consensus introns in S. cerevisiae (Herzel et al.; Neves et al.).

Splicing alterations and cancer development: Splice alterations have been shown to affect the landscape of mRNA isoforms present in both tumor and normal tissues (Wang and Cooper). Mutations in cis can directly affect the use of a splice site or promote the use of an alternative splice site facilitating exon/intron inclusion. The production of inappropriate transcripts in a particular tissue type can result in disease due to its alternate function. Mutations in canonical splice sites have been linked to several diseases, for instance: A familial study found in hSNF5, when the consensus GT dinucleotide is mutated to an AT*,* a deletion of exon 7 occurs. This germline mutation was observed in both affected and unaffected family members, but infant brain tumors in affected members showed a loss of the wild-type allele in the tumor(Taylor et al.; Venables). In colorectal and liver metastases, a 3’ splice site mutation from AG to AT leads to a loss of the adjacent exon. Furthermore a loss of the wild type allele in the tumor suggests that this isoform is under positive selection in the tumor (Kurahashi et al.). In another familial study, a single intronic base change formed a cryptic splice site resulting in the addition of 11 nucleotides to the tumor suppressor BRCA1 gene creating a truncated protein(Findlay et al.; Hoffman et al.). Less studied are mutations in less conserved intronic regions. In *ATM*, a mutation at the 6th position of an intronic region is linked to breast cancer but only causes a proportion of transcripts to form a truncated protein(Broeks et al.).

All of the above studies are one-off cases where a resulting phenotypic effect was observed in an individual which led to the discovery of the mutation inducing the aberrant splicing pattern. Bioinformatic heavy labs are starting to group similar splicing alterations to evaluate genomic signatures that can predict the resulting phenotypic changes. Below we will describe some of the recent novel findings in the field that have started to broadly characterize the normal and mutation induced splicing landscape across populations.

The global effects of mutations on splicing variation is now being widely studied in large-scale RNA-Seq consortia. Rivas et al. predicted the effects of protein-truncating variants on the transcriptome by utilizing data from the Genotype-Tissue Expression (GTEx) and Geuvadis projects (Rivas et al., 2015). Both projects allowed the team to quantitatively evaluate allele specific expression of 13,182 mutations in healthy individuals, primarily focusing on nonsense and splicing mutations. Their findings identified variants within the splicing region with significant splice-disruption events and confirmed their findings in a separate dataset of common variants in Swedish individuals.

Using the same dataset, a recent publication evaluated the level of “deleterious-ness” of mutations near the splice-site to determine a more refined definition for the splice-region (Zhang et al., 2018). In addition to the canonical GT and AG sites, the study found one base in the exon and four bases into the intron after the canonical GT (donor) and the first exonic base adjacent to the AG (acceptor) site are mutation intolerant. Additionally, co-occuring mutations were less likely to occur suggesting one mutation in one of the aforementioned sites was enough to perturb the splicing pattern. Additional studies identified synonymous driver mutations in key oncogenes, and found that nearly half of all synonymous driver mutations affected splicing, specifically last base exonic mutations (Supek et al., 2014a). Together, these findings suggest that synonymous driver mutations likely also contribute to cancer through altering alternative splicing.

Shiraishi et al. developed SAVNet to evaluate somatic variants that induced splice altering variants across 31 cancer types in the TCGA cohort (Shiraishi et al., 2018). The authors found that the smoking signature (C>A substitutions) contributed largely to the variants characterizing splicing differences followed by APOBEC.

Cummings et al. used paired transcriptome analyses for patients with undiagnosed muscle disorders and identified recurrent splice-site-creating variants in collagen VI-related dystrophy and predicted functional variants in mutation-rich genes such as *TTN* (Cummings et al., 2017).

Several studies have tried to integrate computational predictions and massively parallel splicing assays to gain new insight into the splicing code. Soemedi et al. evaluated 4,964 single nucleotide variants via a massively parallel splicing assay (MaPSy) (Soemedi et al., 2017). Approximately 10% of variants were confirmed *in vitro* and *in vivo* to disrupt splicing but their assay is limited in only being able to evaluate mutations in exons that were less than 100 nucleotides in length.

Mutations in Trans Elements & Cancer: Mutations in trans-acting elements or splicing factors will affect a larger subset of genes by disrupting the biological machinery that helps to recognize exons and introns. The most well studied splicing factors include serine/arginine-rich (SR) proteins and heterogenous nuclear ribonuclear proteins (hnRNPs)(Kaida et al.; Matlin et al.). SR proteins bind to ESE's and ISE's and promote splicing and inclusion of exons by recruiting spliceosomal proteins. The family of hnRNP proteins function as splicing silencers by binding to ESS's and ISS's. SR proteins in particular are shown to be differentially regulated in cancer and play a role in neoplasia(Fregoso et al.; Kaida et al.). These splicing elements have been shown to function in a position dependent fashion(Lim et al.). SR proteins are usually bound in exons while hnRNPs are found in introns, but modifications that lead to altering the splicing elements binding sites has been linked to their opposite function. Cis mutations that affect binding of hnRNPs and SR proteins can drastically change whether exon usage is promoted or suppressed. An initial study showed that many disease alleles should be classified as splicing mutations even though they were mis-classified as missense mutations because of the motifs that were changed near the splice site(Lim et al.). Using the pan-cancer dataset by TCGA, more motif alterations can be elucidated and linked to splicing alterations.

Mutations in enhancer and silencer sequences can also affect isoform expression because of the proteins that associate with them. Variants in *BRCA1* (Walker et al.), *SMN1/2*, *PDHA* and *GH* are known to cause exon skipping by disrupting the ESE sequence(Woolfe et al.). This same study showed that splice altering variants (SAVs) are enriched in regions near the splice junctions. Exon skipping SAVs are characterized by a loss of ESE and a gain of ESS while variants with increased exon inclusion are characterized by ESS loss. Alternative isoforms are linked to a number of different diseases which have been shown to contribute to drug resistance, tumor angiogenesis, metastasis, and misregulation of apoptosis(Woolfe et al.). Tools are still needed to identify the functional significance of somatic and germline splice altering mutations within the coding and noncoding regions of a gene.

While the field has made great strides in classifying variants that disrupt alternative splicing, interpreting all variants in a splicing specific context has still been largely ignored in large scale sequencing projects. Variant classification and interpretation is fundamental to understanding the biological consequences of mutations on a gene of interest. If variants are not properly annotated, the biological interpretation is incorrect. Currently, mutations are classified based on their changes to the resulting protein. The degenerate nature of the genetic code whereby 61 codons represent 20 amino acids allows nucleotide changes to affect the sequence of mRNA while leaving the amino acid sequence of the resulting protein unchanged. “Silent” mutations that do not change the amino acid are still to this day commonly overlooked in cancer genomics. A study analyzing mutations in the *CFTR* gene showed that synonymous mutations could alter splicing by leading to the use of a cryptic splice site or by altering the 3’ splice site of exon 9 which results in a truncated protein product(Niksic et al.; Wang and Cooper; Wilschanski et al.) Another recent study showed that synonymous mutations were enriched in oncogenes and some tumor suppressor genes in a cancer type specific manner (Supek et al., 2014a). Furthermore, this study found many recurrent “synonymous” *TP53* mutations are inactivating events that alter canonical splicing of the mRNA. These studies suggest that some seemingly “silent” mutations can affect gene function by modifying splicing, transcription factor binding, or other properties that contribute to mRNA translation(Sauna and Kimchi-Sarfaty, 2011). Using known exonic and intronic mutations that activate cryptic splice sites, Lee et al. developed a machine learning classifier to predict whether a mutation would disrupt alternative splicing (Lee et al., 2017). Interestingly, in analyzing variants in CFTR, 70% of the 47 variants tested, were predicted to be missense variants suggesting many splicing alterations are mis-annotated in well-known disease-related genes.

The previously mentioned studies all emphasize the importance of evaluating synonymous and nonsynonymous mutations across cancer types in order to uncover potential pseudo or non-canonical splice sites that are currently being overlooked in cancer genomics.

My thesis work focused first on generating bioinformatic tools that systematically classify splice site and splice-site-creating mutations by integrating DNA and RNA sequencing and experimentally validating the predicted consequences of the variants using a minigene splicing assay. In Chapter 2, I helped develop SpliceInator a semi-automated tool to systematically detect splicing phenotypes using mutation and gene expression data. SpliceInator, combines two lines of evidence to assess mutant specific aberrant splicing events and their implications, one based on interpretation of RNA-Seq fragment mapping using TopHat and another based on standard statistical hypothesis testing of RSEM expression values. We interrogated 1,146 conserved splice site mutations across 19 cancer types revealing a wide range of complex splicing phenotypes. 521 variants in our dataset conferred a measurable splicing alteration, with 69.29% associated with only one splicing defect, while the remaining were a combination of two to four different splicing events. Another 624 splice site mutations did not confer any measurable splicing defects but 75.6% were classified as having a low variant allele fraction, low exon expressivity or both while 24.8% were undetermined. Furthermore, synonymous and non-synonymous variants genome-wide were evaluated in a splicing context and we discovered 243 mutations that create and strengthen nearby alternative splice sites, respectively, further justifying the demand for a tool that can systematically evaluate all mutations in a splicing context. To meet this demand, in Chapter 3 we developed MiSplice for the first large-scale discovery of splice-creating mutations (SCMs) across 8,656 TCGA tumors. We report 1,964 originally mis-annotated mutations having clear evidence of creating novel splice junctions. *TP53* and *GATA3* have 26 and 18 SCMs, respectively and *ATRX* has 5 from low-grade gliomas. Mutations in 11 genes including *PARP1, BRCA1,* and *BAP1*, were experimentally validated for splice-creating function. Notably, we found neoantigens induced by SCMs are likely several folds more immunogenic compared to missense mutations, exemplified by the recurrent GATA3 SCM. Further, high expression of PD-1 and PD-L1 was observed in tumors with SCMs, suggesting candidates for immune blockade therapy. Our work highlights importance of integrating DNA and RNA data for understanding functional and clinical implications of mutations in human diseases.

Finally, to further capture the full landscape of SCMs, in chapter 4 we evaluated both somatic and germline mutations for splice-site-creating function using MiSplice. Altogether, we have gathered a set of 2,888 SCMs enabling us to effectively compare the landscape of rare and germline SCMs. Interestingly, we found mutations overlapping the splice donor site were sufficient to disrupt the canonical splice site but this phenomenon doesn’t hold true for acceptor splice site mutations in our dataset. Alternatively acceptor SCMs needed to not only strengthen the novel splice site to facilitate the novel site usage, but also disrupt the canonical splice site. This compendium of SCMs has also started to elucidate novel genomic properties of SCM containing exons including an overall decrease in the size of the novel exon post mutation, mimicking a natural evolutionary selective pressure but exploited in the cancer genome to maintain proper alternative splicing. To date, this is the first analysis comparing rare germline SCMs and somatic SCMs revealing their comparable dysregulation to the splicing code in cancer. Evaluating mutation induced events separately from patient specific *de novo* events can provide a focused analysis on the genomic features selecting for SCM+ exons relative to leaky splicing or mutation independent cryptic splice site activation. As tissue type specific datasets continue to increase, developing novel tissue specific signatures will help inform the tissue type specific relevance of mutation induced SCMs. Together my thesis work revealed that splice-site-creating mutants play a much larger role than previously appreciated in contributing to cancer and further expands our understanding of the genetic basis by which mutations can alter the mRNA landscape by dysregulating alternative splicing. More broadly, my work calls for a deeper analysis of seemingly “silent” mutations in any disease as such mutations may alter gene function via alternative splicing and integrating RNA and DNA-Seq can allow for accurate evaluation of mutations in a splicing context.

1.3 References

Uncategorized References

Ast, G. (2004). How did alternative splicing evolve? Nature reviews Genetics *5*, 773-782.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B.*, et al.* (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385 e318.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A *74*, 3171-3175.

Bonomi, S., Gallo, S., Catillo, M., Pignataro, D., Biamonti, G., and Ghigna, C. (2013). Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy. Int J Cell Biol *2013*, 962038.

Broeks, A., Urbanus, J.H.M., de Knijff, P., Devilee, P., Nicke, M., Klöpper, K., Dörk, T., Floore, A.N., and van't Veer, L.J. (2003). IVS10-6T>G, an ancient ATM germline mutation linked with breast cancer. Human mutation *21*, 521-528.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

Chen, W., Lin, H., Feng, P., and Wang, J. (2014). Exon skipping event prediction based on histone modifications. Interdiscip Sci *6*, 241-249.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell *12*, 1-8.

Cooper, L.A., Demicco, E.G., Saltz, J.H., Powell, R.T., Rao, A., and Lazar, A.J. (2018). PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. J Pathol *244*, 512-524.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L.*, et al.* (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med *9*.

Faustino, N.A., and Cooper, T.a. (2003). Pre-mRNA splicing and human disease. Genes & development *17*, 419-437.

Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217-222.

Fregoso, O.I., Das, S., Akerman, M., and Krainer, A.R. (2013). Splicing-factor oncoprotein SRSF1 stabilizes p53 via RPL5 and induces cellular senescence. Mol Cell *50*, 56-66.

Gilbert, W. (1978). Why genes in pieces? Nature *271*, 501.

Hanahan, D., and Weinberg, R.a. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol *18*, 637-650.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V.*, et al.* (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929-944.

Hoffman, J.D., Hallam, S.E., Venne, V.L., Lyon, E., and Ward, K. (1998). Implications of a novel cryptic splice site in the BRCA1 gene. Am J Med Genet *80*, 140-144.

Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N.*, et al.* (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355-370 e314.

Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C.*, et al.* (2014). The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nature communications *5*, 3630-3630.

Irimia, M., and Roy, S.W. (2014). Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol *6*.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Kaida, D., Schneider-Poetsch, T., and Yoshida, M. (2012). Splicing in oncogenesis and tumor suppression. Cancer science *103*, 1611-1616.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature reviews Genetics *11*, 345-355.

Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A *68*, 820-823.

Kurahashi, H., Takami, K., Oue, T., Kusafuka, T., Okada, A., Tawa, A., Okada, S., and Nishisho, I. (1995). Biallelic inactivation of the APC gene in hepatoblastoma. Cancer Res *55*, 5007-5011.

Lee, M., Roos, P., Sharma, N., Atalar, M., Evans, T.A., Pellicore, M.J., Davis, E., Lam, A.N., Stanley, S.E., Khalil, S.E.*, et al.* (2017). Systematic Computational Identification of Variants That Activate Exonic and Intronic Cryptic Splice Sites. Am J Hum Genet *100*, 751-765.

Lim, K., Ferraris, L., Filloux, M.E., Raphael, B.J., and Fairbrother, W.G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. PNAS *108*, 11093-11098.

Lu, C., Xie, M., Wendl, M.C., Wang, J., McLellan, M.D., Leiserson, M.D., Huang, K.L., Wyczalkowski, M.A., Jayasinghe, R., Banerjee, T.*, et al.* (2015). Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun *6*, 10086.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol *6*, 386-398.

Maunakea, A.K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res *23*, 1256-1269.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nature genetics *30*, 13-19.

Murphree, A.L., and Benedict, W.F. (1984). Retinoblastoma: clues to human oncogenesis. Science *223*, 1028-1033.

Neves, L.T., Douglass, S., Spreafico, R., Venkataramanan, S., Kress, T.L., and Johnson, T.L. (2017). The histone variant H2A.Z promotes efficient cotranscriptional splicing in S. cerevisiae. Genes Dev *31*, 702-717.

Niksic, M., Romano, M., Buratti, E., Pagani, F., and Baralle, F.E. (1999). Functional analysis of cis-acting elements regulating the alternative splicing of human CFTR exon 9. Human molecular genetics *8*, 2339-2349.

Oppermann, H., Levinson, A.D., Varmus, H.E., Levintow, L., and Bishop, J.M. (1979). Uninfected vertebrate cells contain a protein that is closely related to the product of the avian sarcoma virus transforming gene (src). Proc Natl Acad Sci U S A *76*, 1804-1808.

Park, S., Supek, F., and Lehner, B. (2018). Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. Nat Commun *9*, 2601.

Rajagopal, N., Ernst, J., Ray, P., Wu, J., Zhang, M., Kellis, M., and Ren, B. (2014). Distinct and predictive histone lysine acetylation patterns at promoters, enhancers, and gene bodies. G3 (Bethesda) *4*, 2051-2063.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., Deluca, D.S., Fromer, M.*, et al.* (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*.

Roy, S.W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet *7*, 211-221.

Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature reviews Genetics *12*, 683-691.

Shiraishi, Y., Kataoka, K., Chiba, K., Okada, A., Kogure, Y., Tanaka, H., Ogawa, S., and Miyano, S. (2018). A comprehensive characterization of cis-acting splicing-associated variants in human cancer. Genome Res.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nature Genetics.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.

Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., and Lehner, B. (2014a). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324-1335.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014b). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324-1335.

Sveen, A., Johannessen, B., Teixeira, M.R., Lothe, R.A., and Skotheim, R.I. (2014). Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. BMC Genomics *15*, 672.

Taylor, M.D., Gokgoz, N., Andrulis, I.L., Mainprize, T.G., Drake, J.M., and Rutka, J.T. (2000). Familial posterior fossa brain tumors of infancy secondary to germline mutation of the hSNF5 gene. Am J Hum Genet *66*, 1403-1406.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Walker, L.C., Whiley, P.J., Couch, F.J., Farrugia, D.J., Healey, S., Eccles, D.M., Lin, F., Butler, S.A., Goff, S.A., Thompson, B.A.*, et al.* (2010). Detection of splicing aberrations caused by BRCA1 and BRCA2 sequence variants encoding missense substitutions: implications for prediction of pathogenicity. Hum Mutat *31*, E1484-1505.

Wang, G.-S., and Cooper, T.a. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviews Genetics *8*, 749-761.

Wilschanski, M., Yahav, Y., Yaacov, Y., Blau, H., Bentur, L., Rivlin, J., Aviram, M., Bdolah-Abram, T., Bebok, Z., Shushi, L.*, et al.* (2003). Gentamicin-induced correction of CFTR function in patients with cystic fibrosis and CFTR stop mutations. N Engl J Med *349*, 1433-1441.

Wong, W.H., Tong, R.S., Young, A.L., and Druley, T.E. (2018). Rare Event Detection Using Error-corrected DNA and RNA Sequencing. J Vis Exp.

Woolfe, A., Mullikin, J.C., and Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. Genome Biol *11*, R20.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, C., McMichael, J.F., Schmidt, H.K., Miller, C.a., Bradley, a.*, et al.* (2014). Age-related cancer mutations associated with clonal hematopoietic expansion. Nature Medicine.

Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun *7*, 12484.

Zhang, S., Samocha, K.E., Rivas, M.A., Karczewski, K.J., Daly, E., Schmandt, B., Neale, B.M., MacArthur, D.G., and Daly, M.J. (2018). Base-specific mutational intolerance near splice sites clarifies the role of nonessential splice nucleotides. Genome Res *28*, 968-974.

Chapter 2: Complex patterns of splice site mutations

Contribution: I scripted and developed the full code for the TopHat Module and partial code to the Expression Level Module and Rule Based Classifier of SpliceInator. I manually reviewed all variants validated by our computational tool, performed downstream analyses and created all Figures other than 2.1A.

2.1 Introduction

Large-scale sequence-based studies are increasingly examining the functional consequences of human genomic variants(Dees et al.; Ding et al.; Hoadley et al.; Huether et al.; Kandoth et al.; Wang and Cooper; Xie et al.), with many focusing on characterizing somatic coding mutations and their amino acid sequence alterations in order to investigate the underlying mechanisms of cancer. Splice mutations are of particular interest because of their dramatic modifications of mature RNA products, including possible retention of large intronic segments and loss of whole exons, coupled with significant downstream consequences for associated surviving proteins. With respect to cancer, splicing alterations can affect both tumor and normal tissues(Wang and Cooper), but determining their functional consequences within and outside of the canonical splice site is still understudied in cancer genomics(Barbaux et al.; Broeks et al.; Holmila et al.; Kurahashi et al.; Steffensen et al.; Venables; Woolfe et al.). By one estimate, around 22% of disease causing alleles are mis-classified as missense mutations because they alter motifs near the canonical site utilized by the splicing machinery. Especially relevant to cancer are synonymous mutations that can be enriched in oncogenes and some tumor suppressor genes in specific cancer types(Supek et al.). Several recurrent mutations in cancer-associated genes, such as *TP53*, were found to be inactivating events, altering canonical mRNA splicing. Silent and missense mutations can also affect gene function by modifying splicing, transcription factor binding, or other aspects of mRNA translation (Sauna and Kimchi-Sarfaty). These studies show that mutations in *cis* can directly affect the use of a splice site, or promote the use of an alternative splice site, thereby facilitating exon skipping or intron inclusion. There is a growing appreciation of the need for reliable splicing analyses to identify alternative biochemical pathways that are causative in contributing to the disease state(Dorman et al.; Lim and Fairbrother; Mort et al.; Rivas et al.; Steffensen et al.). It is likely that many, perhaps most of these pseudo or non-canonical splice sites are currently being overlooked in cancer genomics.

Given the biomedical importance of splice-affecting mutations, a bevy of analysis tools has been developed over the last few decades. Early entrants, including Gene Splicer(Pertea et al.), MaxEntScan(Yeo and Burge), Splice Site Finder(Shapiro and Senapathy), NNSplice(Reese et al.), and Human Splicing Finder(Desmet et al.) sought mainly to identify splice site mutations, but methods have increasingly added new analysis capabilities. For example, tools like Skippy(Woolfe et al.), SpliceTrap(Wu et al.), DEXSeq(Anders et al.) use RNA-Seq data to discern differential exon usage and exon-skipping and SpliceSeq(Ryan et al.) added significant visualization aids. Algorithms are becoming progressively more ambitious, for example SpliceR(Vitting-Seerup et al.) and

SplicingTypesAnno(Sun et al.) attempt to characterize a broader set of events encompassing intron retention and various exon extensions and contractions. Calculations often depend on tally-based heuristics and ad-hoc rules. The overall problem of comprehensively characterizing splice mutations and their downstream implications remains unsolved.

Research increasingly suggests that splice architectures are complex and their implications for cancer are significant(Ast; Bonomi et al.; Faustino and Cooper; Keren et al.; Matlin et al.; Modrek and Lee; Taylor et al.; Venables; Wang and Cooper) . A number of specific mechanisms have already been observed. For example, variants in *BRCA1*, *SMN1/2*, *PDHA* and *GH* cause exon skipping by disrupting an exonic splicing enhancer sequence(Walker et al.; Woolfe et al.). Mutations near pseudo splice sites can also result in activation, for instance the c.190 mutation in BRCA1(Yang et al.), which appears to be under positive selection in the tumor. This mutation is a clear example of a predicted “missense” event that actually functions as a modifier of splicing efficiency in vivo and points to modified alternative splicing as a more important effect in cancer than has been previously appreciated. Importantly, we found that standard annotation tools did not properly identify this variant as splice altering.

Given the importance of splice effects in cancer and the lack of a purely algorithmic solution for finding and characterizing them, we applied a semi-automated approach to this problem based on large batches of case-control RNA-Seq data that are now available. We apply this method here to examine 1,146 candidate mutations over 19 cancer types and describe several novel findings.

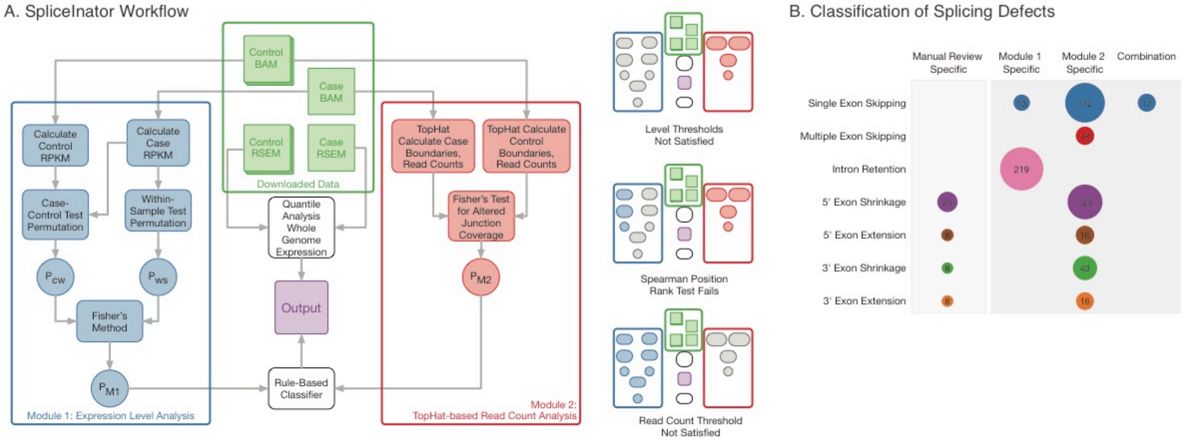

A. SpliceInator Workflow
Control BAM
Case BAM
Control RSEM
Case RSEM
Downloaded Data
Calculate Control RPKM
Calculate Case RPKM
Case-Control Test Permutation
Within-Sample Test Permutation
P_{CW}
P_{WS}
Fisher's Method
P_{M1}
Module 1: Expression Level Analysis
Quantile Analysis Whole Genome Expression
Output
Rule-Based Classifier
TopHat Calculate Case Boundaries, Read Counts
TopHat Calculate Control Boundaries, Read Counts
Fisher's Test for Altered Junction Coverage
P_{M2}
Module 2: TopHat-based Read Count Analysis
Level Thresholds Not Satisfied
Spearman Position Rank Test Fails
Read Count Threshold Not Satisfied
B. Classification of Splicing Defects
Manual Review Specific
Module 1 Specific
Module 2 Specific
Combination
Single Exon Skipping
Multiple Exon Skipping
Intron Retention
5' Exon Shrinkage
5' Exon Extension
3' Exon Shrinkage
3' Exon Extension
13
172
17
18
219
25
123
8
16
8
43
8
16

Figure 2.1: SpliceInator Workflow. The SpliceInator workflow can be grouped into two main modules and post processing analyses. In Module 1, case and control RSEM expression values are compared using standard statistical hypothesis testing to classify intron retention and exon skipping events. In Module 2, BAMs are processed with TopHat using RNA-Seq fragment mapping to identify mutant specific alternative splice junctions to classify intron retention, exon skipping, exon shrinkage and exon extension events. In the post processing steps, a rule based classifier is used to derive support from module 1 and module 2 to classify variants with associated splicing defects. Second, quantile analysis is performed to group genes into expression classifications in a tissue specific manner. After processing, users can compare variants to predicted splicing defects to infer the functional consequences of each variant in a splicing context. The right-hand column highlights several thresholds and internal tests that are utilized in SpliceInator that can alter which steps are run for each module. For example, if read count thresholds are not satisfied in the second module, then only module 1 and the post processing steps will be tested and reported.

2.2 Results

SPLICEINATOR – AUTOMATIC PRIORITIZATION OF SPLICE MUTATION EVENTS

Given the difficulty of rigorously characterizing splice mutation events, we instituted a procedure to automatically prioritize candidates from large lists of inputs (Figure 2.1). Called SpliceInator, it identifies the most promising members, assigning each a preliminary assessment of one or more phenotypes. SpliceInator employs 3 forms of analysis. In the first module, expression levels (units of RPKM) are permutation-tested against corresponding controls, for example intron *x* in the case versus intron *x* in the control group. If 3' bias of the case gene is sufficiently low (Spearman rank test), this between-samples assessment is augmented by a within-sample permutation test, for instance the subject intron against the other introns in the gene. In this latter scenario, the probability results would then be combined using Fisher's chi-square transform method. The second module uses TopHat(Trapnell et al.) assembly to identify boundaries and establish read counts. Given some minimal read thresholds, a Fisher exact test is applied to assess whether patterns of junction coverage are altered between case and controls due to splice site mutation. A third component of SpliceInator classifies whole gene expression values based on Tukey's quartile analysis, including pronouncement of no expression for values more than 1.5 interquartile ranges below the first quartile. Finally, using a rule based classification we then prioritized variants from module 1 and module 2 using the following rules: 1) Events are classified as intron retention if found to be statistically significant by module 1; 2) Exon extension and exon shrinkage events are reported if found to be statistically significant by module 2; 3) Exon skipping events are classified based on a combination approach. If the exon skipping event is found to be statistically significant in module 1 and module 2, the module 1 pvalue is reported and a designation of M1;M2 is reported. If module 2 reports a statistically significant pvalue, an M2 designation is reported. The codebase is available from GitHub (github.com/ding-lab/SpliceInator).

Diverse Splicing Phenotypes Associated with Splice Site Mutations

We collected high quality mutation calls with a UCSC conservation score greater than 99% from 19 cancer types derived from The Cancer Genome Atlas (TCGA). A total of 1,146 splice site mutations currently defined as substitutions, deletions, or insertions overlapped the 2 bp canonical intronic splice donor or acceptor of 624 cancer associated genes. We predicted mutations located at highly conserved sites in cancer genes would be associated with measurable splicing defects. The TopHat module identified alternative junctions near splice site mutations and each prediction was manually reviewed to produce a high confidence list of mutations and associated splicing defects. Our analysis revealed TopHat accurately classifies exon extension, exon shrinkage, and some exon skipping events, but is unable to identify intron retention, another widespread defect observed in many samples by manual review. The statistical based module using RSEM data properly identified intron retention and exon skipping events that were missed in the TopHat alignment analysis.

45.46% of variants in our dataset conferred measurable splicing alterations. The overall landscape of splicing phenotypes for 521 variants included: exon skipping (230), 5’ exon shrinkage (137), intron retention (219), 3’ exon shrinkage (49), multi-exon skipping (32), 3’ exon extension (27), and 5’ exon extension (24) (Figure 2.1a). 69.29% (361) of splice site mutations had only one associated splicing defect, while the remaining were a combination of splicing events including 2 events (24.76%, 129), 3 events (4.99%, 26) and 4 events (0.96%, 5).

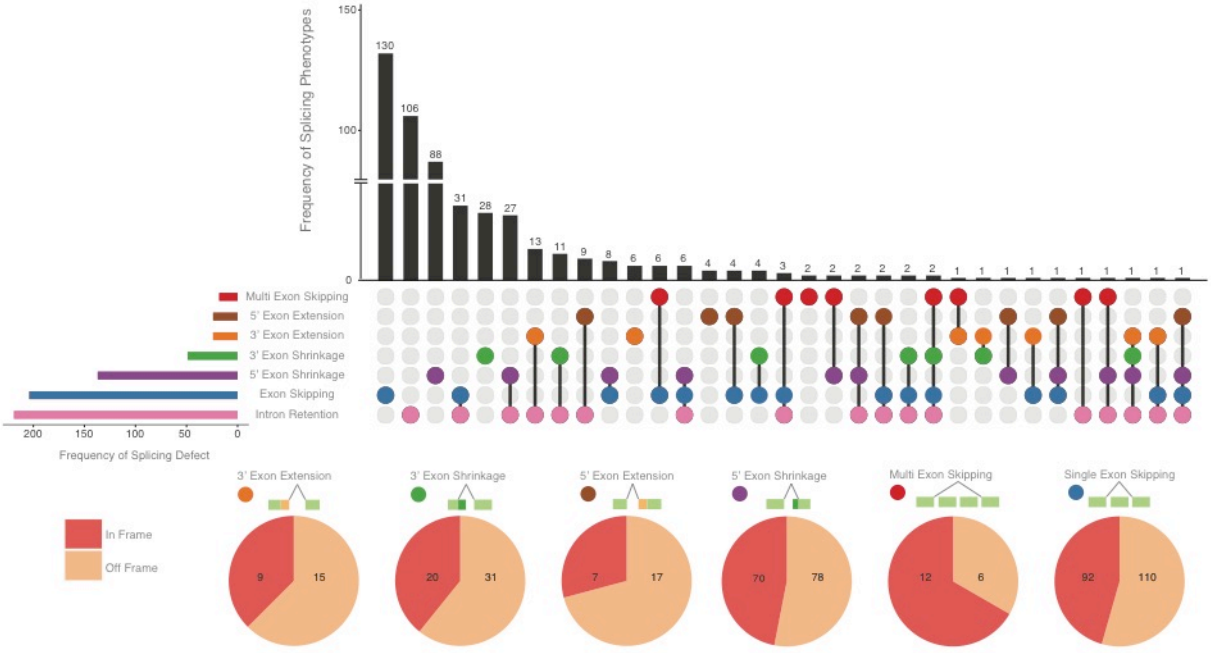


Figure 2.2. Splicing defects associated with splice site mutations. (A) An UpSetR plot highlighting the distribution of splicing phenotypes associated with splice site mutations across the dataset. The bar plot on the left hand side reflects the total amount of sites that fall into each splicing phenotype group. The bar plot at the top depicts the total mutations that are classified as having a corresponding splicing defect depicted in the bubble's below. Colors correspond to a separate splicing phenotype. Black lines between bubbles indicate complex splicing events. (B) For each splicing phenotype, we evaluated how the splicing defect would alter the reading frame of the protein.

SPLICE SITE MUTATIONS LINKED TO COMPLEX SPLICING EVENTS

Current tools and analyses focus on the primary splicing defect, but 30.6% of mutations with a measurable splicing phenotype in our dataset have two to four different splicing alterations linked to the same splice site mutation. The most common complex event involves intron retention and single exon skipping in 29 cases. A 2011 study(Ma et al.) in c.elegans provides evidence that intron retention and exon skipping is regulated by the identity of the 3' splice site due this regions interaction with U2AF splicing factors.

Five mutations in *KDM6A*, *KMT2B*, *CBFB* and *TP53* presented with 3' exon shrinkage and single exon skipping. Two of the mutations sharing this out of frame complex splicing event are annotated to exon 26 of the H3K27 demethylase, *KDM6A*, and have lower expression relative to their BLCA cohort (Figure 2.3a). A closer examination of the RNA-Seq revealed equal support of the 3' exon shrinkage and single exon skipping events, and a lack of reads supporting the canonical junction in both cases. Using MaxEntScan we scored the canonical and alternative splice sites using a sliding window approach to evaluate the strongest alternative splice site score with the introduced mutation. Before the mutations the donor site score is 8.07 but with the introduced mutations the score changes to -0.33 and -6.43 for the point mutation and larger deletion, respectively. Although the alternative splice site utilized in the 3' exon shrinkage event in both cases has a lower splice score of -0.88, Human Splicing Finder scored the alternative site as a potential donor site (70.87 out of 100) taking into consideration additional genomic features such as branch point potential and splice factor protein binding sites(Desmet et al.). Altogether, these results suggest the donor splice site mutation weakens the canonical splice site decreasing the overall expression of KDM6A while also utilizing a nearby alternative donor site leading to a 3' exon shrinkage and exon skipping event. The landscape paper on LUAD(Cancer Genome Atlas Research), and two additional studies(Banka et al.; Cheon et al.) of Korean patients with Kabuki syndrome report the presence of frequent mutations in KDM6A, including mutations overlapping the same splice site, but did not report on the functional consequences of the splice site mutation.

Since the recurrent *KDM6A* splice site mutations had the same splicing phenotype, we next evaluated the landscape of recurrent splicing alterations across the cohort. Our dataset contained 300 recurrent splice site mutations present in two or more samples overlapping the same splice site. Several highly recurrent mutations in the same cancer type showed similar splicing patterns including *GATA3* in BRCA and *MET* in LUAD. The *GATA3* simple indel was identified in 12 breast cancer (BRCA) samples and expressed the same 5' exon shrinkage event across all 13 samples (Arnold et al.; Dorman et al.). Four different mutations overlapping the same splice site in *MET* expressed a similar exon skipping profile that is well supported in the literature(Drilon).

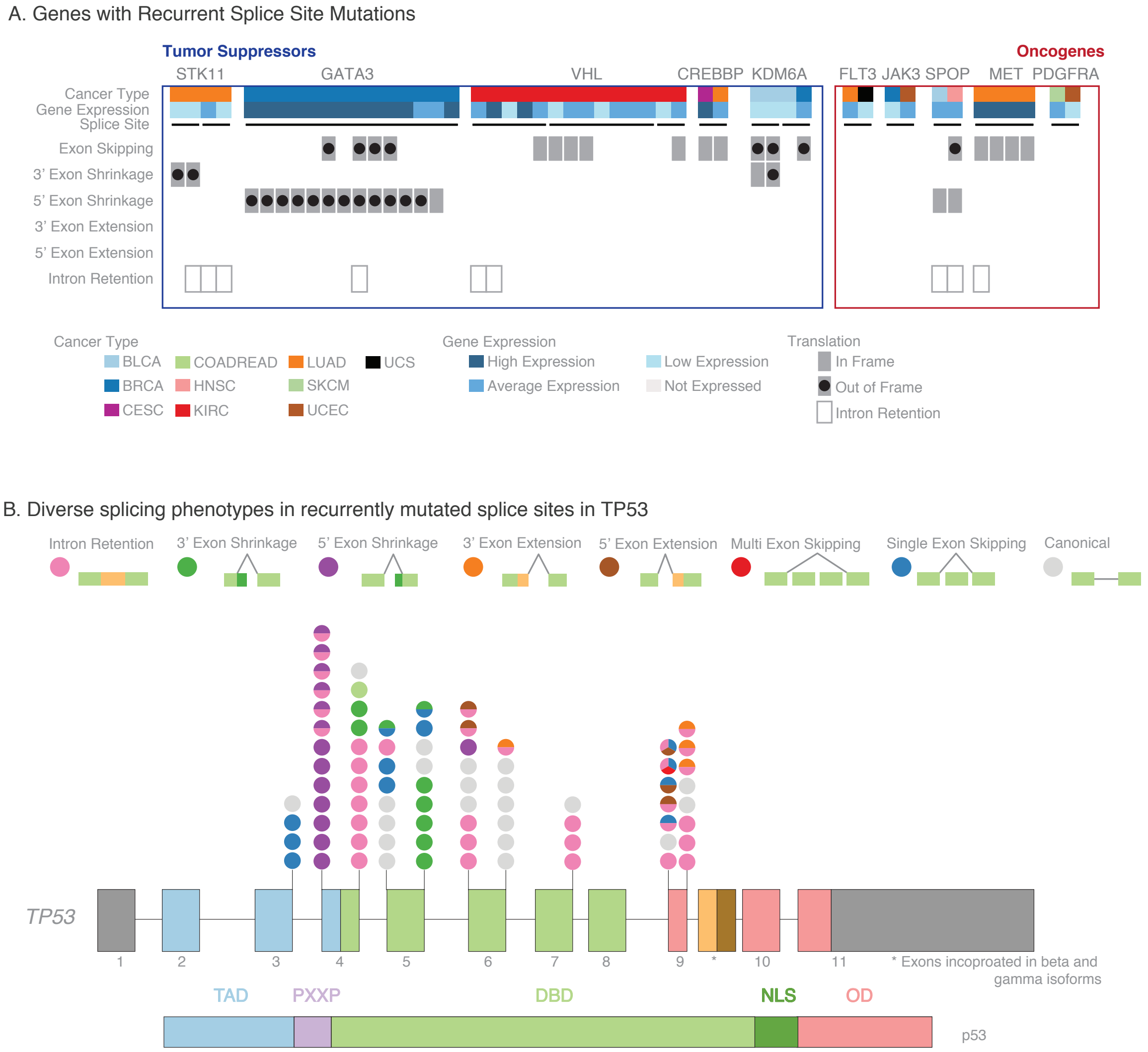


Figure 2.3. Complex Splicing Events. (A) Heatmap showing the distribution of splicing phenotypes associated with select recurrent splice site mutations. Cancer type classification is highlighted above, followed by gene expression classification using quantile based analysis. For select genes, recurrently muatated splice sites in the same gene are grouped by black bands within their gene group. Splicing defects for each site are classified as in frame and out of frame with an empty and filled in square, respectively. (B) Lolliplot with 80 recurrent mutations in TP53 spanning 9 splice sites. Mutations are depicted by a circle, with each circle colored according to the identified splicing defect. If more than one defect is identified to be linked to that mutation, the circle is divided into two or three quadrants, with each quadrant colored according to the splicing defect.

Alternatively, many recurrent sites did not show such consistency across samples and tissue types. *TP53* is highly mutated across cancer types and harbors 80 recurrent splice site mutations across nine splice sites. Evaluating the distribution of splicing phenotypes for each splice site shows a very different effect. Mutations at p.261_splice are derived from BRCA, HNSC, KIRC and UCS and while three cancer types share the same intron retention event, the UCS sample has no measurable splicing alterations. Six of the seven mutations at p.224_splice are derived from four cancer types and all lack a measurable splicing defect, but a single mutation in SKCM expresses 3' exon extension and intron retention which could be explained by the high tumor VAF (94.44%) for this particular mutation. This result highlights the importance of having individualized RNA-Sequencing to complement mutational analysis to evaluate patient specific splicing alterations.

LACK OF MEASURABLE SPLICING DEFECT FOR MANY SPLICE SITE MUTATIONS

We next wanted to explore whether there was a relationship between the presence of an aberrant transcript and the newly formed premature termination codon (PTC). The Nonsense Mediated Decay (NMD) pathway predominantly degrades PTC's and nonsense mutations that differ from the canonical stop codon to reduce expression of potentially damaging transcripts. The general rule of thumb is that PTC's located at least 50-55 bp upstream of the last exon-exon junction drive strong NMD, whereas those out side of this criteria are predicted to escape the degradation process. To evaluate if the presence of a measurable splice defective product is due to the lack of degradation by NMD, we translated the aberrantly spliced product and evaluated the position of the PTC relative to the exon-intron boundary predictions. Our analysis found that X mutations create a PTC that is predicted to escape NMD while X mutations should hypothetically be degraded.

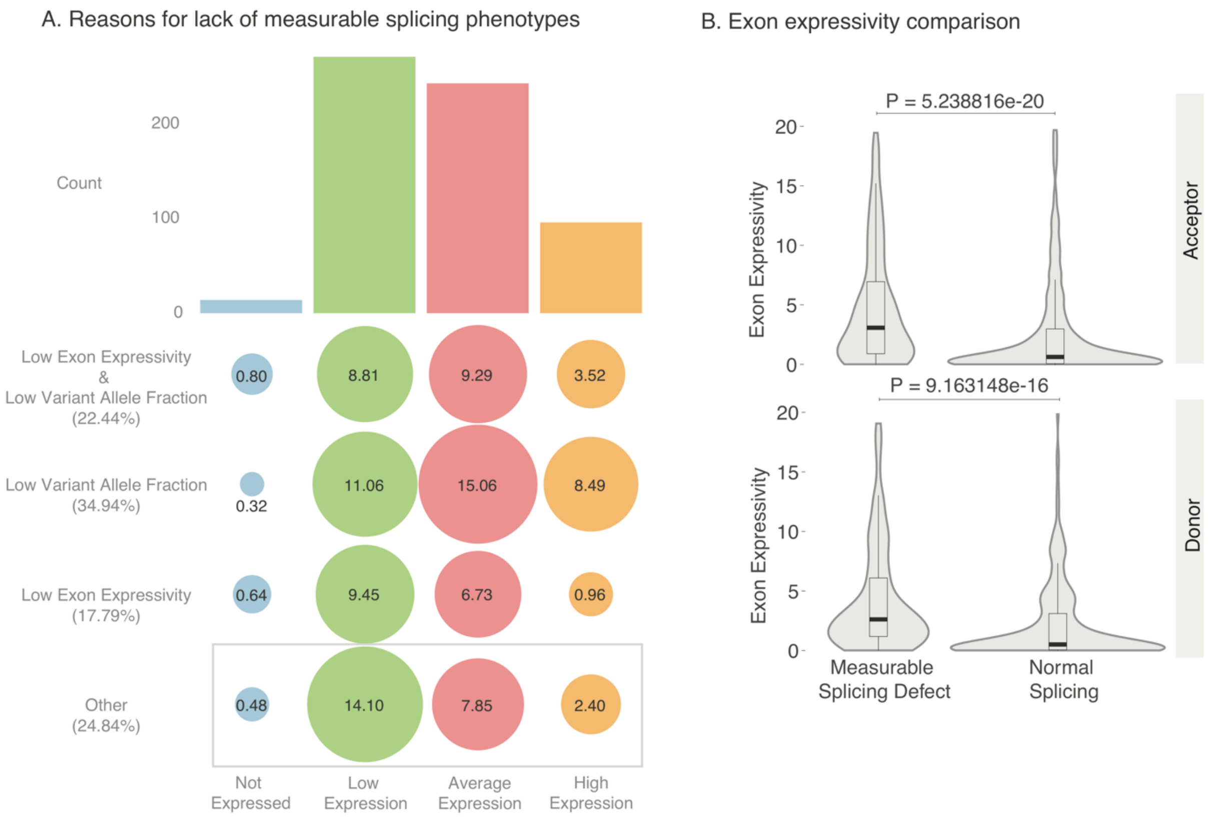


Figure 2.4. Lack of a measurable splicing defect. (A) Mutations which lack a measurable splicing defect are categorized into bins by quantile gene expression classification into not expressed, low expression, average expression and high expression based on their tissue type (x axis). Each variant is then grouped based on exon expression, variant allele fraction and unknown causes to explain the lack of a measurable splicing defect in each gene expression category. (B) Overall distribution of exon expressivity between samples found with and without a measurable splicing defect separated by acceptor and donor site annotation.

Although we identified splicing alterations for 521 variants, 624 splice site mutations were not associated with any measurable splicing defects suggesting one or more of the following: (1) a lack of expression of the transcript; (2) low expression of the variant allele in the tumor; (3) the creation of a highly unstable transcript that is degraded; (4) copy number aberrations; (5) presence of a subclone; or (6) a false positive. To evaluate lack of expression of the case sample relative to the controls, we performed Tukey's quartile analysis, including pronouncement of no expression for values more than 1.5 interquartile ranges below the first quartile (Methods). While 339 sites had high or average expression relative to their cohort, 14 and 272 genes were considered as not expressed or lowly expressed, respectively, suggesting the mutation disrupted overall expression of the transcript. To assess exon expression we evaluated reads supporting the adjacent annotated exon against control samples and normalized by the length of the transcript and total mapped reads (Methods). For 116 of the 624 splice site mutations, the transcript had no expression in the tissue type of interest illuminating the lack of a measurable splicing defect in the presence of a splice site mutation. Furthermore, the group lacking splicing defects had an overall lower exon expressivity and lower variant allele fraction distribution when compared to the group of variants with measurable splicing defects (Figure 2.4). Altogether, our results suggest 75.6% of highly conserved splice site mutations lacking a measurable splicing defect can be classified as having low VAF (<30%), low exon expressivity or both while 24.8% are still undetermined (Figure 2.4).

IMPACT OF SPLICE SITE MUTATION ON GENE EXPRESSION

Gene	Sample	Cancer	VAF %	CN	Purity	Splicing Defects
TP53	TCGA-N5-A4RS	UCS	88.8	2.02	0.6832	🟣
PTEN	TCGA-EB-A44P	SKCM	75	1.26	0.7913	🔵◉ 🩷
TP53	TCGA-22-5482	LUSC	74.07	1.94	0.7887	🟣
KEAP1	TCGA-55-6972	LUAD	72.73	1.38	0.9048	🔴◉ 🔵 🩷
CREBBP	TCGA-FU-A3TX	CESC	63.64	1.97	0.794	🔵
AXIN1	TCGA-CN-A497	HNSC	63.16	2.65	0.7019	🔴◉ 🔵
ACVR2A	TCGA-53-7624	LUAD	62.3	2.45	0.8436	🟤 🩷
KEAP1	TCGA-44-7667	LUAD	60.87	1.78	0.6368	🟣 🟤◉ 🩷
TSC1	TCGA-GV-A3QF	BLCA	60.18	1.94	0.7913	🩷
MGA	TCGA-BT-A20N	BLCA	59.38	1.78	0.7417	🩷
CHD4	TCGA-EA-A4BA	CESC	57.89	1.95	0.9509	🩷
WAS	TCGA-AY-6196	COADREAD	55.26	1.77	0.5077	🩷
CDKN2A	TCGA-EE-A20B	SKCM	53.57	1.83	0.691	🔵◉ 🩷
PTEN	TCGA-B5-A11R	UCEC	53.02	2.47	0.9212	🩷
BCORL1	TCGA-CN-5356	HNSC	52.83	2.01	0.6448	🟠◉ 🩷
BRIP1	TCGA-EE-A2MD	SKCM	52.43	2.62	0.8366	🔵◉ 🟣
MAP4K3	TCGA-GV-A3QK	BLCA	51.52	2.51	0.809	🟣◉ 🩷
KMT2C	TCGA-CZ-4857	KIRC	50.7	3.15	0.61	🔵 🩷
CBLC	TCGA-EB-A5SF	SKCM	50	1.62	0.8304	🔵◉

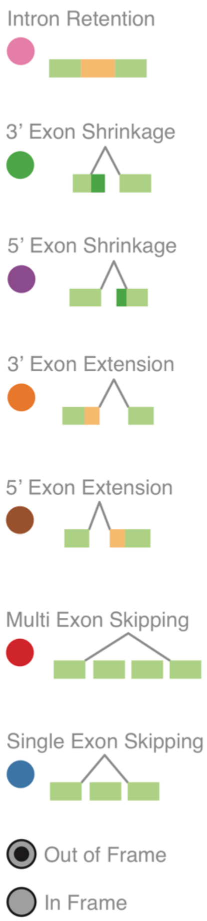


Figure 2.5. Splice altering variants with increased expression. 19 predicted splice eQTLs identified across cohort. Indicated in figure are gene name, sample, cancer, variant allele fraction (VAF), copy number (CN), purity estimate, and the associated splicing defects indicated by color. Color legend on right hand side corresponds to associated splicing defect, and a filled in circle indicates predicted out of frame events.

We next compared cases against cancer specific control cohorts to evaluate the effect of splice site mutations on gene expression. Overall, around 59% of sites had comparable (high or average) expression within their control cohort, while 41% had lower overall expression (low or not expressed). Interestingly of the 59%, 19% had higher gene expression, selecting for an altered transcript that escapes RNA degradation pathways. In the high expression group, we sought to identify splicing expression quantitative trait loci (eQTLs) where the splice site mutation is associated with higher expression in the gene of interest. We hypothesized splice eQTL's would have a measurable splicing defect, elevated DNA VAF (>50%) and increased overall gene expression of case samples relative to the associated control cohort. Out of 526 mutations with measurable splicing alterations, 19 are predicted to be splice eQTLs in our cohort and they reside in the following cancer associated genes: TP53 (2), PTEN (2), KEAP1 (2), CREBBP, AXIN1, ACVR2A, TSC1, MGA, CHD4, WAS, CDKN2A, BCORL1, BRIP1, MAP4K3, KMT2C, and CBLC (Table 1). Interestingly, 10 of these 19 mutations are complex events, having 2 or more splicing phenotypes associated with the mutation, and 13 are classified as having a retained intron.

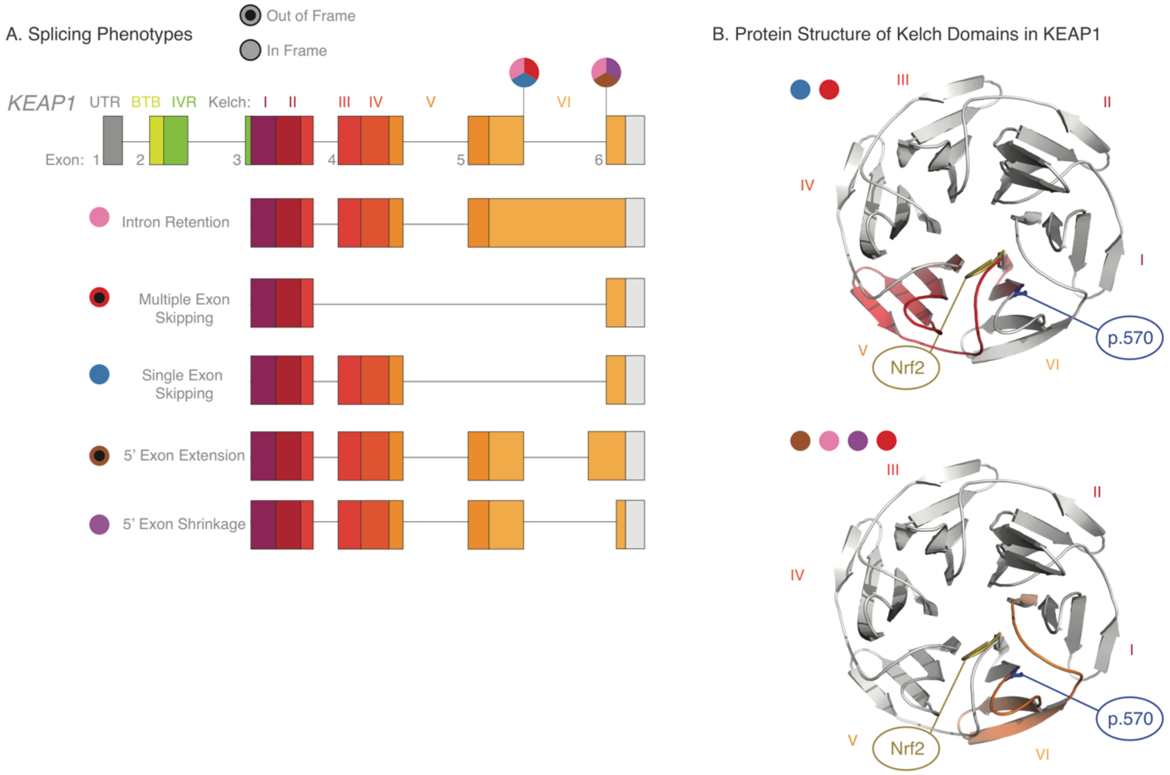


Figure 2.6. KEAP1 highly expressed splice eQTLs. (A) Lolliplot depicting location of highly expressed splice site mutations in KEAP1 (p.570). Each domain is highlighted using a different color and each circle on the lolliplot visualizes the location of the mutation on the exonic sequence and highlights the complex splicing patterns identified at each site. The resulting consequences of each splicing defect on the kelch domains is further annotated below the lolliplot revealing crucial structural elements that are absent under each condition. (B) Protein structure of the kelch domains in the KEAP1 structure. Six kelch domains form a propellar structure within KEAP1 (gray) that is responsible for interacting with Nrf2 (yellow). The circles in the upper left hand corner of each structure highlight the splicing defects that will disrupt the associated colored kelch domain in the 3D depiction. The amino acid position where the splice site mutations fall are highlighted in blue and annotated along with the Nrf2 interacting domain (yellow).

Two highly expressed splice eQTLs were identified in the E3 ubiquitin ligase Kelch-like erythroid cell derived protein with CNC homology (ECH) associated protein 1 otherwise known as KEAP1. KEAP1 sequesters and ubiquitinates Nrf2, a nuclear factor erythroid

2-related factor 2 (NFE2L2). In cancer, many mutants disrupt the binding between KEAP1 and Nrf2. Disrupted binding increases the presence of free Nrf2, thereby activating downstream genes that are conducive for cell growth and responding to oxidative stresses. Three LUAD samples in our dataset harbor splice site mutations in KEAP1, two of which were classified as high expressing splice eQTLs. We hypothesized the two mutations in KEAP1 are disrupting the binding between KEAP1 and Nrf2. One mutation was identified at the acceptor site of exon 6 and another at the donor site of exon 5, with a VAF of 61% and 72%, respectively. Copy number data also suggests the mutant with 72% VAF has undergone a copy number loss, providing further evidence as to the selection of the mutant case in the tumor. Furthermore, while one mutation occurs within the last 50 bp of the last exon-exon junction and the other is present in the last exon of KEAP1, this could explain why the splicing defects are tolerated and escape degradation by nonsense mediated decay.

When comparing the KEAP1 mutants against the control group, KEAP1 mutants had higher overall gene expression (wilcox test=0.04) and lower overall NFE2L2 expression (wilcox test = 0.01). Interestingly, one KEAP1 mutant sample with RPPA data had higher Nrf2 protein expression relative to the controls and lower KEAP1 protein expression, providing an answer as to the mechanism by which the splice eQTL could be disrupting this biological pathway. The discordant relationship between protein and gene expression in this case could be explained by a negative feedback loop leading to a decrease in gene expression under conditions of high protein expression.

Structurally, the mutations occur within the last of the six kelch repeats that fold into a beta propeller tertiary structure, with each blade compromising four anti-parallel beta sheets (Figure 2.6b). In Figure 2.6b we highlight the kelch domains that are predicted to be disrupted due to associated splicing defects. For example, if exon 5 of KEAP1 was skipped, the fifth kelch domain would be completely absent thus disrupting the overall propeller structure and the binding pocket for Nrf2. These results lead us to believe the KEAP1 high expressing splice eQTLs act in a dominant negative manner by disrupting the binding of KEAP1 to Nrf2 thereby increasing the amount of free Nrf2 in the cell.

NON-CANONICAL SPLICE SITE MUTATION DISCOVERY

Finally, we wanted to perform a genome-wide investigation on the creation and strengthening of de novo splice sites by exonic variants. We used two independent methods to interrogate mutations outside of the splice site in a splicing context. Using the entropy based method, for each variant the sequences of each 9-mer (donor) and 23-mer (acceptor) covering the variant position was extracted from the genome for both the mutant and reference sequence. Their splice score as potential donor or acceptor sites were estimated using MaxEntScan. The largest scores of the 9-mer or 23-mer windows were retained for mutant and reference and their difference (mutant - reference) was calculated. Finally, the scores with largest difference between potential donor site and potential acceptor site was retained as the final score. A novel splice site is predicted to be created if the reference score <3 and the mutant score is >8. We further filtered this list by evaluating nearby junctions in the bam files (reads>10) for predicted novel splice sites. For the second method, we used the TopHat module of SpliceInator to collect all alternative junctions within 14 base pairs of the annotated mutations. We manually reviewed all the predictions to evaluate the efficacy of our method.

For the entropy based method, a total of 243 sites were predicted near a novel junction but after manual review the method correctly called 9 de novo donor sites, 6 strengthened alternative acceptor sites, and 7 strengthened alternative donor sites. Of the 22 sites, 15 were specific to the mutated sample and not present in any control samples. For the TopHat method, a total of 192 alternative acceptor sites and 204 alternative donor sites were predicted to fall within 14 bp of a non-splice site mutation. After manual review, 1 *de novo* acceptor site, 35 *de novo* donor sites, 153 strengthened acceptor splice sites, and 145 strengthened donor splice sites were properly detected. Furthermore, 49 of the 154 acceptor sites had no supporting reads supporting the novel junction in the control and 84 of the 180 donor sites were specific to the mutated sample. The MaxEntScan scoring filtered out many sites that were found using our TopHat method, suggesting using the splice score alone isn't sufficient to properly evaluate mutations in a splicing context. Both methods picked up strengthened acceptor sites in the following genes *CTSH*, *LAMC1* and *NUP98;* and 4 created donor sites in *WDR33*, *NOP14*, *RSRC2* and *PARP1*.

The Poly ADP-ribose polymerase 1, PARP1, is an enzyme involved in recruiting proteins involved in DNA repair pathways. Both methods identified a silent mutation in PARP1 in a LUSC patient that acts as a splice altering variant by creating a de novo donor site in exon 21. RNA-Sequencing data supports the use of the *de novo* donor site leading to an 11 amino acid deletion, which falls within the PARP1 catalytic regulatory domain. 180 reads in the mutated sample supported the *de novo* site while 170 controls contained no supporting reads giving us strong evidence that this "silent" mutation is more appropriately a splice altering variant. PARP1 inhibitors are commonly used in cancer treatment to disrupt DNA repair in tumor cells thereby leading to the accumulation of DNA breaks and ultimately cell death(Malyuchenko et al.).

2.3 Methods

<u>SpliceInator Methods:</u>

The analysis procedures use RPKM values, which were derived for the full-length exons and introns of genes using TCGA RNA-Seq bam files. Briefly, bed files were generated for exon and intron coordinates based on the Ensembl 37.75 database. The SamTools package counted raw reads for each exon and intron. Finally, RPKM was calculated as 109·R/(N·L), where R is the number of raw reads mapped to each exon or intron, N is the total number of mapped reads in the project, and L is the length of the exon or intron. Some analyses also consider ±50 bps and ±2 bps around junctions, for which we use RPM (# of mapped reads per million reads) to quantify the expression because the lengths are short and fixed.

The main algorithm, which we call SpliceInator, combines two lines of evidence to assess aberrant splicing events and their implications, one based on standard statistical hypothesis testing of RSEM expression values and the other based on interpretation of RNA-Seq fragment mapping using TopHat. We have found empirically that each of these components excels at a relatively mutually exclusive subset of the various kinds of splicing anomalies that have been observed: intron retention and exon skipping for the statistical method and exon extension and shrinkage for TopHat-based analysis. The latter also can detect some exon skipping events, for which perform ad hoc interpretation in light of the statistical evidence. Consequently, we have not found it necessary to develop any sort of overarching method for combining results of the two calculations.

For the genes of interest, the statistical method takes as input both the gene-wide RSEM and the individual intron and exon RSEM values for each gene over a set of samples. Genes reporting RSEMs below a threshold are considered to have expression too low to properly assess and are skipped. Otherwise, each sample for each gene is processed sequentially for the different kinds of possible splicing anomalies. Each of these calculations proceeds in essentially the same way, the intron retention calculation being representative, as follows. Preliminary exclusion criteria are evaluated first: the intron's RSEM must be above a threshold and must likewise be greater than the average RSEM over the corresponding introns in the control group, of which there must be at least 50 instances. Assuming these criteria are met, the algorithms then evaluates whether the subject intron's RSEM is significantly higher than the average of those in the control group via simple permutation testing, which returns a "case-control" P-value. The method attempts to maximize power when possible. Specifically, if 3' bias within the gene of the current sample is sufficiently low, a within-sample test is also performed. Potential 3' bias is checked via Spearman's rank test between the introns' RSEM values and their location midpoints, both ordered 5' to 3'. Lack of strong bias is inferred if Spearman's coefficient ≤ 0.3, in which case the algorithm runs a second permutation test to check whether the subject intron RSEM is significantly greater than the other intron RSEMs within that gene. This calculation returns the "within-sample" P-value. These tests are essentially independent, by which we finally combine the case-control and within-sample P-values using Fisher's Method. Analysis of exon-skipping is similar. The algorithm is relatively robust against the heuristic parameters quoted above.

For TopHat-based analysis, TopHat Junction output was used for each splice site mutation to identify junction ids where mutations are between the start and stop of the junction feature. Exon intron boundaries were calculated using the overhang value for each junction feature. We enforced a threshold of at least 10 reads aligned to a particular exon-intron boundary. Depths of coverage for each junction for cases and controls were collected using the TopHat junction output. To obtain total unique reads within the junction neighborhood, we performed read counts on the in house bam files and used this value as our total reads for both cases and controls. Reads supporting reference junction versus alternative junction were compared using Fisher's 2x2 exact test between cases and controls to discern altered junction coverage due to presence of the splice site mutation.

Methods Novel splice site mutation discovery:

We propose to use the TopHat Module of SpliceInator to identify all alternative junction predictions across the entire cancer genome atlas RNA-Seq sample set (>7,000 samples) and pediatric cancer dataset (>500 samples). To hone in on mutations that are directly disrupting pseudo splice sites, we will first identify missense and silent variants that are within a threshold distance of an alternative junction prediction. After identifying the alternative junction, we searched for nearby donor and acceptor sites that may have been strengthened or weakened by the introduced mutation. By comparing cases to controls within the same cancer type, we were able to determine if the novel junction is used within the cancer type at a low level or only present in the case with the mutation. Finally, each *de novo* donor and acceptor site is scored using a scoring algorithm. By using entropy and information analysis of the consensus base sequences, we can evaluate the strength of a particular site based on its nucleotide frequencies(Burset et al.). Using a weighted scoring method for each possible acceptor/donor sequence, we multiplied the information of each base by the information weight of the site. The acceptor and donor information and weight is derived from over 25,000 different splice donor/acceptor sites to come up with the information weights and content at each position. To be classified as creating or strengthening a novel acceptor or donor site the mutation fell within 8 bp of the novel predicted donor site, 14 bp of the novel predicted acceptor site, had supporting reads from both the positive and negative strand, and create or strengthen a nearby site that utilizes the canonical GT or AG splice site.

Methods to evaluate exon expression:

Github/Exon-Expressivity

We developed a script which take as inputs a Mutation Annotation File (MAF) and an RNA bamlist to compare total reads supporting an exon against the total mapped reads (normalized by length) to evaluate exon expressivity across cancer types and samples in our dataset.

- Exon_Expressivity.pl : Inputs 1) RNA-Seq bam locations 2) Mutation Annotation File (MAF). This script will identify all exons nearest to the splice site mutation of interest designated in the MAF file and collect exon read count data across samples in the RNA-Seq bam list and total mapped reads from the flagstat or alignment stats files in the designated bam directories. The exon expressivity coefficient (Ei) is calculated by dividing the total reads aligned to exon (E) over the total mapped reads of the imported bam (B), normalized by the length of the exon and multiplied by a scaling factor (see equation below). The Ei coefficients are then averaged across all samples for each cancer type to determine the overall exon expressivity of the exon of interest. If multiple transcripts share the same exon, multiple Ei values will be reported for each individual exon.
- $Exon\ Expressivity\ (Ei) = \frac{\left(\frac{Total\ Reads\ Aligned\ to\ Exon\ (E)}{Total\ Mapped\ Reads\ of\ Imported\ Bam\ (B)}\right) Length\ of\ Exon\ (bp)}{} * 10^9$

To determine a cut off for the exon expressivity (Ei) we evaluated Ei scores across all 1,159 mutations in our dataset and associated control samples. We determined Ei of case samples with detectable splicing alterations to define samples with known "expressed junctions".

Alignment guided junction analysis to identify mutations linked to splicing defects: We propose to identify junctions between two exon-exon boundaries genome wide. A junction is defined by RNA-Seq reads that are split between two exon-intron boundaries, joining two exons. TopHat[41] reports the number of reads supporting the canonical junction shared between two exons, or any deviations including exon skipping, intron retention, 5' exon shrinkage, 3' exon shrinkage, multi exon skipping, 3' exon extension, and 5' exon extension. The number of reads supporting each junction for cases and controls were collected for each sample. Controls were defined for each gene based on tissue type and only include samples lacking mutations in the gene of interest. Reads supporting the reference junction and alternative junctions were compared using Fisher's Exact Test between cases and controls to determine which junctions showed statistically significant altered junction usage. This calculation indicates whether the altered junction is potentially due to the mutation in the case, or if a background level of alternative splicing is occurring at this junction due to a tissue type specific event.

Part 1. SpliceInator: Splice site mutations that lead to aberrant signals have cancer implications. We developed a tool combining TopHat RNA-seq analysis with case-control and within-sample statistical testing of expression data to identify and classify such events.

RNA-Seq Guided Junction Analysis using SpliceInator to identify mutations contributing to splicing defects: SpliceInator is a statistical tool that we are developing to classify variants into the following categories: 1) expression loss, 2) intron retention, 3) exon skipping, 4) exon extension, 5) exon shrinkage, and 6) unknown, using patient specific RNA-seq data. It performs permutation based testing on normalized expression data derived from exons, introns, and splice junctions between mutant and control samples. Additionally, when there is minimal 3' bias, within-sample testing is also performed to add statistical power.

First, reads per kilobase per million (RPKM) are derived for each gene using patient specific RNA-Seq. Next, mutations are tested for intron retention. Within this test, we first check for 3' bias within the given sample according to Spearman's Rank Correlation test of each exon across the gene in a 5' to 3' manner. Absent significant bias, RPKM of each intron is tested against all others within the gene to check for elevated expression. Next, the full intron expression of the mutant sample is tested against the same intron in the control cases. For exon skipping, the same type of calculations are performed, except full exon expression is used. After an initial analysis of a highly recurrent GATA3 acceptor variant, we noted that exon shrinkage and extension events could not be properly identified when comparing the full exon and intron expression between cases and controls. For the exon shrinkage and extension tests, we decided to additionally extract RPKM values associated with +/- 2bp and +/- 50 bp of all defined exon boundaries. This allowed us to test smaller events that could be occurring near a splice site, such as the 7 bp shrinkage identified in the GATA3 variants, or larger events with the +/- 50 bp data.

For exon shrinkage and exon extension, the full exon expression, 2 bp expression and 50 bp expression are all compared between cases and controls. This analysis will allow us to connect our expression analysis with junction-based prediction and identify cases of intron retention, which are outside of TopHat's predictive capability.

Intron retention events are common splicing aberrations: SpliceInator identified 104 intron retention events highlighted in Figure 2.1. An additional 44 exon skipping events were identified by SpliceInator, further supporting the findings of TopHat predicted exon skipping events. Our findings emphasize the importance of using both tools in identifying splicing defects associated with mutations. [add more examples here about interesting intron retention events].

3.4 Discussion

Splice alterations have been shown to affect the landscape of mRNA isoforms present in both tumor and normal tissues. Unlike many missense mutations known to disrupt the protein structure or associated binding sites, an initial review of splice site mutations quickly revealed position and tissue type alone couldn't explain observed splicing patterns. Since looking solely at genomic context could not directly inform whether a splice site mutation would confer a splicing defect we developed SpliceInator, a semi-automated tool to interrogate mutations in a splicing context.

Evaluating splicing defects associated with highly conserved splice site mutations revealed diverse splicing patterns. While some well known splicing patterns were observed, more than 30% of variants with a known splicing defect had a more complex phenotype consisting of two or more splicing defects. Although some recurrent splice site mutations at the same position conferred similar splicing defects, many exhibited different patterns, suggesting tumor heterogeneity makes predicting the effects of mutations on the transcriptome very difficult without patient specific RNA-Sequencing data. In particular, TP53 harbored many recurrent splice site mutations across cancer types and didn’t show consistent splicing patterns while recurrent mutations in KDM6A did.

Even more interesting was the lack of a measurable splicing defect for 54% of highly conserved splice site variants. While we were able to classify approximately 75% of variants from this group into samples having low variant expression and low exon expressivity, approximately 25% were still unclassified. The unclassified variants could be creating a highly unstable transcript that is quickly degraded by the cell, but follow up functional analysis is needed to confirm this prediction revealing the limitation of RNA-Seq data. We were also able to demonstrate mis-annotated variants by evaluating all mutations in a splicing context using SpliceInator. By more accurately classifying variants in genes such as PARP1 and PTEN we can better understand the biological implications of variants disrupting the splicing process and determine patient specific treatments.

Altogether our findings suggest you can’t predict what to expect due to the complexity of the splicing process and tumor heterogeneity. While our tool makes a significant contribution by providing an automated method to determine functionally relevant mutations using matched RNA-Seq data there is still much to learn about how cancer can abrogate splicing mechanisms for tumor growth and proliferation. For example, while we focused on defining specific defects such as exon skipping, exon shrinkage, exon extension and intron retention, there were several events that were not easily classified into one or more of the aforementioned categories. Further developing our tool to accommodate more complex splicing defects will be a necessity to better gauge the true landscape of splicing phenotypes in disease.

3.5 References

Uncategorized References

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res *22*, 2008-2017.

Arnold, J.M., Choong, D.Y., Thompson, E.R., kConFab, Waddell, N., Lindeman, G.J., Visvader, J.E., Campbell, I.G., and Chenevix-Trench, G. (2010). Frequent somatic mutations of GATA3 in non-BRCA1/BRCA2 familial breast tumors, but not in BRCA1-, BRCA2- or sporadic breast tumors. Breast Cancer Res Treat *119*, 491-496.

Ast, G. (2004). How did alternative splicing evolve? Nature reviews Genetics *5*, 773-782.

Banka, S., Lederer, D., Benoit, V., Jenkins, E., Howard, E., Bunstone, S., Kerr, B., McKee, S., Lloyd, I.C., Shears, D.*, et al.* (2015). Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). Clin Genet *87*, 252-258.

Barbaux, S., Niaudet, P., Gubler, M.C., Grunfeld, J.P., Jaubert, F., Kuttenn, F., Fekete, C.N., Souleyreau-Therville, N., Thibaud, E., Fellous, M.*, et al.* (1997). Donor splice-site mutations in WT1 are responsible for Frasier syndrome. Nat Genet *17*, 467-470.

Bonomi, S., Gallo, S., Catillo, M., Pignataro, D., Biamonti, G., and Ghigna, C. (2013). Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy. Int J Cell Biol *2013*, 962038.

Broeks, A., Urbanus, J.H.M., de Knijff, P., Devilee, P., Nicke, M., Klöpper, K., Dörk, T., Floore, A.N., and van't Veer, L.J. (2003). IVS10-6T>G, an ancient ATM germline mutation linked with breast cancer. Human mutation *21*, 521-528.

Burset, M., Seledtsov, I.A., and Solovyev, V.V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res *29*, 255-259.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.

Cheon, C.K., Sohn, Y.B., Ko, J.M., Lee, Y.J., Song, J.S., Moon, J.W., Yang, B.K., Ha, I.S., Bae, E.J., Jin, H.S.*, et al.* (2014). Identification of KMT2D and KDM6A mutations by exome sequencing in Korean patients with Kabuki syndrome. J Hum Genet *59*, 321-325.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R.*, et al.* (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res *22*, 1589-1598.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., and Beroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res *37*, e67.

Ding, L., Wendl, M.C., McMichael, J.F., and Raphael, B.J. (2014). Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet *15*, 556-570.

Dorman, S.N., Viner, C., and Rogan, P.K. (2014). Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. Sci Rep *4*, 7063.

Drilon, A. (2016). MET Exon 14 Alterations in Lung Cancer: Exon Skipping Extends Half-Life. Clin Cancer Res *22*, 2832-2834.

Faustino, N.A., and Cooper, T.a. (2003). Pre-mRNA splicing and human disease. Genes & development *17*, 419-437.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V.*, et al.* (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929-944.

Holmila, R., Fouquet, C., Cadranel, J., Zalcman, G., and Soussi, T. (2003). Splice mutations in the p53 gene: case report and review of the literature. Human mutation *21*, 101-102.

Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C.*, et al.* (2014). The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nature communications *5*, 3630-3630.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature reviews Genetics *11*, 345-355.

Kurahashi, H., Takami, K., Oue, T., Kusafuka, T., Okada, A., Tawa, A., Okada, S., and Nishisho, I. (1995). Biallelic inactivation of the APC gene in hepatoblastoma. Cancer Res *55*, 5007-5011.

Lim, K.H., and Fairbrother, W.G. (2012). Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing. Bioinformatics *28*, 1031-1032.

Ma, L., Tan, Z., Teng, Y., Hoersch, S., and Horvitz, H.R. (2011). In vivo effects on intron retention and exon skipping by the U2AF large subunit and SF1/BBP in the nematode Caenorhabditis elegans. RNA *17*, 2201-2211.

Malyuchenko, N.V., Kotova, E.Y., Kulaeva, O.I., Kirpichnikov, M.P., and Studitskiy, V.M. (2015). PARP1 Inhibitors: antitumor drug design. Acta Naturae *7*, 27-37.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol *6*, 386-398.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nature genetics *30*, 13-19.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., and Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome biology *15*, R19-R19.

Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res *29*, 1185-1190.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J Comput Biol *4*, 311-323.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M*., et al.* (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*, 666-669.

Ryan, M.C., Cleland, J., Kim, R., Wong, W.C., and Weinstein, J.N. (2012). SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics *28*, 2385-2387.

Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature reviews Genetics *12*, 683-691.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res *15*, 7155-7174.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T.V. (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. European journal of human genetics : EJHG *3*, 1-7.

Sun, X., Zuo, F., Ru, Y., Guo, J., Yan, X., and Sablok, G. (2015). SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data. Comput Methods Programs Biomed *119*, 53-62.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324-1335.

Taylor, M.D., Gokgoz, N., Andrulis, I.L., Mainprize, T.G., Drake, J.M., and Rutka, J.T. (2000). Familial posterior fossa brain tumors of infancy secondary to germline mutation of the hSNF5 gene. Am J Hum Genet *66*, 1403-1406.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Vitting-Seerup, K., Porse, B.T., Sandelin, A., and Waage, J. (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics *15*, 81.

Walker, L.C., Whiley, P.J., Couch, F.J., Farrugia, D.J., Healey, S., Eccles, D.M., Lin, F., Butler, S.A., Goff, S.A., Thompson, B.A.*, et al.* (2010). Detection of splicing aberrations caused by BRCA1 and BRCA2 sequence variants encoding missense substitutions: implications for prediction of pathogenicity. Hum Mutat *31*, E1484-1505.

Wang, G.-S., and Cooper, T.a. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviews Genetics *8*, 749-761.

Woolfe, A., Mullikin, J.C., and Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. Genome Biol *11*, R20.

Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R., and Zhang, M.Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics *27*, 3010-3016.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A.*, et al.* (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med *20*, 1472-1478.

Yang, Y., Swaminathan, S., Martin, B.K., and Sharan, S.K. (2003). Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. Human molecular genetics *12*, 2121-2131.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

Chapter 3: Discovery of splice-site-creating mutations

Contribution: I developed 1/3 the code for MiSplice with the other parts developed by Qingsong Gao and Song Cao. I manually reviewed all variants validated by our computational tool, performed downstream analyses, created all Figures and performed entire mini-gene splicing assay for 11 variants tested.

3.1 Introduction

Large-scale sequencing studies, such as The Cancer Genome Atlas (TCGA) project, have worked to address the functional consequences of genomic mutations in tumors (Dees et al., 2012; Kandoth et al., 2013; Lawrence et al., 2013; Niu et al., 2016), with the larger goal of determining the underlying mechanisms of cancer initiation and progression. Many studies have focused on characterizing non-synonymous somatic mutations that alter amino acid sequence, as well as splice disrupting mutations at splice donors and acceptors (Jung et al. 2015). Current annotation methods typically classify mutations as disruptors of splicing if they fall on either the consensus intronic dinucleotide splice donor, GT, or the splice acceptor, AG. As a group, splice site mutations have been presumed to be invariably deleterious because of their disruption of the conserved sequences that are used to identify exon-intron boundaries.

While this classification method has been useful, increasing evidence suggests that splice site mutations can lead to transcriptional changes beyond disruption of the canonical junction (Lim and Fairbrother, 2012; Mort et al., 2014; Rivas et al., 2015; Sauna and Kimchi-Sarfaty, 2011; Steffensen et al., 2014). One such example is the c.190 mutation in *BRCA1*. Conventional annotation had predicted a missense mutation, p.C64G, but our analysis of RNA-seq data in ovarian tumors harboring p.C64G and a published mouse model (Yang et al., 2003) suggested the germline c.190 mutation leads to the creation of an alternative splice junction, resulting in a truncated null protein. There have been several case studies reporting observations of missense and silent mutations activating cryptic splice sites in *MLH1* (Nyström-Lahti et al., 1999), *LMNA* (Woolfe et al., 2010), *RB1* (Zhang et al., 2008), *RNASEH2A* (Rice et al., 2013), *MECP2* (Taimoor I Sheikh, 2013), *BAP1* (Wadt et al., 2012), *KIT* (Chen et al., 2005), as well as other studies relating missense and silent mutations with splicing changes (Jung et al., 2015; Kahles et al., 2016; Soemedi et al., 2017; Supek et al., 2014). Despite the broad clinical ramifications of mutation-induced altered splicing, systematic evaluation of their occurrence and resulting effects in cancer has yet to be undertaken, nor have there been significant bioinformatics platforms for doing so.

We applied a newly developed bioinformatic tool called MiSplice, which integrates DNA and RNA-Seq data across thousands of samples to discover mutations that induce splice site creation. In our large-scale analysis across 8,656 tumor samples, we report 1,964 such somatic mutations that had originally been mis-annotated. Splice site-creating mutations (SCMs) are enriched in the new introns, with the highest rate at the -3 nucleotide position of acceptors with two-thirds of such events at aGag and agGag repeats by creating an alternative junction 2 nucleotides away. Partial and full splice creation capabilities across these 1,964 sites are evaluated by measuring the fraction of reads supporting the alternative junction, which we term the Junction Allele Fraction (JAF) and which is found to be negatively correlated with distance to the new splice site. In total, 1,607 genes harbor SCMs, with 248 of them having more than one mutation, including *TP53, GATA3, ATRX*, and *NF1*. Recurrent SCMs were found in *TP53*, *GATA3*, *DDX5*, *KDM6A*, *PTEN*, *SETD2*, *SMAD4*, *BCOR*, *SPOP,* and *BAP1,* suggesting association with cancer development. Broadly speaking, integrated DNA and RNA data can furnish a sound basis for discovering SCMs and for accurately understanding functional consequences of mutations in cancer and other human diseases.

3.2 Results

Splice-Site-Creating Mutation Discovery

We collected high quality mutation calls from 8,656 tumors across 33 cancer types derived from The Cancer Genome Atlas having available TCGA RNA-Seq data (Methods). For every mutation, we defined a set of control samples in the same cancer cohort that lacked the same mutation in the gene of interest. We sought to assess the landscape of SCMs across cancer genomes by evaluating all mutations already having conventional annotations and their potential splice site-creating effects (Fig. 3.1A). To achieve this goal, we conducted analysis using a bioinformatic tool, MiSplice (Mutation Induced Splicing), which systematically evaluates mutations in a splicing context using RNA-Sequencing data (Fig. 3.1B).

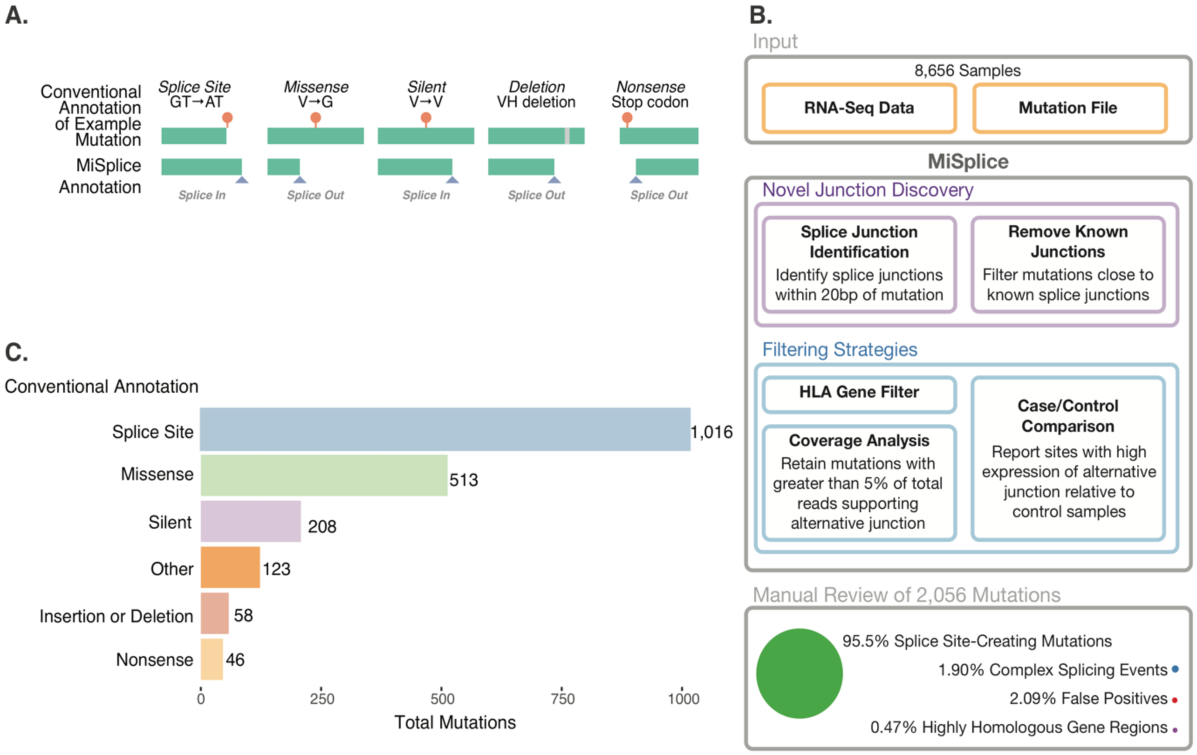


Figure 3.1. Splice site-creating mutation discovery. (A) Examples of splice site-creating mutations for different conventionally annotated mutation types. Splice-in is defined as mutations contained within the newly created exons and splice-out is when the mutation is present in the newly created intron. (B) The MiSplice work flow consists of three steps: alternative junction discovery, filtering, and manual review. First, the user inputs the locations of RNA-Seq BAM files along with a mutation file. MiSplice searches the BAM file to identify any alternative splice junctions near the mutation of interest, while filtering out known splice junctions and calculates the number of alternative junction supporting reads for case and control samples. For the filtering step, the following sites are removed: mutations in HLA genes, low fraction of reads supporting the alternative splice junction, and sites expressed in controls. Finally we manually reviewed all sites to validate the in silico predictions. (C) Breakdown of 2,056 manually validated splice site-creating mutations by conventional annotation.

MiSplice manages large analyses using parallel computation to search for alternative splice junctions within windows of ±20 bp from the mutation of interest. For example, of the 1,416,566 candidate mutations examined here, 4,448 had 5 or more unique RNA-Sequencing reads supporting the predicted alternative junction in proximity to the mutation. MiSplice then conducts a series of further evaluations, including Ensembl-based filtering of canonical junctions, establishing observational significance by case comparison to a matched set of controls, and assessing score and depth of each cryptic site using MaxEntScan (Yeo and Burge, 2004) and SamTools (Li et al., 2009). From the resulting subset, MiSplice filters out HLA genes and sites whose junctions have insufficient difference of expression, as judged from the case-control assessment. Here, we evaluated promising alternative junctions with at least 5% of paired end RNA-Sequencing reads at the genomic location supporting the alternative junction of interest. MiSplice processing revealed 2,056 mutations (Table S1) potentially creating an alternative splice site. Manual review indicated a 2.09% false positive rate, suggesting high specificity of the MiSplice algorithm for discovering these types of mutation-induced splicing events. Of these putative splice events, 1.90% and 0.47% are considered complex and are in highly homologous gene regions, respectively, so they were excluded from further analyses (Methods).

Of the final 1,964 splice site-creating mutations (SCMs) passing manual review (Table S1), 52% (1,016) are in annotated splice sites, suggesting disruption of the canonical splice site and selection of a the alternative splice site nearby (Fig. 3.1C). Importantly, 26% (513) and 11% (208) of the SCMs had previously been mis-annotated as missense and silent mutations, respectively. In addition, we found 58 insertions or deletions, 46 nonsense, and 123 non-coding region mutations that likewise create cryptic splicing sites.

MOLECULAR AND BIOLOGICAL PATTERNS OF SPLICE SITE-CREATING MUTATIONS

We next characterized the sequence context for the 1,790 SCMs corresponding to single nucleotide mutations. The sequences of each 9-mer (donor) and 23-mer (acceptor) covering the mutation position were extracted for both the mutant and reference sequences. Their splice scores as potential donor or acceptor sites were then estimated using MaxEntScan (Table S1).

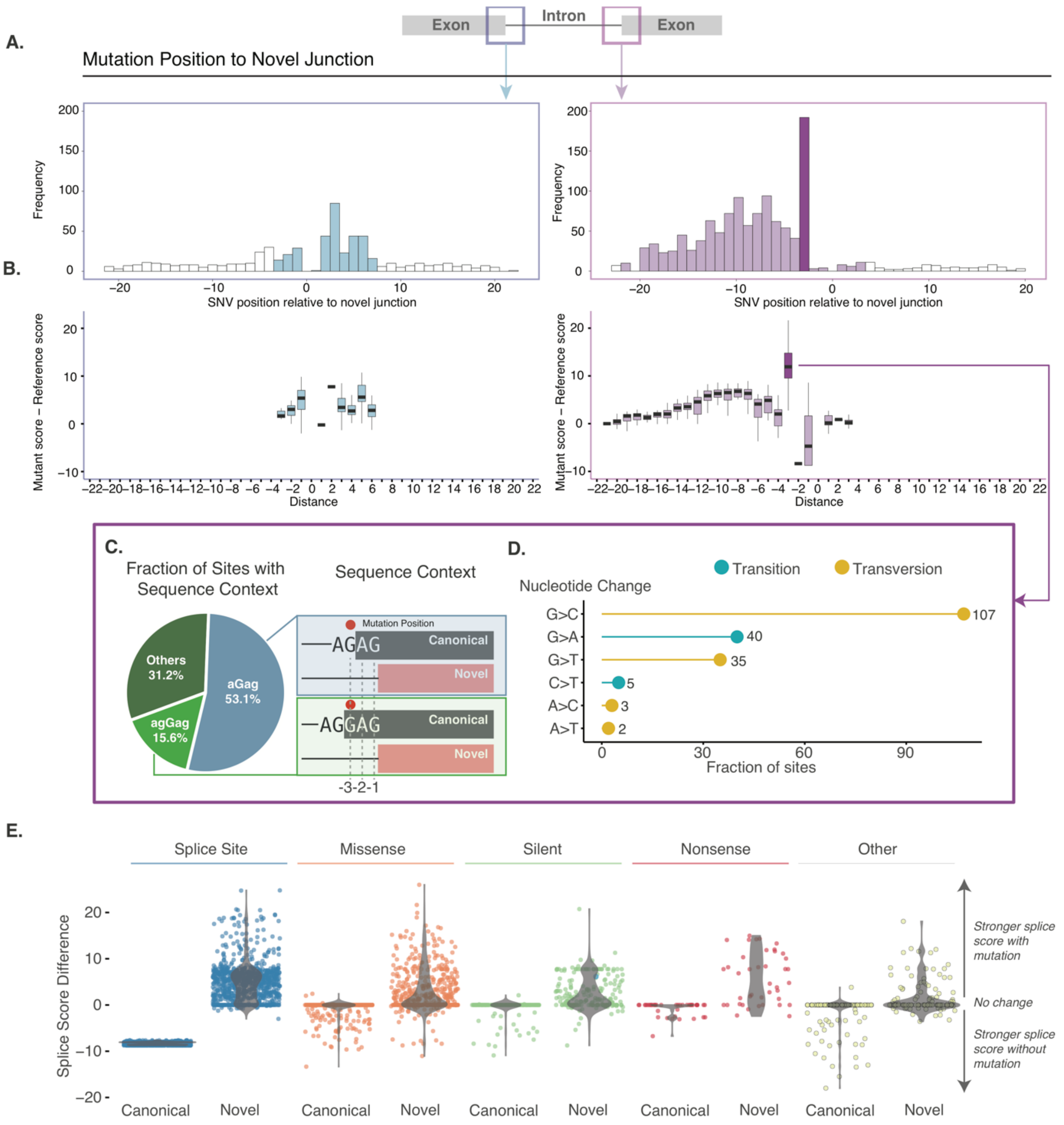


Figure 3.2. Sequence contexts and characteristics of splice site-creating mutations. (A) Frequency distribution of splice site-creating mutations relative to the newly created splice junction, with high frequency shown at the 3rd nucleotide position in the newly created intron. (B) Comparison of splicing scores for the newly created splice site, before (reference) and after the mutation (mutant). Showing larger effect of mutations at the 3rd nucleotide position in the intron (especially for the 3' splice

sites). (C) Dominant nucleotide sequence context for splice site-creating mutations at -3 position of the 3' splice site. Mutation position (red dot) is present 3 base pairs away from the newly created exon. (D) Transition and transversion rate at the -3 position of the 3' splice site. Most mutations are G>C transversions, strengthening the consensus sequence of the splicing factor U2AF1. (E) Comparison of splicing scores between the nearest canonical splice junction with and without mutation compared to the newly created splice junction with and without the mutation. Most mutations strengthen the alternative splice junction relative to the canonical splice junction.

Mutations near the alternative splice junctions show higher mutation rates in the introns for both 5' ($p<1e-5$, binomial test) and 3' splice site ($p<1e-6$) (Fig. 3.2A). More interestingly, we found an enrichment of mutations at the third nucleotide position in the intron, but depletion at the first and second positions (especially for 3' splice site) (Fig. 3.2A). Comparison of splicing scores between splice site-creating mutants and reference forms shows that most mutants have stronger splice signals than the reference (Fig. 3.2B). Mutations that create a G or T to produce an alternative 5' splice site dramatically increase splice site strength. For 3' splice sites, mutations enriched on the third nucleotide of the newly created intron showed the largest increase of splicing score (Fig. 3.2B). Further examination of the sequence context around mutations at the third nucleotide of 3' splice sites shows that 53% have a mutation on aGag repeats and another 16% are mutated on agGag repeats, all creating alternative junctions 2 nucleotides away from the annotated ones (Fig. 3.2C). Mutations at the -3 position of the alternative acceptor site would potentially enhance *U2AF1* recognition of the acceptor splice site. Previous studies have reported S34F *U2AF1* mutants preferentially skip exons that contain a T nucleotide at the -3 position (Okeyo-Owuor et al., 2015). Of the 192 mutations that are located at the

-3 position from the alternative junction and contain an AG in the -2 and -1 positions, 56% undergo a G>C transversion (21% G>A, 18% G>T, 3% C>T, 2% A>C, 1% A>T) with C being the preferred base at the -3 position for *U2AF1* binding (Fig. 3.2D).

We also explored the relationship between the alternative and canonical splice junctions. As expected, mutations at splice sites dramatically reduced splice scores of the canonical splice junctions, while strengthening those at the alternative splice junctions in most cases. In contrast, a subset of missense and silent mutations did not drastically alter the canonical junction, but instead preferentially strengthened a nearby alternative splice site (Fig. 3.2E). When analyzing the raw splicing scores (canonical and alternative site before and after mutation), we found that 1,089 out of 1,790 (61%) events showed higher splice score for the alternative splice site than the canonical site, indicating inclination for the alternative sites. Further, while 485 (27%) events saw lower post-mutation alternative splice score, differences between alternative and canonical scores had decreased, suggesting that these mutations are still likely enhancing the preference for the alternative site. Only 214 (12%) events did not show evidence suggesting increased post-mutational preference for using the alternative site. These cases are a good illustration of the fact that many other genomic splicing features are also relevant, including exonic splicing enhancers (ESE), polypyrimidine tract, branch point, and RNA-binding proteins. They are also consistent with the general view that splice score is not definitive (Jian et al., 2014). We emphasize that all 1,790 alternative splice sites demonstrated usage based on patient RNA-Seq data and that 10 out of 11 (>90%) identified splice site-creating mutations were validated experimentally (see below).

EXPRESSIVITY AND PENETRANCE OF SPLICE-SITE-CREATING MUTATIONS

In the presence of the mutation, alternative splice junctions exhibited a wide range of expression. To quantify this effect, we measured alternative junction expression as the fraction of alternatively spliced junction spanning reads over the total number of reads at the genomic location, what we refer to as the Junction Allele Fraction (JAF). Fig. 3.3A shows the distribution of JAF's for all high confidence MiSplice predicted alternative junctions, separated by conventional mutation annotations (Fig. 3.3A). Currently, we use a JAF cut-off of 5% for reporting the final high-confidence sites. However, there are some potential alternative sites excluded by this cut-off. Our analysis revealed alternative junction expression varies widely. As expected, DNA variant allele fraction (VAF) and JAF have a generally positive correlation (Fig. 3.3B), with SCMs in *KDM6A* and *FGFR2* having >75% DNA VAF and JAF. However, a SCM in *ARID1A* has a DNA VAF of 23% and JAF of 67%. Such large ranges have been noted for mutations outside of the splice site (Broeks et al., 2003; Clarke et al., 2000; Venables, 2004). Both the truncated and normal spliced products can be observed for many variants, either due to the wild type allele or leaky splicing, for example as observed in *RNASEH2A, NFU1, SMN1, CFTR,* and *NF2* (Boerkoel et al., 1995; Caminsky et al., 2014; Ferrer-Cortes et al., 2016; Lohmann and Gallie, 2004; Mautner et al., 1996; Pagani et al., 2003; Rice et al., 2013; Svenson et al., 2001; Vezain et al., 2011) .

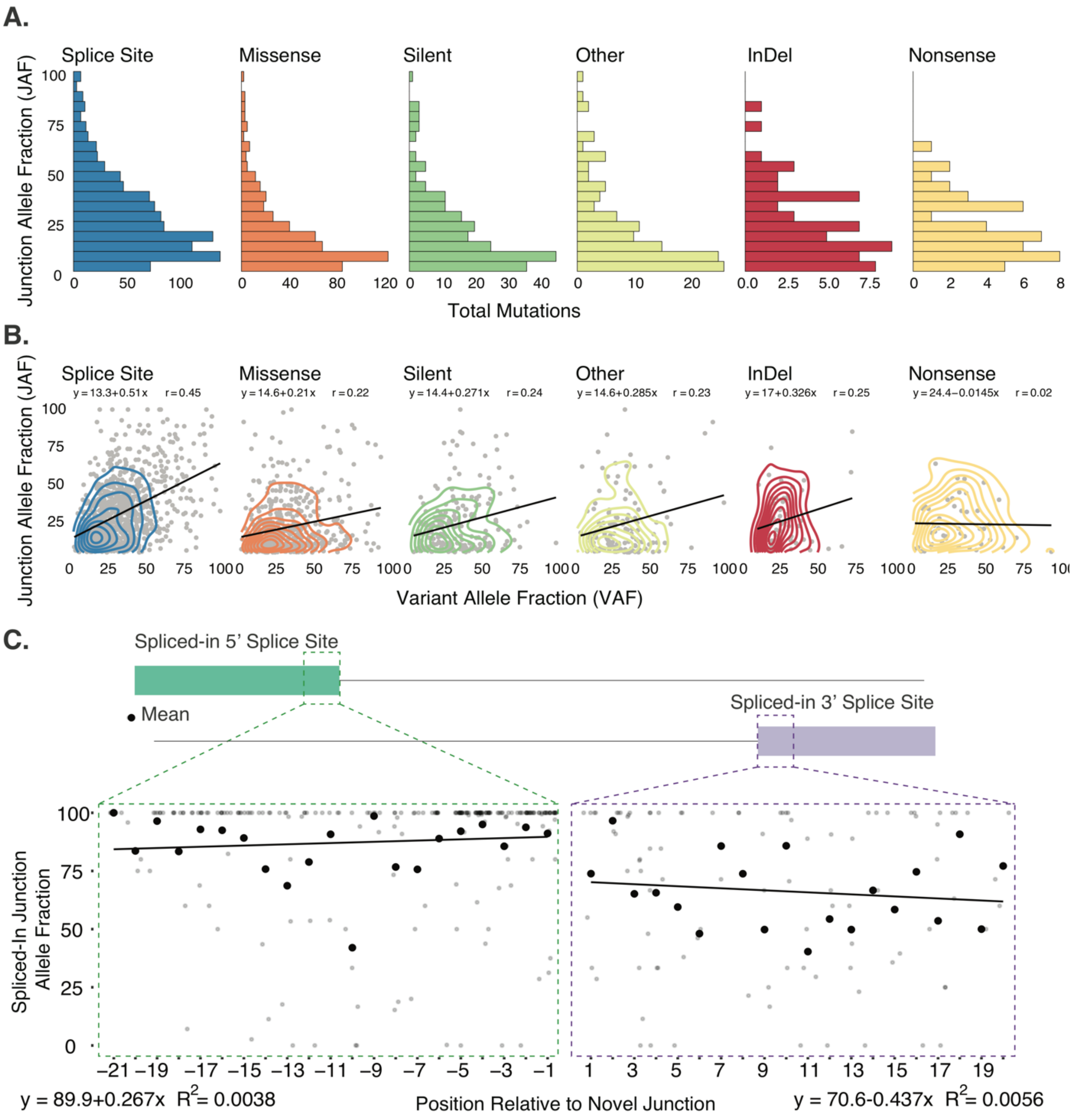


Figure 3.3. Junction allele fraction of splice site-creating mutations. (A) The junction allele fraction (JAF) is defined as the number of reads supporting the alternative spliced junction relative to total junction spanning reads. Distribution of JAF values separated by conventional annotation type. (B) JAF vs. DNA Variant Allele Fraction (VAF) comparison by conventional annotation type. Most mutation types show a generally positive correlation between JAF and VAF values. (C) Splice site-creating mutations expressed in the newly created exon of the alternative splice junction. Comparison of

mutation position relative to the percent of reads supporting the alternative junction and mutation (Spliced-In JAF). The mean of each position is highlighted by the black point. For all positions, there is a strong correlation between the presence of the splice site-creating mutation and the alternative splice junction.

We next considered the expression of mutations that are spliced-in, i.e. mutations located within the exon of the alternatively spliced product. To this end, we determined the ratio of the number of alternative junction reads containing the mutation versus total number of reads supporting the alternative junction (Fig. 3.3C). Overall, most of the reads supporting the alternative junction also support the mutation, which suggests a strong association between the mutation and alternative splice junction. Regarding the 5' splice site, mutations within the first 6 bp of the new exon junction have a much higher fraction of alternative junction reads supporting them; and we see an inverse correlation between the mutation and the junction as the distance between them increases. For the 3' splice site, we observe a similar trend, although with a higher variability as a function of the distance from the alternative junction.

SPLICE SITE-CREATING MUTATIONS ACROSS GENES AND CANCER TYPES

A total of 1,607 unique genes harbored SCMs, with 85% (1,359) having one mutation and 15% (248) having two or more. *TP53* contained the greatest number (26), followed by *GATA3* (18). While most SCMs were found outside the current cancer gene compendium (Table S1), Fig. 3.4A shows that a remarkable number of cancer genes harbor splice altering variants, a phenomenon supported in the literature (Sebestyen et al., 2016). A pan-cancer view reveals that *TP53* was the most mutated across cancer types, while 18 *GATA3* mutations and 6 *ATRX* mutations were specific to breast cancer (BRCA) and low grade glioma (LGG), respectively.

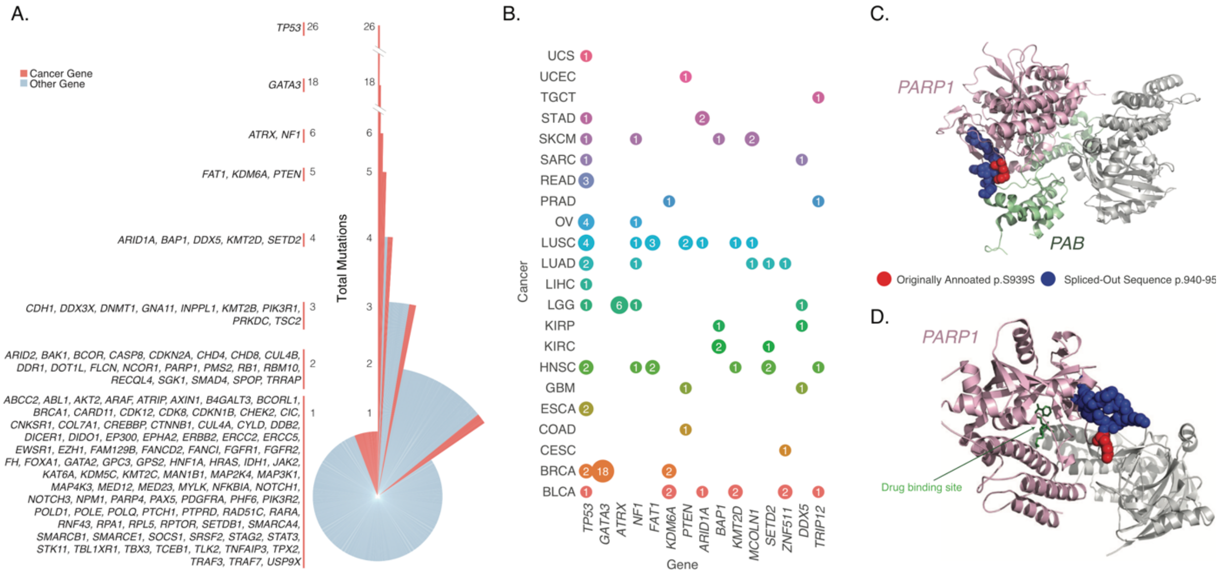


Figure 3.4. Splice site-creating mutations across genes and cancer types. (A) Distribution of splice site-creating mutations in each gene separated by total number of mutations in each gene. *TP53* has the largest number of splice site-creating mutations, followed by *GATA3* and *ATRX*. (B) Genes with the highest number of pancancer splice site-creating mutations. Circle size correlates with total number of mutations for each gene (labeled inside circle), and colored by cancer type. Splice site-creating mutations in *TP53* are present in many cancer types, while mutations in *ATRX* and *GATA3* are specific to LGG and BRCA, respectively. (C) Proteins Timeless (PAB domain) and PARP1 (Chain A) are colored green and pink, respectively. Originally annotated p.S939S mutation (red) and spliced-out sequence (blue) are highlighted on PARP1 (Chain A). (D) 3D protein structure of PARP1 in complex with an inhibitor (PDB ID: 5WRQ). Drug inhibitor and PARP1 (Chain A) are indicated with green and pink, respectively.

We observed 137 mutations nearby to one another (+/- 5 bp) which lead to the creation of the same recurrent splice site-creating events, not only in *TP53*, but also *GATA3*, *DDX5*, *KDM6A*, *SETD2*, *PTEN, SPOP,* and *BAP1*. While some mutations did not occur at the same position, 14 mutations creating the same alternative splice junction were found in the same exon, including 2 mutations in the third exon of *BAK1*. While most mutations in close proximity created the same alternative splice junction, two adjacent splice site-creating mutations in *CTNND1* and 2 nearby exonic mutations in *ACP2* and *GMPPB* created different alternative junctions.

SCMs can impact protein structure and have potential therapeutic implications. Poly ADP-ribose polymerase 1 (*PARP1*) is an enzyme involved in recruiting protein members of DNA repair pathways including Timeless PAB (PARP1 binding domain) (Fig. 3.4C) (Xie et al., 2015). Since *PARP1* is essential to many cellular processes, including DNA repair, it is commonly targeted by antitumor agents (Malyuchenko et al., 2015). *PARP1* inhibitors targeting the catalytic domain disrupt DNA repair mechanisms thereby increasing the effectiveness of chemotherapeutic agents (Fig. 3.4D). Identifying mutations that disrupt inhibitor binding are essential to properly evaluate treatment options. MiSplice identified a conventionally annotated silent *PARP1* mutation (p.S939S) in a lung squamous cell carcinoma (LUSC) patient that acts as a splice site-creating variant by creating a *de novo* donor site (Fig. 5A). 82 reads supported the *de novo* donor site, which results in a 10 amino acid deletion (p.940-p.950) that falls within the catalytic domain (Fig. 3.4D). Out of 173 LUSC control samples, none contained reads supporting the alternative junction, providing strong evidence that the annotated “silent” mutation is actually a SCM. Previous reports of missense mutations at p.940 are predicted to reduce *PARP1* enzymatic activity by disrupting the binding affinity of *PARP1* to its substrate NAD+ (Alshammari et al., 2014). The in-frame SCM is likely disturbing the local structure of *PARP1* and thereby disrupting the interactions between PARP1, its protein binding partners, and drugs binding within the pocket (Figs. 3.4C and 3.4D).

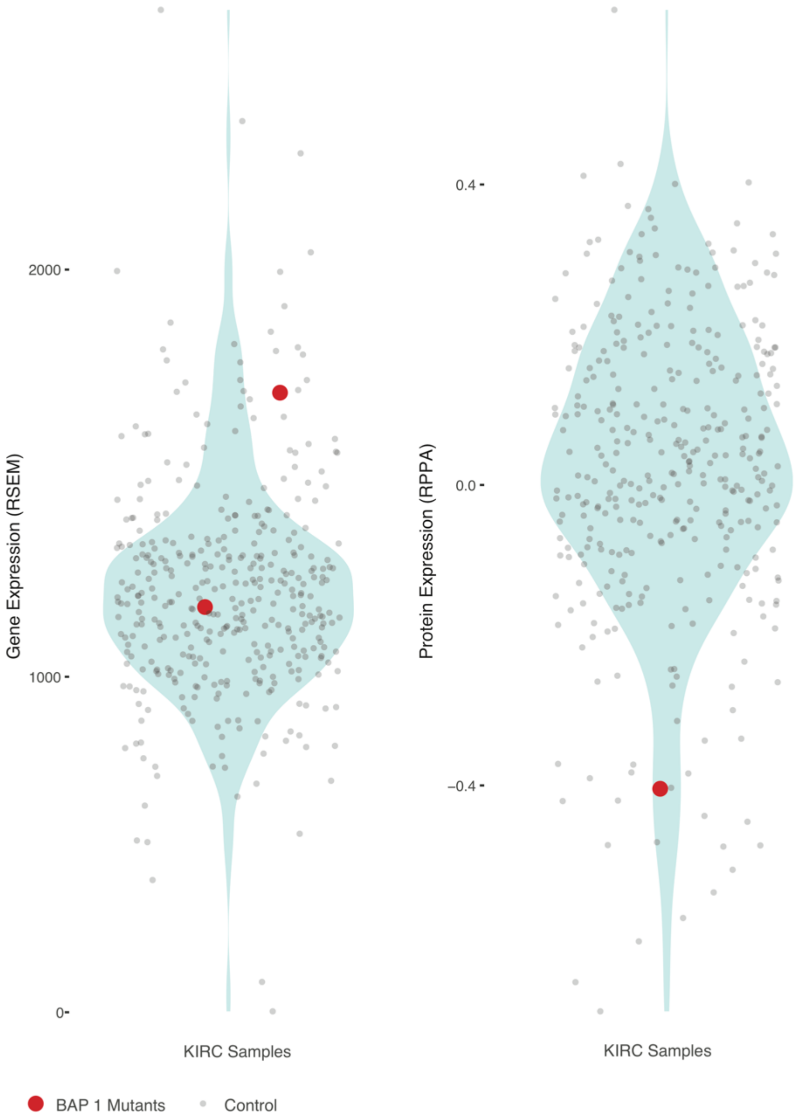


Figure S1. BAP1 gene and protein expression. Related to Figure 3.5. Violin plot of RSEM and RPPA data for control samples (grey) and novel splice creating mutant samples (red).

We identified two kidney renal clear cell carcinoma (KIRC) samples having the same conventionally annotated missense mutation (c.233A>G, p.N78S) in *BAP1*, a nuclear deubiquitinase, that created the same spliced-out alternative splicing product (Fig. 3.5B). Inactivation of *BAP1* is prevalent among renal cell carcinomas (Pena-Llopis et al., 2012) and an annotated missense mutation (p.L570V) has been reported to create a cryptic splice site in melanoma (Wadt et al., 2012). At the transcriptional level, the expressions of the case and control samples are relatively comparable, but at the translational level, one case with available protein data (RPPA) showed significantly lower expression (p=0.044, permutation test) relative to the controls (Table S2). This result suggests the conventionally annotated missense mutations in *BAP1* are likely creating an alternatively spliced transcript that is not readily expressed at the protein level.

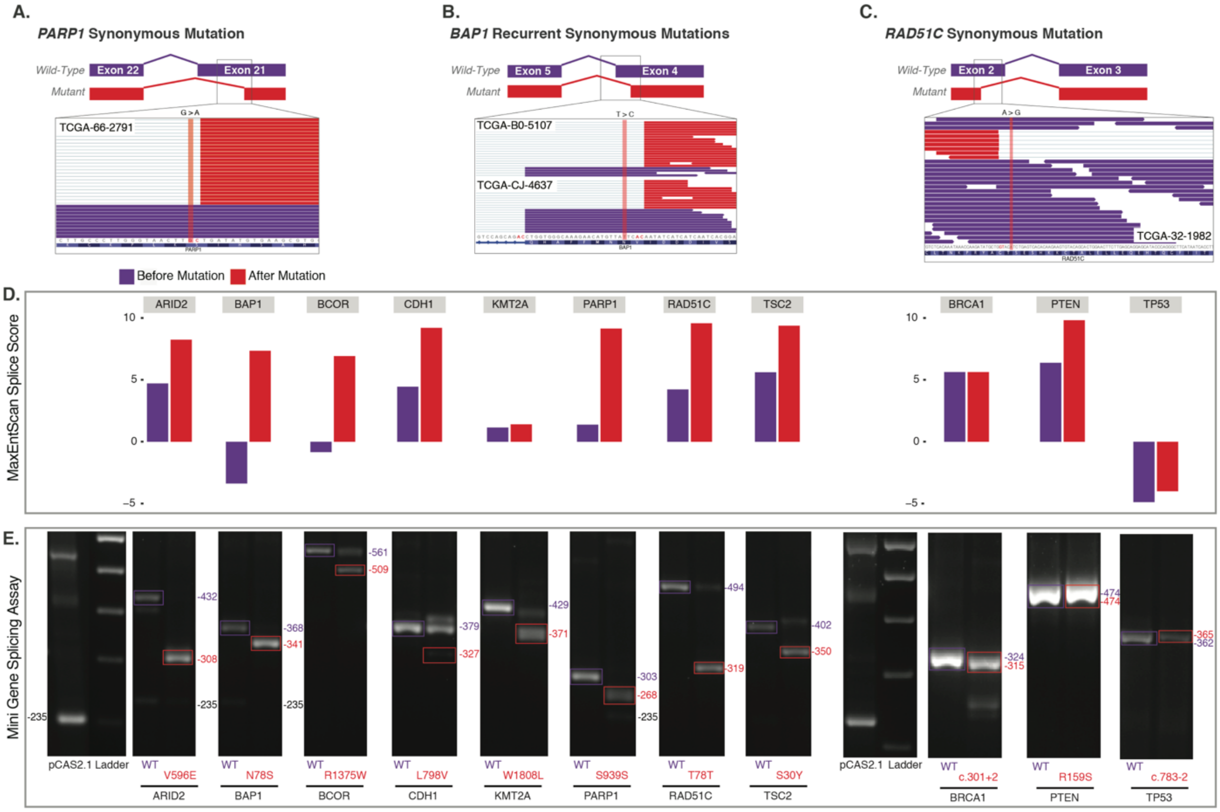


Figure 3.5. Minigene functional assay of splice site-creating mutations. (A) Integrative genomics viewer (IGV) screenshot of the conventionally annotated synonymous mutation in *PARP1* in exon 21. RNA-Seq reads of the candidate splice site-creating mutation reveals the creation of an alternative splice site (red reads) created by the conventionally annotated synonymous mutation. (B) Candidate recurrent splice site-creating mutations in *BAP1*. Conventionally annotated as synonymous variants, *BAP1* mutated region shows alternatively spliced reads (red reads) in the IGV screenshot for each sample with the splice site-creating mutation. (C) IGV screenshot of a conventionally annotated synonymous mutation in *RAD51C* in exon 2. (D) Maximum entropy score of the splice site-creating variant before (purple) and after (red) the introduced mutation for each variant functionally validated in the mini-gene splicing assay. In silico predictions suggest all mutations strengthen the alternative splice site. (E) Candidate splice site-creating mutations validated by the mini-gene splicing assay. Exons of interest were cloned into the pCAS2.1 vector and mutant (red) and wildtype (purple) plasmids were transfected into 293T cells and total RNA was extracted to identify mutation induced alternatively spliced products.

We used a pCAS2.1 splicing reporter mini-gene functional assay, adapted from previous publications (Bonnet et al., 2008; Gaildrat et al., 2010; Malone et al., 2016; Tournier et al., 2008; Vreeswijk and van der Klift, 2012), to validate SCMs in eleven cancer genes, including two originally annotated silent mutations in *PARP1*, *RAD51C*, two splice site mutations in *TP53* and *BRCA1,* and several missense mutations in *ARID2*, *BAP1*, *BCOR*, *CDH1*, *KMT2A, PTEN*, and *TSC2.* Wild-type and mutant exons were cloned into a pCAS2.1 vector (Gaildrat et al., 2010) and transiently transfected into HEK293T cells. Total RNA was extracted to evaluate alternatively spliced products by RT-PCR. Examining the change in the MaxEntScan score for the 11 genes revealed mutations in *ARID2*, *BAP1*, *BCOR*, *CDH1*, *PARP1*, *RAD51C*, *PTEN,* and *TSC2* having dramatically stronger splice scores in the presence of the mutation, while mutations in *BRCA1*, *KMT2A*, and *TP53* did not (Fig. 3.5D). Except for *PTEN*, variants with stronger splice scores showed higher levels of the alternatively spliced product in the mini-gene assay when compared to the wild-type. Variants with moderate changes in splice score still showed evidence of alternatively spliced transcripts, revealing the importance of utilizing functional assays to evaluate the effect of mutations in a splicing context in addition to *in-silico* predictions. The mini-gene assay confirmed 91% (10/11 genes) splicing alterations in all tested genes and sequencing confirmed the alternatively spliced products (Fig. 3.5E, Methods), suggesting a strong concordance between MiSplice predicted splice site-creating mutations and the functional assay.

NEO-ANTIGENS INTRODUCED BY SPLICE SITE-CREATING MUTATIONS

We have further investigated neoantigens produced by splice site-creating mutations. By using the RefSeq transcript database, a total of 2,993 protein sequences were translated for transcripts containing mutation-induced alternative splice forms (Table S3). In the translation, we allowed for different transcripts from each SCM. The HLA types for each sample were adopted from the TCGA pancan immune working group (Synapse ID: syn5974636). NetMHC4 and NetMHCpan-3.0 (Andreatta and Nielsen, 2016) were used to predict the binding affinity between epitopes and the major histocompatibility complex (MHC) and showed a high concordance in total predicted neoantigens (Pearson = 0.94, Supplementary Figure S2). We found that alternative splice forms for some important genes related to tumorigenesis, including *SMARC1*, *KDM6A*, and *NOTCH1,* are highly immunogenic and can contain 40 or more unique neoantigens (Fig. 3.6A) (Dalgliesh et al., 2010; Papadakis et al., 2015). In addition, the mean number of neoantigens across SCMs from NetMHCpan-4.0 and NetMHCpan-3.0 are 2.0 and 2.5, respectively, which are both higher than the average number of around 1 for non-synonymous mutations. Furthermore, 28 genes contain recurrent neoantigen events (more than or equal to three) across samples (Fig. 3.6B). In particular, *GATA3* has the highest recurrence and *GATA3* SCMs were mutually exclusive with other mutation types (Fig. 3.6C). The CA deletion at chr8:8111433 disrupts the canonical splice site and an alternative splice site is used for creating the alternative splice form, which results in a frame shifted protein product spanning the Zinc-finger domain (Figs. 3.6D and 3.6E). 19 unique neoantigen peptide sequences were mapped to the frameshifted protein product for the 16 samples (Fig. 3.6F). We were further able to validate one alternative peptide sequence using mass spectrometry data from a recent proteogenomics study on 77 TCGA breast cancer patients (Mertins et al., 2016). For one sample with the highly recurrent and expressed *GATA3* SCM, we used MSGF+ to search publicly available mass spectrometry data for evidence of alternative GATA3 peptides. Fig. 3.6G shows one identified mass spectrum supporting one alternative *GATA3* peptide, which covers two immunogenic peptides (KPKRRLPG and LIKPKRRLPG) predicted in TCGA-AR-A1AP.

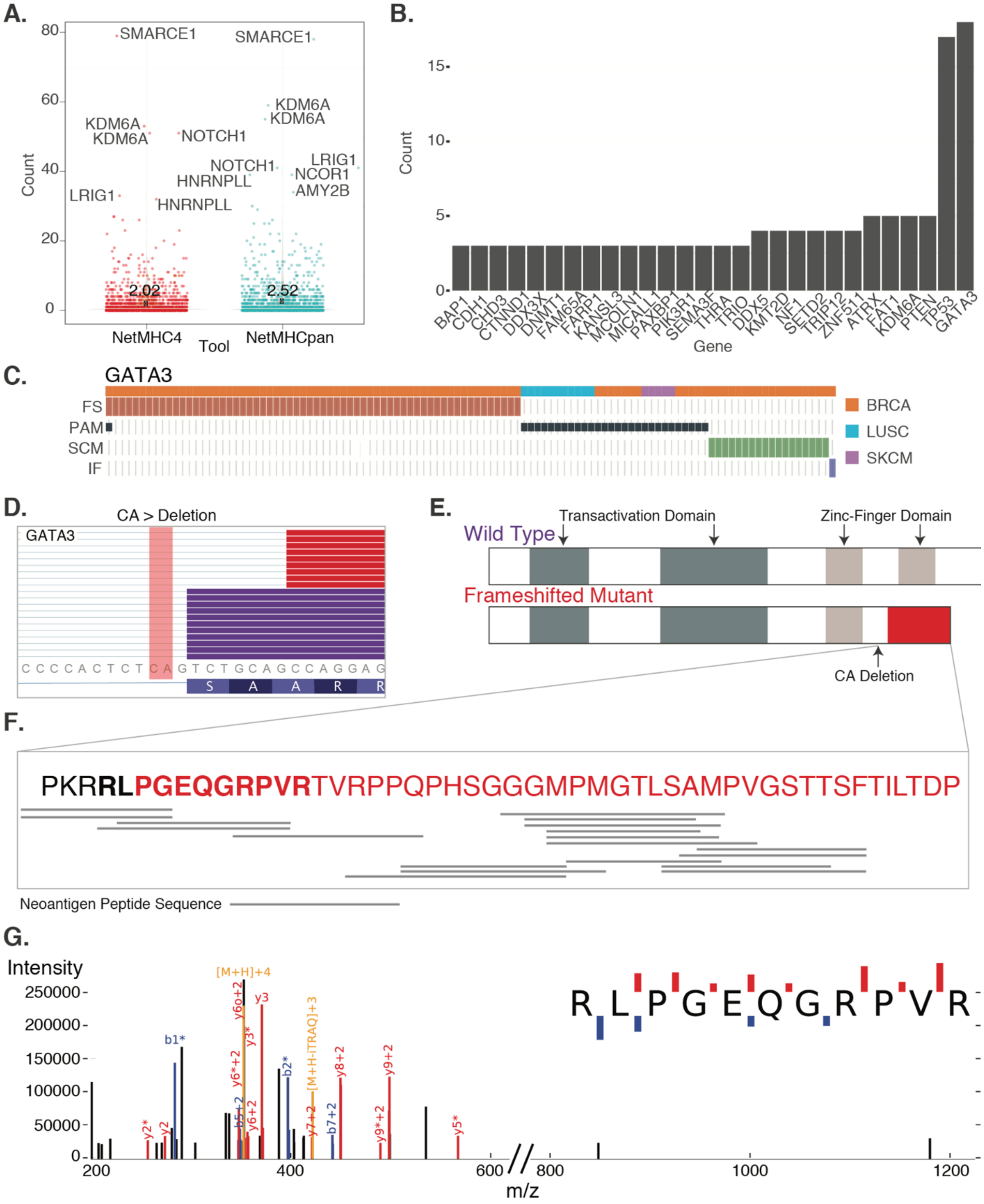


Figure 3.6. Schematic of GATA3 splice site-creating mutations and neoantigen predictions. (A)

Distribution of neoantigens predicted by NetMHCpan and NetMHC4. Genes with highest number of neoantigens labeled. Mean value for each tool indicated by X and labeled. (B) Genes with the largest recurrence of predicted neoantigens across the dataset. *GATA3* shows the highest recurrence. (C) Mutual exclusivity of protein-affecting mutation (PAM), frame-shifting indel (FS), in-frame indel (IF) and splice site-creating mutations (SCM) in *GATA3*. (D) IGV screenshot of *GATA3* splice site-creating mutation which disrupts the canonical splice site and utilizes a cryptic splice site 7 bp downstream. Mutant reads highlighted in red and normal reads in purple. CA deletion indicated in figure. (E) Predicted functional domains disrupted due to the recurrent splice site-creating mutation in *GATA3*. (F) Predicted neoantigen peptide sequences mapped to the frameshifted protein product for samples with *GATA3* splice site-creating mutations. (G) Mass spectrum of GATA3 peptide in TCGA-AR-A1AP.

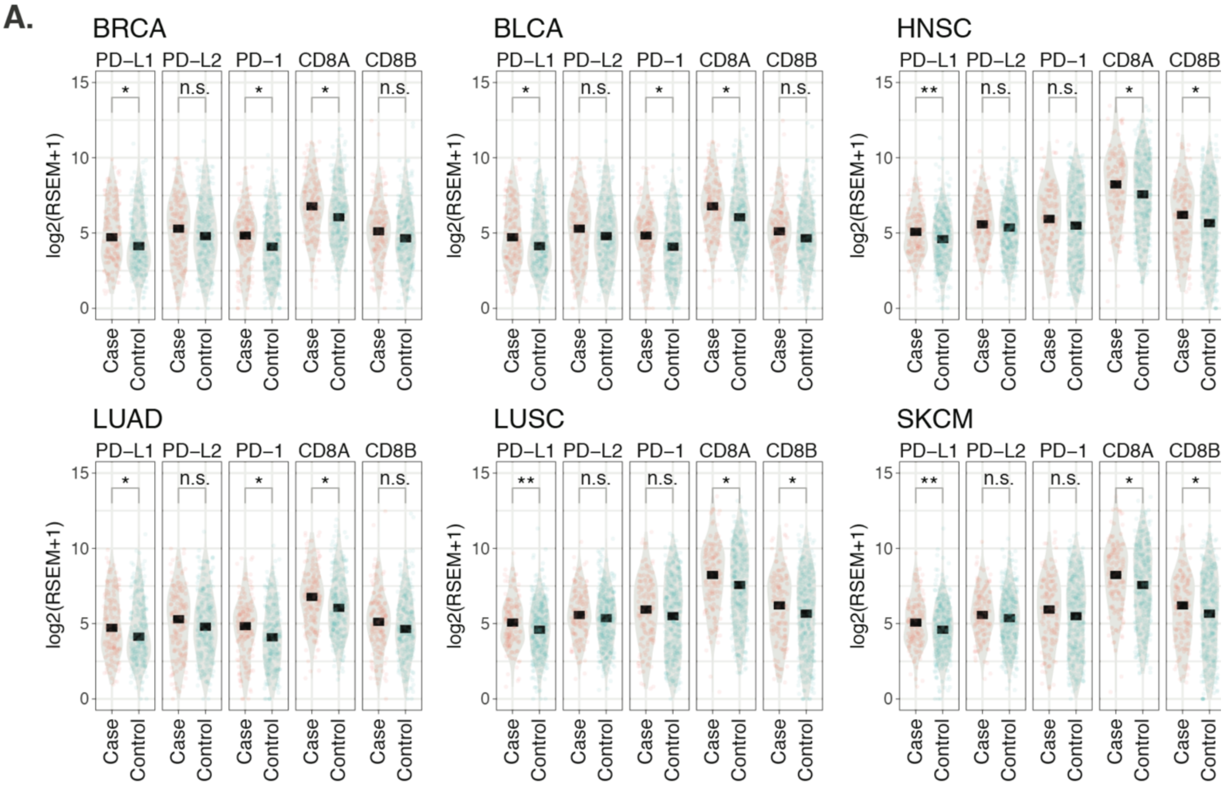


Figure 3.7. PD-L1, PD-L2, PD-1, CD8A and CD8B expression. (A) Expression comparison of PD-L1, PD-L2, and T cell markers: PD-1, CD8A, CD8B between samples with (case) and without (control) splice site-creating mutations across 6 cancer types. Symbols: * is p-value less than 0.05, ** is p-value <0.01, *** is p-value <0.001, n.s. is not significant.

High neoantigen burden is associated with an elevated immune response (Turajlic et al., 2017). To test whether SCMs affect immune response, we compared the expression of T-Cell markers PD-1, CD8A and CD8B and PD1 immune checkpoint blockades PD-L1 and PD-L2 (Fig. 3.7). We selected 6 cancer types (BRCA, BLCA, HNSC, LUAD, LUSC and SKCM) with sufficient samples containing SCMs for adequate statistical power. Both T-Cell markers (PD-1, CD8A and CD8B) and immune checkpoint blockade PD-L1 show increased expression in samples with SCMs compared to samples without SCMs (Fig. 3.7), indicating alternative splice forms induced by SCMs increase the overall immunogenicity of these cancers. The highly expressed PD-L1 suggests PD-L1 immunotherapy as potential treatments for samples containing SCMs.

3.3 Methods

DATASET DESCRIPTION

Aligned RNA-Seq bam files were analyzed using the ISB google. These cancer types are Acute Myeloid Leukemia [LAML], Adrenocortical carcinoma [ACC], Bladder Urothelial Carcinoma [BLCA], Brain Lower Grade Glioma [LGG], Breast invasive carcinoma

[BRCA], Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], Cholangiocarcinoma [CHOL], Colon adenocarcinoma [COAD], Esophageal carcinoma [ESCA], Glioblastoma multiforme [GBM], Head and Neck squamous cell carcinoma [HNSC], Kidney Chromophobe [KICH], Kidney renal clear cell carcinoma [KIRC], Kidney renal papillary cell carcinoma [KIRP], Liver hepatocellular carcinoma [LIHC], Lung adenocarcinoma [LUAD], Lung squamous cell carcinoma [LUSC], Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC], Mesothelioma [MESO], Ovarian serous cystadenocarcinoma [OV], Pancreatic adenocarcinoma [PAAD], Pheochromocytoma and Paraganglioma [PCPG], Prostate adenocarcinoma [PRAD], Rectum adenocarcinoma [READ], Sarcoma [SARC], Skin Cutaneous Melanoma [SKCM], Stomach adenocarcinoma [STAD], Testicular Germ Cell Tumors [TGCT], Thymoma [THYM], Thyroid carcinoma [THCA], Uterine Carcinosarcoma [UCS], Uterine Corpus Endometrial Carcinoma [UCEC], Uveal Melanoma [UVM]

MISPLICE PIPELINE

The MiSplice pipeline was developed to detect mutation-induced splicing events from RNA-Seq data. It is written in Perl and incorporates two standard tools, samtools and MaxEntScan. The pipeline is fully automated and can run multiple jobs in parallel on LSF cluster. It executes the following steps:

1) Splitting large maf file into multiple smaller files with less mutations (currently, the default setting is 200).

2) Discovering splicing junctions within 20bps of the mutation with at least 5 supporting reads with mapping quality Q20 and then filtering canonical junctions by using the Ensembl 37.75 database. We selected 20bp as a cut-off since it is the farthest distance from the splice junction in a splice region.

3) Computing the number of supporting reads of above cryptic splice sites for control samples without mutations (Table S1).

4) Calculating the splicing scores for the cryptic splice sites via MaxEntScan.

5) Reporting the depth of each cryptic splice site via Samtools.

6) Filtering cryptic sites which fall in HLA loci or less than 5% of reads at the genomic location supporting the alternative junction of interest.

7) Further filtering cryptic sites by comparing the supporting reads in control samples. The final reported cryptic sites must stand as top 5% for the number of supporting reads in the case (with mutation).

SPLICE SITE SCORE ESTIMATION

For each cryptic splice site and nearby canonical splice site, the corresponding nucleotide sequences were first extracted for both the mutant and reference sequences (9-mer and 23-mer for donor and acceptor, respectively). Their splice scores as potential donor or acceptor sites were then estimated using MaxEntScan.

NEOANTIGEN PREDICTION

For each predicted SCM, we use a curated RefSeq transcript database (version 20130722) to obtain the translated protein sequences for transcript containing alternative splice forms induced by SCMs. Different length of epitopes (8mer, 9mer, 10mer and 11mer) are constructed from the translated protein sequence. We use NetMHC3pan(Nielsen and Andreatta, 2016) and NetMHC4(Andreatta and Nielsen, 2016) to predict the binding affinity between epitopes and MHC. Epitopes with binding affinity <=500nM which are also not present in the wild-type transcript are extracted from the following neoantigen analysis.

MANUAL REVIEW

All splice site-creating mutations were manually reviewed using the integrative genomics viewer (http://software.broadinstitute.org/software/igv/). Mutations were placed into one of three categories: Pass, Complex, and No Support. Mutations were classified as complex if more than one alternatively spliced product was observed for the mutated sample.

CODE AVAILABILITY

MiSplice is written in Perl and is freely available from GitHub at https://github.com/ding-lab/misplice under the GNU general public license. MiSplice uses several independent tools and packages, including SamTools and MaxEntScan, all of which are likewise freely available, but which must be obtained from their respective developers. The MiSplice documentation contains complete instructions for obtaining and linking these applications into MiSplice.

MINI GENE SPLICING ASSAY

Exons of interest and approximately 150 bp of their flanking intron sequences were PCR amplified from HEK293T genomic DNA using primers carrying restriction enzyme sites for BamH1 and MluI. PCR products were cleaned up using NucleoSpin PCR Cleanup (Macherey-Nagel) or DNA Clean and Concentrator-5 Kit (Zymo Research) and digested with BamHI and MluI. The digested pCAS2.1 vector and PCR products were ligated using T4 DNA Ligase (NEB). Mutations were introduced via Q5 Site-Directed Mutagenesis (NEB). WT and MUT constructs were confirmed by sequencing of the insert region. The plasmids were transiently transfected into HEK293T cells using Lipofectamine 2000 (ThermoFisher Scientific). 24 hours post transfection, cDNA was synthesized using 2 to 3 ug of total RNA with the Superscript III First-Strand Synthesis System (ThermoFisher Scientific) and priming with Oligo(dT)20. Finally, cDNA was amplified using pCAS-KO1-(5′-TGACGTCGCCGCCCATCAC-3′) and pCAS-R (5′-ATTGGTTGTTGAGTTGGTTGTC-3′) and the alternative splicing patterns were evaluated on a 2.5% agarose gel with ethidium bromide. Qiaquick Gel Extraction Kit (Qiagen) was used to purify bands for sequencing (Table S3.1-3.2, Figure S3-S6).

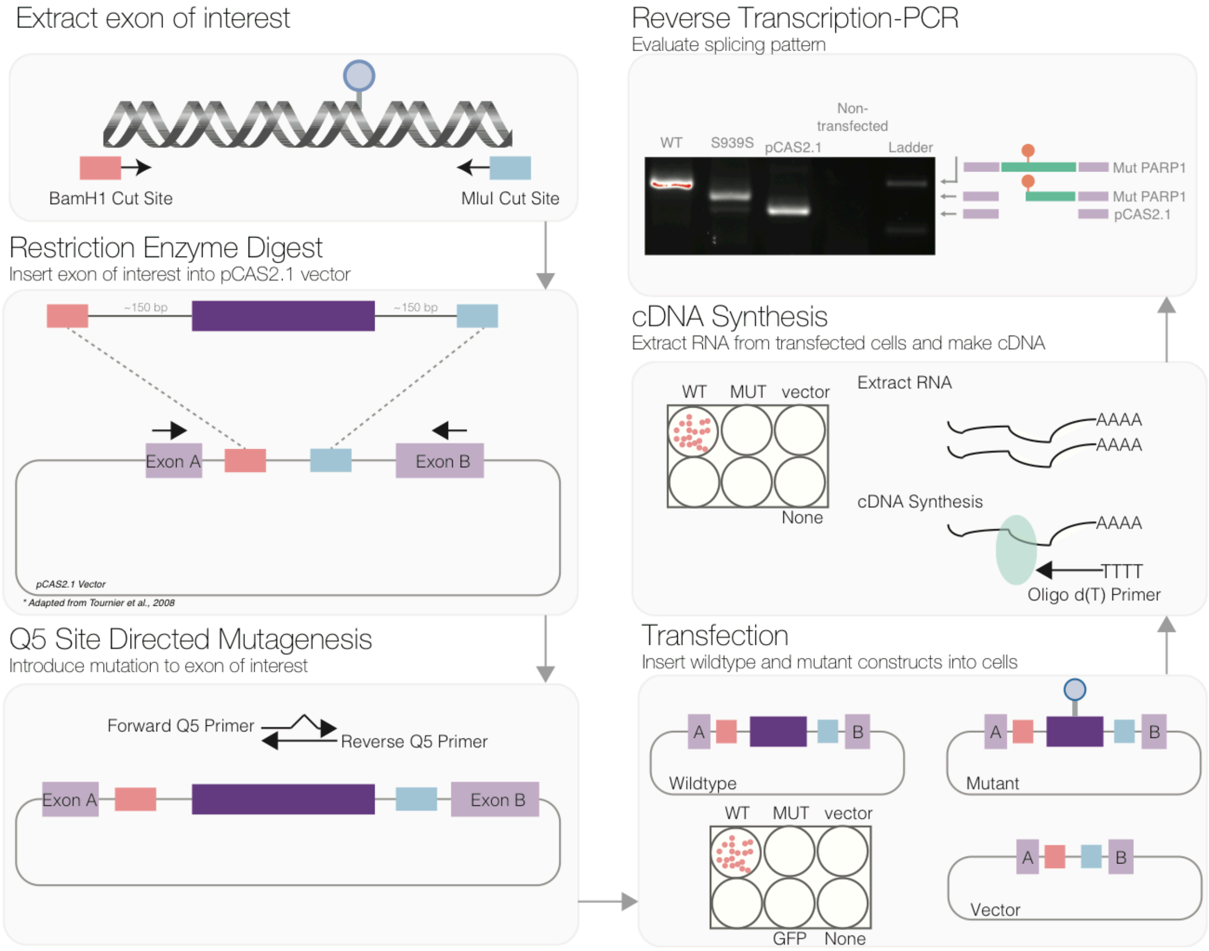


Figure S2: Mini-gene splicing assay.

Table 3.1: Mutation information for splice site-creating mutations validated in minigene assay.

Gene	Sample	Position	Reference	Mutant	Coding Change	Amino Acid Change	Transcript
ARID2	TCGA-FS-A1Z3	12: 46243434	T	A	c.1787T>A	p.Val596Glu	ENST00000334344

TSC2	TCGA-EE-A17Y	16:2098705	C	A	c.89C>A	p.Ser30Tyr	ENST00000219476
CDH1	TCGA-GC-A3I6	16:68863653	C	G	c.2392C>G	p.Leu798Val	ENST00000219476
TP53	TCGA-FG-A60J	17:7577157	T	G	c.783-2A>C	p.X261_splice	ENST00000269305
RAD51C	TCGA-32-1982	17:56772380	A	G	c.234A>G	p.Thr78Thr	ENST00000337432
BCOR	TCGA-DM-A1HA	X:39922049	A	G	c.4123C>T	p.Arg1375Trp	ENST00000378444.4
BAP1	TCGA-CJ-4637 and TCGA-B0-5107	3: 52442512	T	C	c.233A>G	p.Asn78Ser	ENST00000460680
PARP1	TCGA-66-2791	1:226550831	G	A	c.2817C>T	p.Ser939Ser	ENST00000366794
BRCA1	TCGA-D6-6823	17:41256883	A	C	c.301+2T>G	p.X101_splice	ENST00000471181
PTEN	TCGA-06-2559	10:89692993	G	T	c.477G>T	p.Arg159Ser	ENST00000371953
KMT2A	TCGA-55-7994	11:118366474	G	T	c.5423G>T	p.Trp1808Leu	ENST00000534358

Table 3.2: Predicted alternative and wild-type RT-PCR splice products from mini-gene splicing assay.

Type	RT-PCR Sequence
pCAS2.1	TGACGTCGCCGCCCATCACGCCTCCAGGCTGACCCTGCTGACCCTCCTGCTGCTGCTGCTGGCTGG**GG**ATAGAGCCTCCTCAAATCCAAATGCTACCAGCTCCAG

	CAGCCAAGATCCAGAGAGTTTGCAAGACAGAGGCGAAGGGAAGGTCGCAACAACAGTTATCTCCAAGATGCTATTCGTTGAACCCATCCTGGAGGTTTCCAGCTTGCCGACAACCAACTCAACAACCAAT
BAP1 Wild Type	CCCTGTATATGGATTTATCTTCCTGTTCAAATGGATCGAAGAGCGCCGGTCCCGGCGAAAGGTCTCTACCTTGGTGGATGATACGTCCGTGATTGATGATGATATTGTGAATAACATGTTCTTTGCCCACCAG
BAP1 Mutant	CCCTGTATATGGATTTATCTTCCTGTTCAAATGGATCGAAGAGCGCCGGTCCCGGCGAAAGGTCTCTACCTTGGTGGATGATACGTCCGTGATTGATGATGATATT
TP53 Wild Type	CTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCTCCCCTTTCTTGCGGAGATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCA
TP53 Mutant	CTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCTCCCCTTTCTTGCGGAGATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCACGA
BCOR Wild Type	CTTGCCATCGGCATTCTCCACGTAGTATTCCCCTGTCAGTGGCAATCCCCGCCTGGACTCCTGAGGGATCAAGTGTTTGGTTTTGCACAGTCTCTTCCCGGATGGCTTCTCGCTGTTGTCGGTGTATTTCTGCAGCAGGGAGGCAGCCTGGCAATCCTCTTCTTCGTCTGCACACAGCACATCTGTCTTCTGGTTTTCTTTAATTTTCTGCTGTTTGGCAGGCGGCCTGGAGGCTGGTGCGCAGCTTGGCTGAGCCTGCTTTTTGCCGCCTGCACTGGTGGATGAAAGACTCTTCATGGGCGGAGAGCCGGAGAACACAGGCAAGC
BCOR Mutant	CTGGACTCCTGAGGGATCAAGTGTTTGGTTTTGCACAGTCTCTTCCCGGATGGCTTCTCGCTGTTGTCGGTGTATTTCTGCAGCAGGGAGGCAGCCTGGCAATCCTCTTCTTCGTCTGCACACAGCACATCTGTCTTCTGGTTTTCTTTAATTTTCTGCTGTTTGGCAGGCGGCCTGGAGGCTGGTGCGCAGCTTGGCTGAGCCTGCTTTTTGCCGCCTGCACTGGTGGATGAAAGACTCTTCATGGGCGGAGAGCCGGAGAACACAGGCAAGC

RAD51C Wild Type	AAGTTGGGATATCTAAAGCAGAAGCCTTAGAAACTCTGCAAATTATCAGAAGAGAATGTCTCACAAATAAACCAAGATATGCTGGTACATCTGAGTCACACAAGAAGTGTACAGCACTGGAACTTCTTGAGCAGGAGCATACCCAGGGCTTCATAATCACCTTCTGTTCAGCACTAGATGATATTCTTGGGGGTGGAGTGCCCTTAATGAAAACAACAGAAATTTGTGGTGCACCAGGTGTTGGAAAAACACAATTATG
RAD51C Mutant	AAGTTGGGATATCTAAAGCAGAAGCCTTAGAAACTCTGCAAATTATCAGAAGAGAATGTCTCACAAATAAACCAAGATATGCTG
KMT2A Wild Type	CAGTGGGATGTTACCAAACGCAGTGCTTCCACCTTCACTTGACCATAATTATGCTCAGTGGCAGGAGCGAGAGGAAAACAGCCACACTGAGCAGCCTCCTTTAATGAAGAAAATCATTCCAGCTCCCAAACCCAAAGGTCCTGGAGAACCAGACTCACCAACTCCTCTGCATCCTCCTACACCACCAATTTTGA
KMT2A Mutant	TTGCAGGAGCGAGAGGAAAACAGCCACACTGAGCAGCCTCCTTTAATGAAGAAAATCATTCCAGCTCCCAAACCCAAAGGTCCTGGAGAACCAGACTCACCAACTCCTCTGCATCCTCCTACACCACCAATTTTGA
PARP1 Wild Type	CTTTGACACTGTGCTTGCCCTTGGGTAACTTGCTGATATGTGAAGCGTGCTTCAGTTCATAC
PARP1 Mutant	TGATATGTGAAGCGTGCTTCAGTTCATAC
BRCA1 Wild Type	ACTCCAAACCTGTGTCAAGCTGAAAAGCACAAATGATTTTCAATAGCTCTTCAACAAGTTGACTAAATCTCGTACTTTCTTGTAGGCTC
BRCA1 Mutant	CTGTGTCAAGCTGAAAAGCACAAATGATTTTCAATAGCTCTTCAACAAGTTGACTAAATCTCGTACTTTCTTGTAGGCTC
ARID2 Wild Type	AACGGTCTTTCCAAATCATACAGTGAAGAGAGTGGAGGATTCCAGTAGCAATGGGCAGGCACATATTCATGTGGTAGGAGTAAAACGGAGGGCTATACCACTTCCCATTCAGATGTACTATCAGCAGCAACCAGTTTCTACTTCTGTTGTTCGTGTTGATTCTGTTCCTGATGTATCTCCTGCTCCTTCACCTGCAG
ARID2 Mutant	AACGGTCTTTCCAAATCATACAGTGAAGAGAGTGGAGGATTCCAGTAGCAATGGGCAGGCACATATTCATGAG

PTEN Wild Type	TTGCACAATATCCTTTTGAAGACCATAACCCACCACAGCTAGAACTTATCAAACCCTTTTGTGAAGATCTTGACCAATGGCTAAGTGAAGATGACAATCATGTTGCAGCAATTCACTGTAAAGCTGGAAAGGGACGAACTGGTGTAATGATATGTGCATATTTATTACATCGGGGCAAATTTTTAAAGGCACAAGAGGCCCTAGATTTCTATGGGGAAGTAAGGACCAGAGACAAAAAG
PTEN Mutant	TTGCACAATATCCTTTTGAAGACCATAACCCACCACAGCTAGAACTTATCAAACCCTTTTGTGAAGATCTTGACCAATGGCTAAGTGAAGATGACAATCATGTTGCAGCAATTCACTGTAAAGCTGGAAAGGGACGAACTGGTGTAATGATATGTGCATATTTATTACATCGGGGCAAATTTTTAAAGGCACAAGAGGCCCTAGATTTCTATGGGGAA
CDH1 Wild Type	GACTTTGACTTGAGCCAGCTGCACAGGGGCCTGGACGCTCGGCCTGAAGTGACTCGTAACGACGTTGCACCAACCCTCATGAGTGTCCCCCGGTATCTTCCCCGCCCTGCCAATCCCGATGAAATTGGAAATTTTATTGATGAA
CDH1 Mutant	GACTTTGACTTGAGCCAGCTGCACAGGGGCCTGGACGCTCGGCCTGAAGTGACTCGTAACGACGTTGCACCAACCCTCATGAGTGTCCCCCG
TSC2 Wild Type	AGGGGTTTTCTGGTGCGTCCTGGTCCACCATGGCCAAACCAACAAGCAAAGATTCAGGCTTGAAGGAGAAGTTTAAGATTCTGTTGGGACTGGGAACACCGAGGCCAAATCCCAGGTCTGCAGAGGGTAAACAGACGGAGTTTATCATCACCGCGGAAATACTGAGA
TSC2 Mutant	AGGGGTTTTCTGGTGCGTCCTGGTCCACCATGGCCAAACCAACAAGCAAAGATTCAGGCTTGAAGGAGAAGTTTAAGATTCTGTTGGGACTGGGAACACCGAGGCCAAATCCCAG

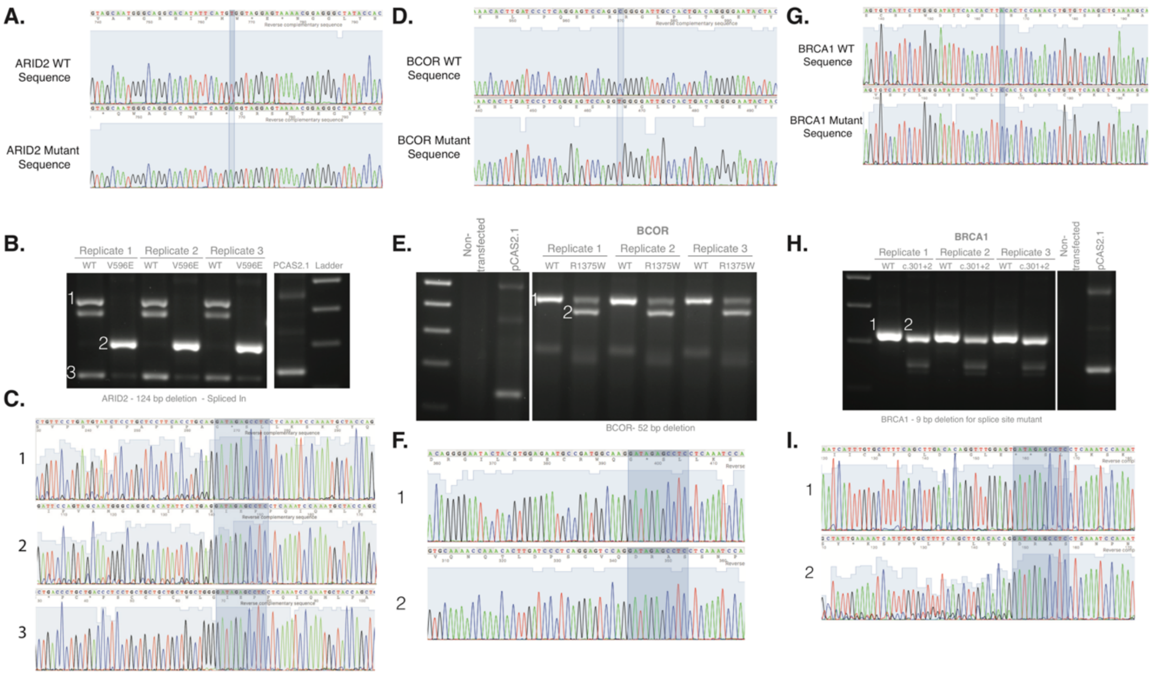


Figure S3. ***ARID2, BCOR, BRCA1*** **Mini-Gene results. Related to Figure 3.5.** (A,D,G) DNA chromatograms verifying *ARID2, BCOR, and BRCA1* wildtype and mutant sequencing results, respectively. Mutation position is highlighted. (B,E,H) Reverse transcriptase PCR (RT-PCR) with wild type and mutant plasmids, results in triplicate. Numbered bands are sequenced for confirmation. (C,F,I) DNA chromatograms of RT-PCR bands sequenced. Highlighted sequence indicates boundary of pCAS2.1 plasmid.

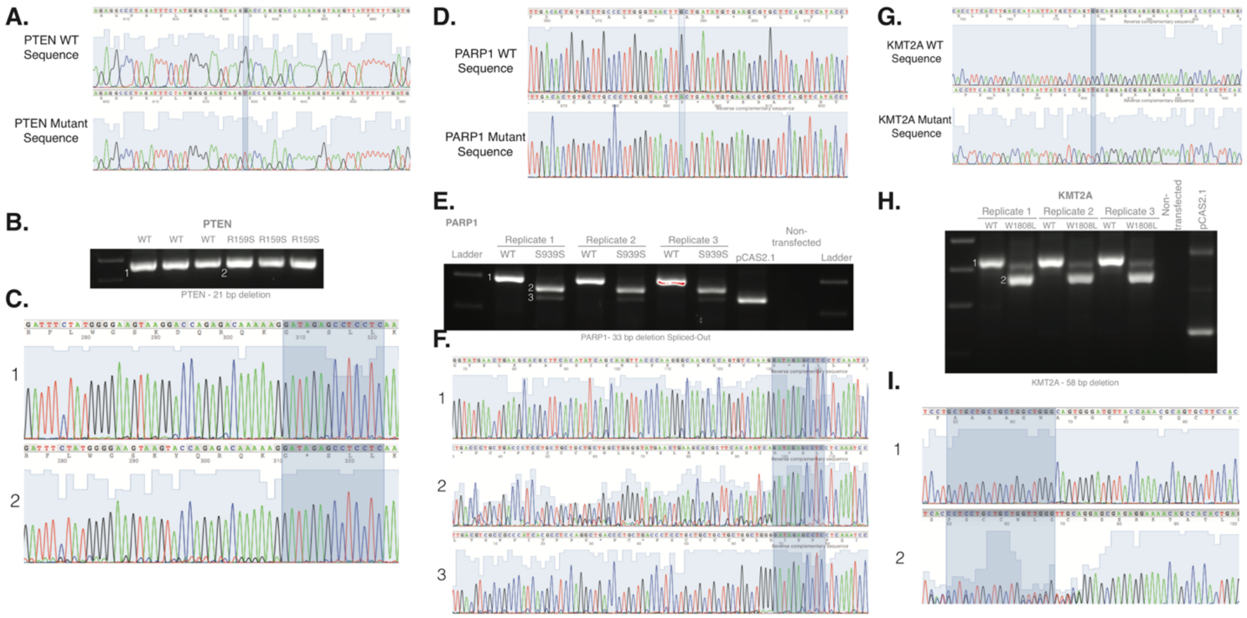


Figure S4. *PTEN, PARP1, KMT2A* Mini-Gene results. Related to Figure 3.5. (A,D,G) DNA chromatograms verifying *PTEN, PARP1, and KMT2A* wildtype and mutant sequencing results, respectively. Mutation position is highlighted. (B,E,H) Reverse transcriptase PCR (RT-PCR) with wild type and mutant plasmids, results in triplicate. Numbered bands are sequenced for confirmation. (C,F,**I**) DNA chromatograms of RT-PCR bands sequenced. Highlighted sequence indicates boundary of pCAS2.1 plasmid.

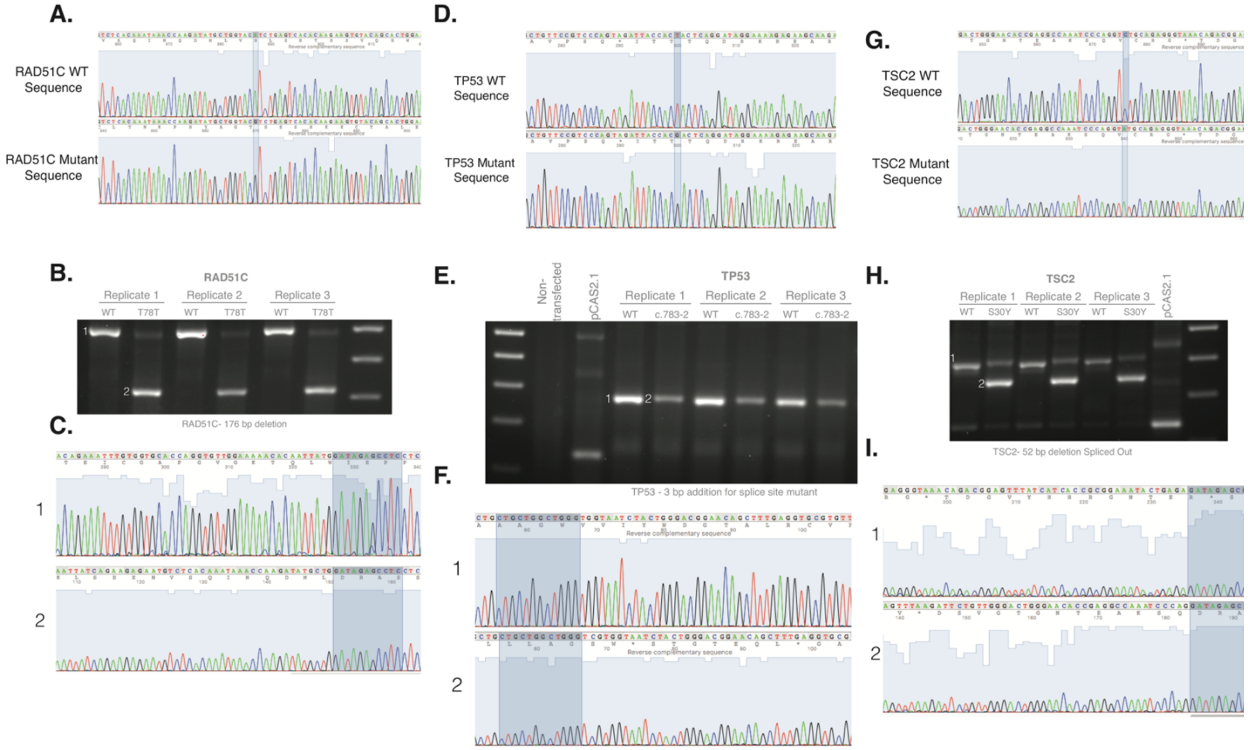


Figure S5. ***RAD51C, TP53, TSC2*** **Mini-Gene results. Related to Figure 3.5.** (A,D,G) DNA chromatograms verifying *RAD51C, TP53, and TSC2* wildtype and mutant sequencing results, respectively. Mutation position is highlighted. (B,E,H) Reverse transcriptase PCR (RT-PCR) with wild type and mutant plasmids, results in triplicate. Numbered bands are sequenced for confirmation. (C,F,**I**) DNA chromatograms of RT-PCR bands sequenced. Highlighted sequence indicates boundary of pCAS2.1 plasmid.

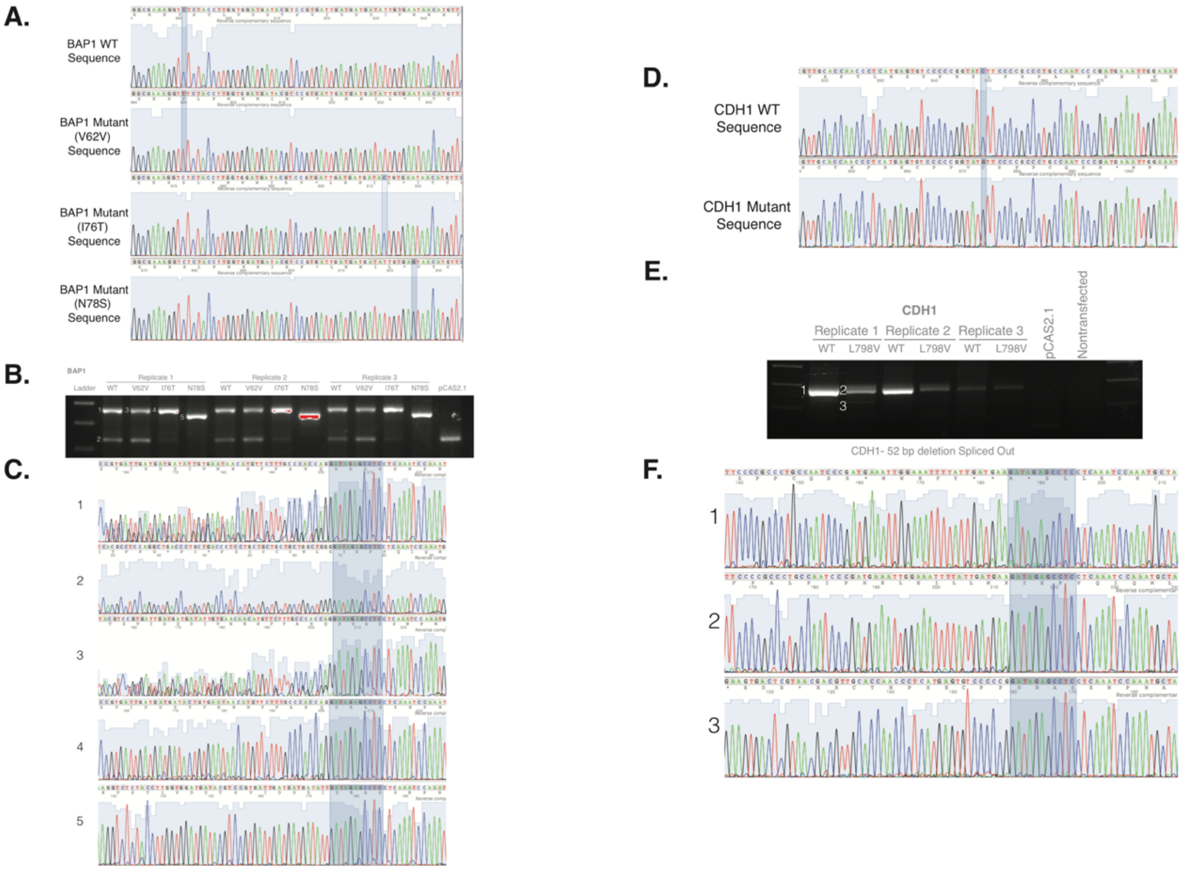


Figure S6. *BAP1 and CDH1* Mini-Gene results. Related to Figure 3.5. (A,D) DNA chromatograms verifying *BAP1* and *CDH1* wildtype and mutant sequencing results, respectively. Mutation position is highlighted. (B,E) Reverse transcriptase PCR (RT-PCR) with wild type and mutant plasmids, results in triplicate. Numbered bands are sequenced for confirmation. (C,F) DNA chromatograms of RT-PCR bands sequenced. Highlighted sequence indicates boundary of pCAS2.1 plasmid.

CELL CULTURE

HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with fetal bovine serum (FBS) and penicillin streptomycin.

STATISTICAL ANALYSES

MiSplice assesses the significance of the number of reads supporting the predicted alternative splice junction by comparing to read counts from a control cohort. Specifically, a frequency distribution is constructed from the control cohort, from which threshold values for 5% and 95% tails on the left and right, respectively, are determined. A series of logic tests is then conducted to discern the best explanation of the data. Possible verdicts are low or high expression if the datum is outside the 5% or 95% thresholds, respectively, average expression if no thresholds are exceeded, or no expression in this tissue if the thresholds are zero.

3.4 Discussion

In this study, we applied our newly developed bioinformatics tool called MiSplice (Mutation Induced Splicing) to systematically analyze splice site-creating events that arise from somatic mutations. Our analysis shows MiSplice reliably identifies SCMs across multiple cancer types. Existing studies have largely focused on splice-disrupting events in known splice sites, but the current study substantially extends our knowledge into the realm of splice site-creating mutations in human cancer. For instance, we found 1,016 splice site mutations not only disrupt the canonical splice site, but also create an alternative splice site. We also found that hundreds of mutations that would traditionally be classified as missense, silent, indel, and nonsense are really acting as SCMs. Many important cancer-related genes harbor these mutations, such as *TP53*, *ATRX*, *BAP1*, *CTNNB1*, *RB1*, etc. It is noteworthy that we found five splice site-creating mutations in *ATRX* among 288 lower grade glioma cases, likely leading to the disruption of ATRX function. A previous study has shown that loss of wild-type *ATRX* is associated with tumor growth in glioma (Koschmann et al., 2016).

Characterization of these alternative splice events show that most SCMs have a higher splice score, as measured by MaxEntScan, in the post-mutation alternative splice site as compared to the reference. These results are consistent with the preferential selection of these alternative sites as new splicing forms. For the splice-site mutation, the splice score associated with the canonical junction is coincidently decreased after mutation. However, while there is no difference in splice scores of canonical junctions before and after missense and silent mutations, the alternative splice site was often strengthened after mutation. This suggests silent and missense mutations instead act as modifiers of splicing by creating or strengthening cryptic sites within the exon as opposed to disrupting the canonical splice site. In addition, we found a significant enrichment of mutations at the -3 position in the 3' splice site, the two dominant sequence contexts being aGag and agGag, where G is at the -3 position.

In cases where the mutation is retained in the alternative splice junction, we distinguish mutations with two further categories, splice-in and splice-out. For splice-in mutations, we can characterize the association between mutations and cryptic splicing forms. For example, we found high concordance for RNA-seq reads supporting alternatively spliced junctions and mutations, suggesting the association between mutations and cryptic splicing forms.

The current study has greatly extended the insight into the transcriptional ramifications of genomic alterations by identifying nearly 1,964 alternative splice sites introduced by somatic mutations and functionally validating ten of eleven variants in a mini-gene splicing assay. These events were conventionally annotated as missense, silent, splice site, nonsense, or other mutations when, in fact, we have shown that they often create cryptic splice sites. The relative abundance of the alternative and wild-type product suggests varying levels of junction usage, depending on the context of the mutation, and emphasizes the importance of validating predictions using a functional assay to understand the full biological consequence. The alternative products may be therapeutically targetable in some cancer patients. For example, the targeting of neoantigens shows promising result in treating melanoma patients (Carreno et al., 2015). By further evaluating human leukocyte antigen (HLA) genotypes and binding affinities to the major histocompatibility complex (MHC), it is likely that new neoantigens from cryptic splice sites may be discovered. The current study reveals that alternative splice forms induced by splice site-creating mutations are highly immunogenic and correlated with a high T-Cell immune response and an elevated PD-L1 expression, suggesting potential for immunotherapy in these samples. Further investigation of the cryptic splice sites by mass spectra

3.5 References

Uncategorized References

Alshammari, A.H., Shalaby, M.A., Alanazi, M.S., and Saeed, H.M. (2014). Novel mutations of the PARP-1 gene associated with colorectal cancer in the Saudi population. Asian Pac J Cancer Prev *15*, 3667-3673.

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511-517.

Boerkoel, C.F., Exelbert, R., Nicastri, C., Nichols, R.C., Miller, F.W., Plotz, P.H., and Raben, N. (1995). Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II. Am J Hum Genet *56*, 887-897.

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V.*, et al.* (2008). Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. J Med Genet *45*, 438-446.

Broeks, A., Urbanus, J.H.M., de Knijff, P., Devilee, P., Nicke, M., Klöpper, K., Dörk, T., Floore, A.N., and van't Veer, L.J. (2003). IVS10-6T>G, an ancient ATM germline mutation linked with breast cancer. Human mutation *21*, 521-528.

Caminsky, N., Mucaki, E.J., and Rogan, P.K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Res *3*, 282.

Carreno, B.M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A.A., Ly, A., Lie, W.R., Hildebrand, W.H., Mardis, E.R.*, et al.* (2015). Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science *348*, 803-808.

Chen, L.L., Sabripour, M., Wu, E.F., Prieto, V.G., Fuller, G.N., and Frazier, M.L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. Oncogene *24*, 4271-4280.

Clarke, L.a., Veiga, I., Isidro, G., Jordan, P., Ramos, J.S., Castedo, S., and Boavida, M.G. (2000). Pathological exon skipping in an HNPCC proband with MLH1 splice acceptor site mutation. Genes, chromosomes & cancer *29*, 367-370.

Dalgliesh, G.L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C.*, et al.* (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature *463*, 360-363.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R.*, et al.* (2012). MuSiC: identifying mutational significance in cancer genomes. Genome research *22*, 1589-1598.

Ferrer-Cortes, X., Narbona, J., Bujan, N., Matalonga, L., Del Toro, M., Arranz, J.A., Riudor, E., Garcia-Cazorla, A., Jou, C., O'Callaghan, M.*, et al.* (2016). A leaky splicing mutation in NFU1 is associated with a particular biochemical phenotype. Consequences for the diagnosis. Mitochondrion *26*, 72-80.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frebourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol Biol *653*, 249-257.

Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res *42*, 13534-13544.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat Genet *47*, 1242-1248.

Kahles, A., Ong, C.S., Zhong, Y., and Ratsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics *32*, 1840-1847.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.a.*, et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Koschmann, C., Calinescu, A.A., Nunez, F.J., Mackay, A., Fazal-Salom, J., Thomas, D., Mendez, F., Kamran, N., Dzaman, M., Mulpuri, L.*, et al.* (2016). ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma. Sci Transl Med *8*, 328ra328.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.a.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lim, K.H., and Fairbrother, W.G. (2012). Spliceman-A computational web server that predicts sequence variations in pre-mRNA splicing. Bioinformatics *28*, 1031-1032.

Lohmann, D.R., and Gallie, B.L. (2004). Retinoblastoma: revisiting the model prototype of inherited cancer. Am J Med Genet C Semin Med Genet *129C*, 23-28.

Malone, A.F., Funk, S.D., Alhamad, T., and Miner, J.H. (2016). Functional assessment of a novel COL4A5 splice region variant and immunostaining of plucked hair follicles as an alternative method of diagnosis in X-linked Alport syndrome. Pediatr Nephrol.

Malyuchenko, N.V., Kotova, E.Y., Kulaeva, O.I., Kirpichnikov, M.P., and Studitskiy, V.M. (2015). PARP1 Inhibitors: antitumor drug design. Acta Naturae *7*, 27-37.

Mautner, V.F., Baser, M.E., and Kluwe, L. (1996). Phenotypic variability in two families with novel splice-site and frameshift NF2 mutations. Hum Genet *98*, 203-206.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F.*, et al.* (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55-62.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., and Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome biology *15*, R19-R19.

Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med *8*, 33.

Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.-W., Zhang, Q., McLellan, M.D.*, et al.* (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nature Genetics.

Nyström-Lahti, M., Holmberg, M., Fidalgo, P., Salovaara, R., de la Chapelle, a., Jiricny, J., and Peltomäki, P. (1999). Missense and nonsense mutations in codon 659 of MLH1 cause aberrant splicing of messenger RNA in HNPCC kindreds. Genes, chromosomes & cancer *26*, 372-375.

Okeyo-Owuor, T., White, B.S., Chatrikhi, R., Mohan, D.R., Kim, S., Griffith, M., Ding, L., Ketkar-Kulkarni, S., Hundal, J., Laird, K.M.*, et al.* (2015). U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. Leukemia *29*, 909-917.

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., and Baralle, F.E. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum Mol Genet *12*, 1111-1120.

Papadakis, A.I., Sun, C., Knijnenburg, T.A., Xue, Y., Grernrum, W., Holzel, M., Nijkamp, W., Wessels, L.F., Beijersbergen, R.L., Bernards, R.*, et al.* (2015). SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer. Cell Res *25*, 445-458.

Pena-Llopis, S., Vega-Rubin-de-Celis, S., Liao, A., Leng, N., Pavia-Jimenez, A., Wang, S., Yamasaki, T., Zhrebker, L., Sivanand, S., Spence, P.*, et al.* (2012). BAP1 loss defines a new class of renal cell carcinoma. Nat Genet *44*, 751-759.

Rice, G.I., Reijns, M.A., Coffin, S.R., Forte, G.M., Anderson, B.H., Szynkiewicz, M., Gornall, H., Gent, D., Leitch, A., Botella, M.P.*, et al.* (2013). Synonymous mutations in RNASEH2A create cryptic splice sites impairing RNase H2 enzyme function in Aicardi-Goutieres syndrome. Hum Mutat *34*, 1066-1070.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., Deluca, D.S., Fromer, M.*, et al.* (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*.

Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature reviews Genetics *12*, 683-691.

Sebestyen, E., Singh, B., Minana, B., Pages, A., Mateo, F., Pujana, M.A., Valcarcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res *26*, 732-744.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nature Genetics.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T.V. (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. European journal of human genetics : EJHG *3*, 1-7.

Supek, F., Minana, B., Valcarcel, J., Gabaldon, T., and Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324-1335.

Svenson, I.K., Ashley-Koch, A.E., Pericak-Vance, M.A., and Marchuk, D.A. (2001). A second leaky splice-site mutation in the spastin gene. Am J Hum Genet *69*, 1407-1409.

Taimoor I Sheikh, K.M., Mary J Willis and John B Vincent (2013). A synonymous change, p.Gly16Gly in MECP2 Exon 1, causes a cryptic splice event in a Rett syndrome patient. Orphanet Journal of Rare Diseases *8*.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J.*, et al.* (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum Mutat *29*, 1412-1424.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M.*, et al.* (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol *18*, 1009-1021.

Venables, J.P. (2004). Aberrant and Alternative Splicing in Cancer. Cancer research *64*, 7647-7654.

Vezain, M., Gerard, B., Drunat, S., Funalot, B., Fehrenbach, S., N'Guyen-Viet, V., Vallat, J.M., Frebourg, T., Tosi, M., Martins, A.*, et al.* (2011). A leaky splicing mutation affecting SMN1 exon 7 inclusion explains an unexpected mild case of spinal muscular atrophy. Hum Mutat *32*, 989-994.

Vreeswijk, M.P., and van der Klift, H.M. (2012). Analysis and interpretation of RNA splicing alterations in genes involved in genetic disorders. Methods Mol Biol *867*, 49-63.

Wadt, K., Choi, J., Chung, J.Y., Kiilgaard, J., Heegaard, S., Drzewiecki, K.T., Trent, J.M., Hewitt, S.M., Hayward, N.K., Gerdes, A.M.*, et al.* (2012). A cryptic BAP1 splice mutation in a family with uveal and cutaneous melanoma, and paraganglioma. Pigment Cell Melanoma Res *25*, 815-818.

Woolfe, A., Mullikin, J.C., and Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. Genome biology *11*, R20-R20.

Xie, S., Mortusewicz, O., Ma, H.T., Herr, P., Poon, R.Y., Helleday, T., and Qian, C. (2015). Timeless Interacts with PARP-1 to Promote Homologous Recombination Repair. Mol Cell *60*, 163-176.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology *11*, 377-394.

Zhang, K., Nowak, I., Rushlow, D., Gallie, B.L., and Lohmann, D.R. (2008). Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression. Hum Mutat *29*, 475-484.

Chapter 4: Rare germline and somatic splice-site-creating mutations disrupt exon definition in cancer genomes

Contribution: I developed the entire code for translation of MiSplice variants. I manually reviewed many variants validated by our computational tool, performed all downstream analyses, wrote the entire text, created all Figures, and performed entire mini-gene splicing assay for 7 variants tested.

4.1 Introduction

Mutations can be germline or somatic in nature. Germline mutations are inherited from parents or acquired *de novo*. Somatic mutations are acquired throughout an organism's lifetime in individual cells due to genetic and environmental factors such as chemicals and radiation. Most of the damage in the DNA is repaired, but sometimes the alterations are fixed. Mutations in genic and regulatory regions can affect gene function in several ways by causing loss or gain of function, altering transcript splicing, and increasing or decreasing gene expression level. In combination, such variations can disrupt normal gene function and alter cellular response to regulation giving the cell a selective advantage to proliferate autonomously.

The underlying genomic architecture of individual species alters the resulting phenotypic variability. Understanding how differences to gene architecture can alter a genes function is essential to understanding many biological questions including disease diagnosis, progression and treatment. Current studies suggest that between 70-95% of human genes harbor multiple mRNA transcripts (Johnson et al.; Matlin et al.; Modrek and Lee). Different mRNA transcripts are created by the process of alternative splicing which expands the complexity and information content of the eukaryotic genome and allows for tissue, developmental, or temporally expressed isoforms which can perform alternate functions. Understanding how genomic variation in tumors can contribute to alternative splicing defects is an actively expanding field.

For the past decade, cancer genomics studies have focused on identifying and validating germline and somatic mutations by comparing patients tumor samples to their normal tissue. The integration of transcriptomic and proteomic data provides valuable insight as to the biological factors contributing to the cancer genome (Cieslik and Chinnaiyan). A database to explore splice-junction usage across TCGA (Sun et al.) as well as a global analysis of splicing across the TCGA cohort (Kahles et al.) were recently published expanding the current paradigm about the exceptions and rules by which splicing is dysregulated in cancer. While this recent publication analyzed the global landscape of alternative splicing across the TCGA pan-cancer cohort, they limited their analysis to germline mutations that are known to also be somatic in origin, greatly limiting their prediction power (Kahles et al.).

Earlier this year, we published a bioinformatic tool that systematically predicts somatic splice-site-creating mutations across 33 cancer types (Jayasinghe et al.). We identified nearly 2,000 somatic splice-site-creating mutations with 26% and 11% of mutations being initially mis-classified as missense and silent mutations, respectively. This novel finding lead us to question the burden of similar mutations in the germline of cancer patients, as this is an underexplored question in the field. Recently, The Cancer Genome Atlas consortium published a curated mutation profile of germline (Huang et al.) and somatic (Bailey et al.; Ellrott et al.) mutations across cancer genomes. These timely and thorough scientific publications have provided a sound basis for exploring the splice-site-creating landscape in both a somatic and germline context.

Here we developed additional modules for MiSplice to facilitate accurate identification of thousands of rare germline and somatic splice-site-creating events while providing necessary supplemental scripts to infer the resulting transcriptional and translational effects. Using our stringent pipeline we identified 2,888 rare germline and somatic SCMs which were initially inaccurately annotated. We observed interesting patterns in genomic sequence contexts surrounding SCM containing exons (SCM+ exons) including increased nucleosome occupancy and an overall decrease in the novel splice form relative to the original exon size. The discovery of rgSCMs genome wide facilitated the proper annotation of hundreds of variants that could have direct implications in cancer. For example, one recurrent synonymous variant was identified in *RAD54L,* generally characterized as a DNA repair gene, suggesting this variant could be a novel pathogenic germline variant that has been missed in previous studies. Finally, we validated 6 rare germline SCMs (rgSCMs) in cancer associated genes, including the *RAD54L* silent variant, and one common germline SCM (cgSCM) validating our computational pipeline and workflow for proper and efficient SCM identification.

4.2 Results

Dataset Description

To further explore the germline and somatic splice-site-creating landscape across cancer samples, we collected somatic and germline mutations calls from the MC3 working group (Ellrott et al.) and the Cancer Genome Atlas (TCGA) germline working group (Huang et al.). We sought to evaluate the splice-creating potential of variants conventionally annotated as SpliceDonorSNV, SpliceAcceptorSNV, Missense, Synonymous, IntronicSNV (near the splice site) and Nonsense mutations (Figure 4.1A). With stringent filtering (Methods) we evaluated >3,000,000 somatic mutations, > 10,000,000 rare germline mutations and >20,000,000 common germline variants in a splicing context using the MiSplice pipeline (Jayasinghe et al.) (Figure 4.1C). For rare germline mutations and somatic mutations, we defined a set of control samples in the same cancer type that lacked the same mutation of interest. With respect to common germline mutations, samples were placed in a case group or control group depending on their mutation status for each mutation of interest. We conducted the splice-site-creating mutation analysis using the ISB google cloud using MiSplice (mutation-induced splicing) that systematically evaluates mutations in a splicing context using RNA-seq data. MiSplice manages large analyses using parallel computation to search for alternative splice junctions within windows of ±20 bp from the mutation of interest. We identified high confidence rare and common germline splice-site-creating mutations (rgSCMs, cgSCMs) and somatic splice-site-creating mutations (sSCMs).

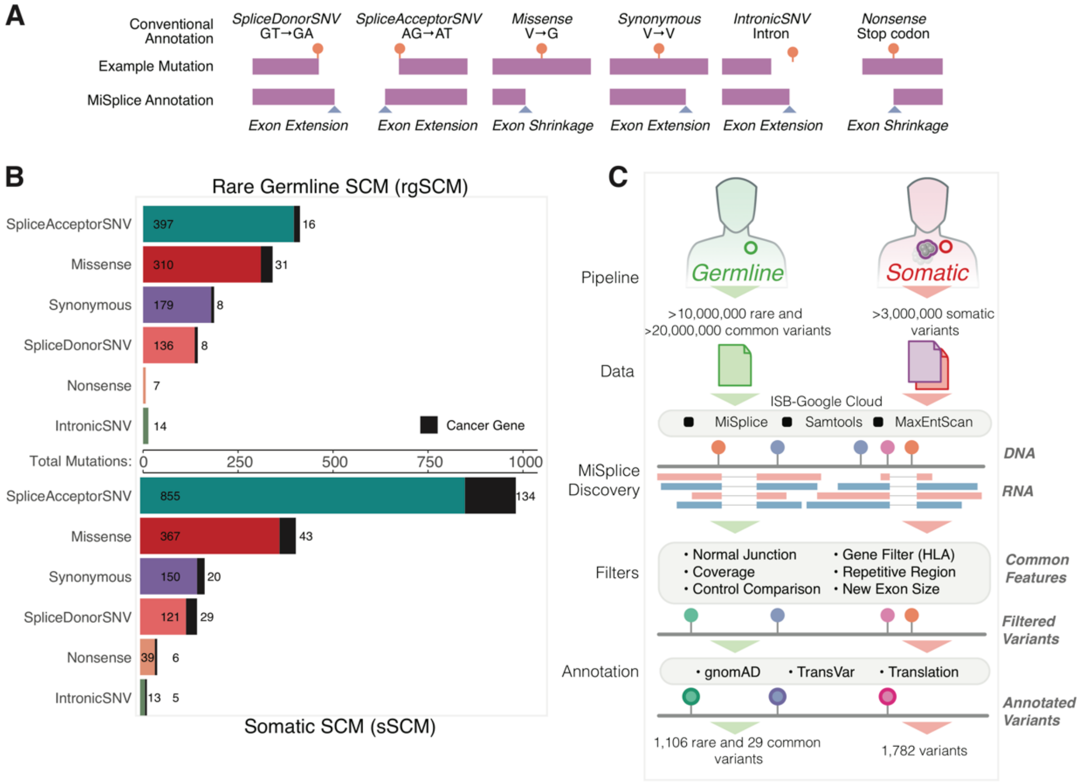


Figure 4.1. Somatic and Germline splice-site-creating variant discovery. (A) Examples of splice-site-creating mutations (SCM) for different annotated mutation types. (B) Breakdown of rare germline SCMs and somatic SCMs, plotted by conventional annotation type. Cancer associated genes are highlighted in black. (C) The updated MiSplice workflow can process millions of somatic and germline variants on the ISB-Google Cloud or on a local compute cluster. Mutation calls and RNA-Sequencing bams are provided as input, and variants in close proximity to a detected splice junction are maintained for additional filtering. Canonical junctions, low coverage, novel junction presence in the controls, large predicted exon sizes and certain genes including HLA are filtered out. Finally an added annotation

module determines allele frequency for each variant in the gnomAD database, predicts the mutation effect using TransVar and predicts the resulting wildtype and novel protein isoforms

GERMLINE AND SOMATIC SPLICE-SITE-CREATING MUTATION DISCOVERY

The expanded discovery of somatic splice-site-creating mutations across 33 cancer types identified 1,782 sSCMs (Figure 4.1B). 237 SCMs are annotated to genes commonly associated with either adult or pediatric cancer. Some of the most highly recurrent events were reported in our previous publication (Jayasinghe et al.) but novel sSCMs were identified by expanding our analysis to this larger MC3 dataset (Table 1). Newly identified sSCMs in cancer associated genes were manually reviewed for accuracy.

In expanding our analysis of splice-site-creating mutations to the germline, we identified mis-annotated germline mutations that show evidence of creating new splice sites. After processing rare variants from 33 cancer types a total of 14,709 cancer specific unique variants were identified as having splice-creating-potential. Additional filters included removing sites with more than 5% of controls containing the alternative splicing event (>2 reads), samples with less than 20 controls, and finally filtered out splicing events identified in highly homologous genes including MUC*, AHNAK* and CRIPAK (Methods) leaving 4,295 potential splice-site-creating variants. Of the 4,295 variants, 1,121 variants had a maximum allele frequency across populations in the controls subset less than 0.05% as derived from gnomAD browser - version 2.1 (Lek et al.). Finally after peptide translation and additional filtering, we were left with 1,106 high confidence single nucleotide variants classified as rgSCM events encompassing 995 unique variants from 852 genes, and 63 in cancer associated genes.

Additional filters were incorporated into MiSplice to filter out novel splice isoforms present at a higher frequency in the cancer cohort population. Finally all variants underwent additional annotation post filtering to capture filter out events present at a high frequency in the population, and we incorporated TransVar to provide the most up to date annotation and predict the resulting peptide change. For the entirety of this analysis we will focus on single nucleotide variants.

A majority of SCMs were conventionally annotated as SpliceAcceptorSNVs followed by Missense, Synonymous, SpliceDonorSNVs, Nonsense and a handful IntronicSNVs (Figure 4.1B). This distribution held true for both rare germline and somatic events. Interestingly, our expanded analysis identified 170 novel rare germline SCMs conventionally annotated as synonymous variants and an additional X somatic SCMs not identified in our previous work. A remarkable number of SCMs were annotated to cancer associated genes (Table 1) including 63 rare germline SCMs and 237 somatic SCMs. We will take a closer look at the biological relevance of these events in a later section.

COMPARATIVE SEQUENCE CONTEXTS AND CHARACTERISTICS

With this large and high confidence list of somatic and germline splice-site-creating mutations across thousands of exons, we have the unique opportunity to collect relevant genomic information to characterize SCM positive exons. First we looked to determine the location of the novel splice junction relative to the canonical splice site (Figure 4.2A). In knowing that the strongest splicing enhancer elements are located near the canonical splice site, we predicted many SCMs would cluster near the canonical site and taper off with increasing distance from the canonical site. As expected splice acceptor and splice donor variants are densely populated near the canonical splice site while synonymous and missense mutations are found deeper within the exonic region. After the mutation is introduced, the largest differences in transcript size post mutation (here on in referred to as effect size) are shared by conventionally annotated missense, synonymous and nonsense mutations, while smaller alterations are observed for mutations in close proximity to the splice site (Supplementary Figure 4.1A). Due to our restricted search window for identifying novel splice sites in proximity to mutations, we anticipated and gladly observed a direct correlation between the effect size and distance from the mutation to the canonical splice site (Supplementary Figure 1B).

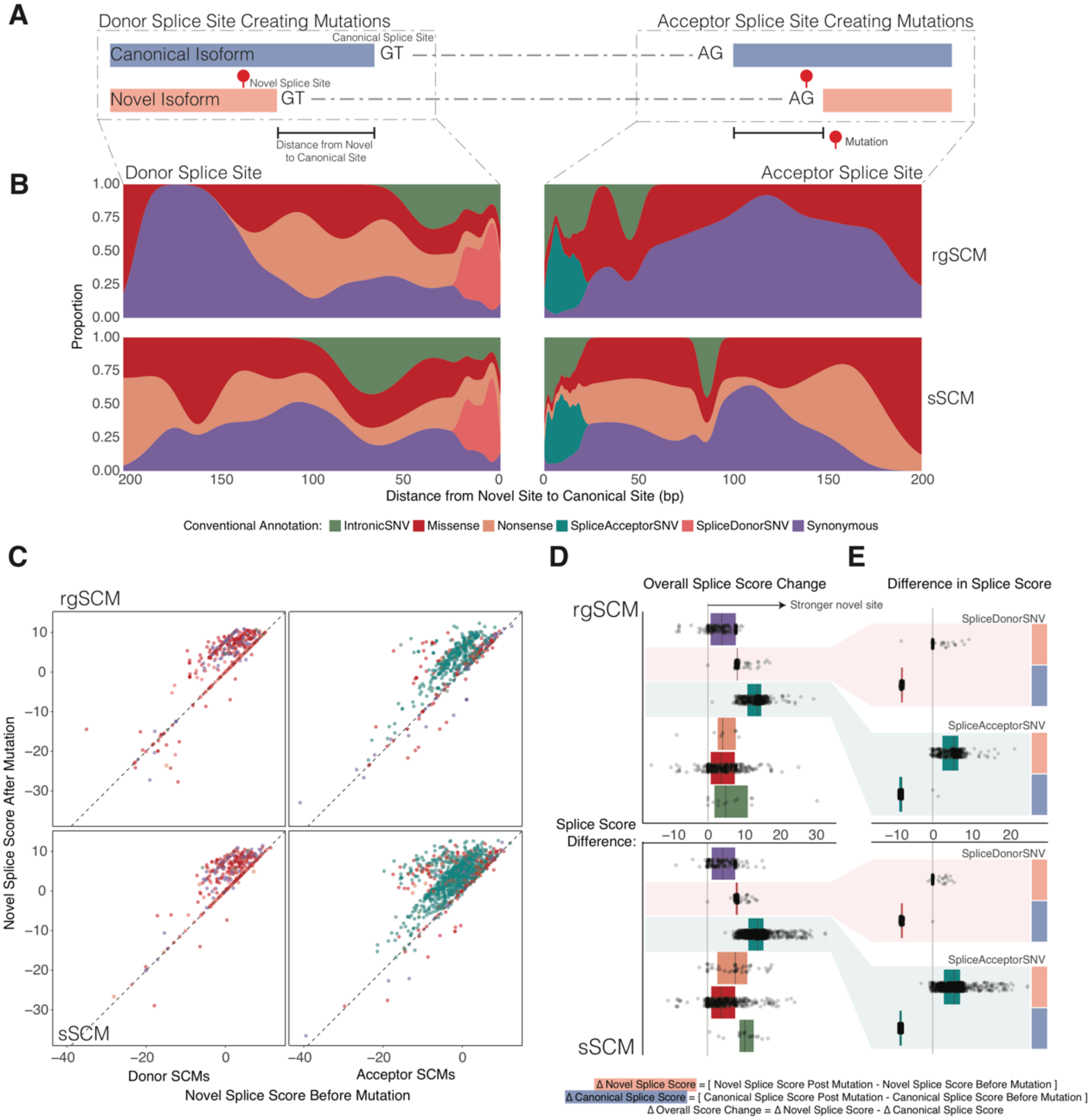


Figure 4.2. Sequence contexts and characteristics of splice-site-creating mutations. (A) Cartoon of donor and acceptor splice-site-creating mutations (SCM). Example mutation is plotted in red relative to the novel splice site and canonical splice site. (B) Distribution of rgSCM and sSCM events relative to the distance from the canoncial splice site. Proportion of variants at each position is colored by conventional annotation type. (C) Comparison of the novel splice site score before and after mutation for rgSCM and sSCM broken down by the creation of a novel donor or acceptor splice site. Each point is colored by conventional annotation type. (D) Contrast the overall change in splice score before and after mutation for the novel and canoncial junctions. Positive values indicate a stronger novel splice score

change post mutation relative to canonical splice site. (E) For variants conventionally annotated as SpliceDonorSNV and SpliceAcceptorSNV, compare the difference in splice score pre and post mutation for the novel and canonical splice site.

To confirm the mutation is indeed a splice-site-creating mutation, we compared the novel splice score of the genomic sequence before and after mutagenesis (Figure 4.2A). Rare germline and somatic SCMs alike globally exhibited an increase in the novel splice score after mutation (deviating in the positive direction above the xy line), with few exceptions (Figure 4.2C). While strengthening a novel or cryptic splice site is necessary to induce alternative isoform usage, we wanted to determine if the canonical splice site was also disrupted due to the mutation. To characterize an overall splice score change, the difference in the splice score was calculated for both the novel (ΔN) and canonical (ΔC) splice site. The difference between ΔN-ΔC would be positive under conditions where the novel splice score change post mutation was stronger than the canonical site in the presence of the mutation, capturing the overall mutations effect on both sites. Figure 4.2D displays the distribution of the overall combined splice score change for each site by conventional annotation type. While all values trend in the positive direction, suggesting a stronger novel site change relative to canonical site change, conventionally annotated splice donor variants are densely clustered together. By taking a step back we can compare ΔN and ΔC separately. In Figure 4.2E, the difference in the novel score (ΔN) increases dramatically, while the canonical site (ΔC) is completely disrupted. On the other hand, for the donor site, the novel splice site doesn’t change in the presence of the mutation, but the canonical site is disrupted.

This interesting phenomenon for the donor splice site holds true for the 294 conventionally annotated donor mutations and suggests that by simply disrupting the donor splice site, it's usage can be completely ablated and becomes unrecognizable, placing a stronger emphasis on the donor site relative to the acceptor splice site. Supporting this hypothesis, recently Wong et al. performed a massively parallel splicing assay on the introns of three genes to monitor the effects of all predicted changes to the 9 nucleotide 5' splice site, surveying a total of 32,768 unique donor sites (Wong et al.). From the assay, they determined disruption of the 5' splice site alone was enough to disrupt splicing, despite mutating nearby genomic regions.

The average exon in our dataset containing a splice-site-creating mutation is 289 bp long with a standard deviation of 668 bp. By comparing the relative exon sizes between each of the conventionally annotated mutation types, an increased canonical exon size was observed for missense (p=2.1e-08,1.9e-13) and synonymous sites (2.9e-05,9.8e-11) in comparison to the acceptor and donor splice sites, respectively. Nonsense mutations also showed evidence of increased exon size relative to the donor splice site (p=0.033) (Supplementary Figure 1A).

Next we chose to explore beyond the exon to determine if introns adjacent to the SCM positive exon contained additional information for SCM classification. Interestingly, splice acceptor mutations were more likely to be situated downstream of a larger intron relative to the intron upstream of the synonymous mutation (p=5.4e-05, Figure 4.1C). Similarly the downstream intron size was larger for splice acceptor SNVs relative to introns downstream of missense (p=3.4e-05), synonymous (p=2.5e-09) and splice donor mutations (Supplementary Figure 1D). These initial observations suggest that while evaluating the splicing score in the presence and absence of the mutation is very informative, with a large enough dataset, we may be able to glean new information from the adjacent genomic sequences to strengthen the identification of putative SCM containing exons.

TRANSLATIONAL AND TRANSCRIPTIONAL IMPLICATIONS OF SCMS

Each variant was annotated according to their resulting translational consequence to interrogate overall trends in splicing dysregulation (Methods). For both somatic and rare germline SCMs, exon shrinkage events were observed at a higher frequency than exon extension. 91% and 89% of translated sites resulted in exon shrinkage trending towards an overall decrease in the novel exon size relative to the canonical exon (Figure 4.3A). The novel exon size decreased by an average of 21 and 33 bp for sSCM and rgSCM, respectively. From an evolutionary standpoint, exon size tends to be far more constrained than intron size. One hypothesis for this constraint is due to the intricate dance between chromatin remodeling and alternative splicing.

Within the nucleus, DNA is wrapped around histones in segments of <150 nucleotides. Nucleosomes are made up of eight core histones and are conveniently positioned on exons. The size of the average exon is ~150 nucleotides long providing strong evidence as to the evolutionary pressure selecting for exons around this length(Amit et al.;

Schwartz et al.). Since splicing is a co-transcriptional process, splicing factors and histone marks work together to properly facilitate the splicing process(Jeong). With the aforementioned evidence suggesting a link between nucleosome occupancy and splicing, we performed a nucleosome occupancy enrichment analysis for the rgSCM and sSCMs using NucMap (Zhao et al.). The enrichment scores were much stronger at regions overlapping the exons containing splice-site-creating mutations relative to a 2kb window surrounding the exon of interest (Figure 4.3B). Additional studies have revealed stronger nucleosome enrichment for exons that are adjacent to long introns or have weaker splice sites (Spies et al.). To explore this phenomenon in our dataset, SCMs were binned by the canonical splice site score, and nucleosome enrichment was compared among the four binned groups by quartile, but no correlation was observed in our dataset (Supplemental Figure S7).

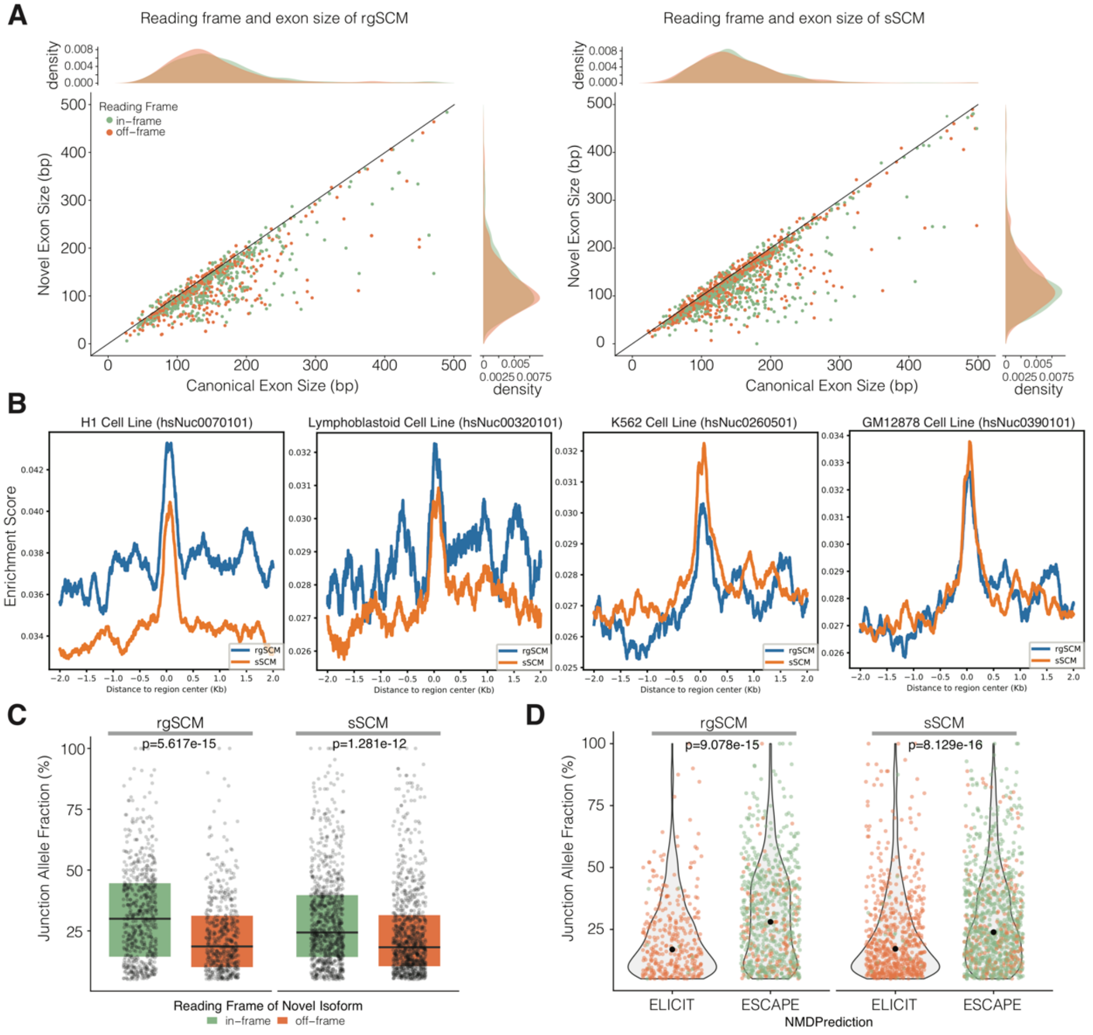


Figure 4.3. Novel spliced isoform exon expression and predicted translation. (A) Contrasting overall novel exon size relative to canonical exon size for rgSCM and sSCM events. Each point is colored by the predicted reading frame of the novel spliced isoform. (B) Enrichment analysis of nucleosome occupancy by NucMap on four different cell lines for all 2,888 predicted SCM containing-exons. Each plot is centered on the exon containing the novel splice-site-creating mutation. (C) Junction allele fraction (JAF) distribution by predicted reading frame. P-value reported by wilcox test comparing in frame and off-frame events within each SCM group. (D) Based on the location of the premature termination codon, each predicted protein product was classified as eliciting or escaping nonsense mediated decay based

on the 50 bp rule. P-value reported by wilcox test comparing both groups within each SCM group. Each point is colored by the predicted reading frame of the novel spliced isoform.

The resulting reading frame of the novel splice isoform was annotated as off-frame or in-frame based on the size of the exon shrinkage or extension event. As expected, the junction allele fraction, or fraction of reads supporting the alternative splice form, was significantly lower for predicted off-frame SCMs compared to in-frame SCMs (wilcox test, sSCM:*p=1.281e-12*, rgSCM:*p=5.617e-15*) (Figure 4.3C). The expression of the resulting off-frame events is predicted to be lower at the RNA level because aberrantly spliced transcripts are often degraded by nonsense mediated decay (NMD), a process that identifies and degrades transcripts containing premature termination codons (PTCs). The general rule of thumb is PTCs located at least 50 bp upstream of the last exon-exon junction drive strong degradation, whereas those outside of this criteria are predicted to escape the degradation process(Brogna and Wen; Lewis et al.; Maquat et al.; Nagy and Maquat; Popp and Maquat; Venables; Weischenfeldt et al.). When applying the 50 bp rule to to our set, 97% of the off-frame spliced products are predicted to elicit-NMD (990 off-frame, 33 in-frame) while 79% of in-frame events are expected to escape-NMD (1465 in-frame, 400 off-frame) (Supplemental Figure S7). For escapee transcripts, we would expect to observe a higher expression of the novel splice form. Indeed, expression as measured by the junction allele fraction supported NMD-escaping transcripts maintaining a higher expression relative to the non-escapee group (wilcox, rgSCM:*p=9.078e-15*, sSCM:*p=8.129e-16*) (Figure 4.3D).

Even though some sites are predicted to undergo degradation, all SCMs were identified due to their expression in the RNA-Sequencing data. Most sites have JAF's lower than 50% suggesting some post transcriptional mechanisms by which the transcript is being degraded, although not to completion. It is important to note that it is very likely we are not capturing the full landscape of SCMs due to efficient degradation of mutation induced alternative transcripts.

Landscape of somatic and germline SCMs genome-wide

We next explored the landscape of SCMs genome-wide. Figure 4.4A highlights the most highly recurrent genes in our dataset with *TP53* harboring somatic SCMs spanning several cancer types while *CBWD5* has many SCMs predominantly derived from BRCA. For the subset of highly recurrent SCM containing genes, rgSCM and sSCM almost always do not co-occur in the same gene with the exception of: *LZTR1* and *PTPN13*. Out of 2,163 unique genes harboring at least one SCM, 161 genes encompassing 417 variants had both a germline and somatic SCM reported. Several shared SCM positive genes overlap cancer associated genes including *CHEK2, CHD8, CASP8, FANCL, MUTYH, RAD51C, RPA1, BRCA1, EML4, FANCI, PARP3*, and *PARP4*. Interestingly, only *CHEK2* shared a mutation at the same splice donor site, but both events produced different transcriptional effects. Additionally, a conventionally annotated missense variant in *MLLT10* contained both an rgSCM and sSCM resulting in the same 251 bp exon shrinkage, and was validated in our minigene splicing assay (Figure 4.5D).

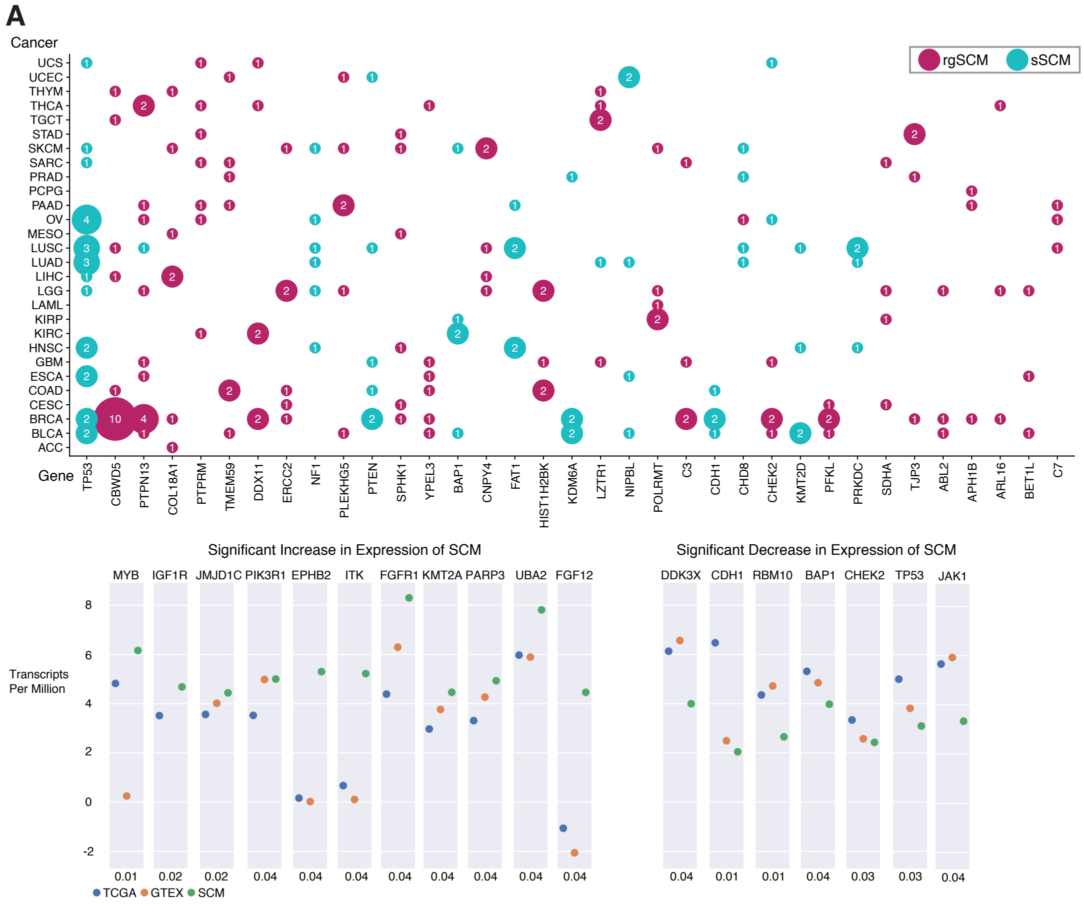


Figure 4.4. Splice-site-creating variants across genes and cancer types. (A) Overview of most highly recurrent splice-site-creating mutations colored by SCM type (germline or somatic). Values in plot indicate total number of SCMs identified for each gene-cancer type pair. (B) Overall change in gene expression (TPM) between SCMs and their cancer type cohort (wilcox test). Pvalue is indicated at the bottom of each plot and the average of each group are plotted as one value. Grouped by change in expression with a significant increase in expression being on the left and a significant decrease on the right.

Next we wanted to explore changes in gene expression relative to SCM presence. Using gene expression data from the UCSC XenaBrowser, we compared transcripts per million (tpm) values between the SCM mutants, TCGA Cancer Cohort and associated GTEX tissue cohort, when available. By comparing the expression of the SCM mutant to the cancer type cohort, several SCMs exhibited differential expression including those in cancer associated genes (Figure 4.4B) (Wilcox test). The biological implications of the altered expression will need to formally evaluated in follow up studies.

HYPERMUTATOR PHENOTYPE OF RAD54L RGSCM

RAD54L is involved in DNA repair via homologous recombination and frequently undergoes copy number alterations depending on the tissue type of interest. For example, loss of heterozygosity in *RAD54L* is common in breast cancer and lower-grade gliomas(Nowacka-Zawisza et al.) and associates with longer progression free survival and chemosensitivity(Tang et al.), respectively, suggesting a tumor suppressor phenotype. More recently, the mechanism of cell proliferation and radio-resistance in glioblastoma was linked to CDC7 expression, a gene directly regulated by RAD54L(Li et al.). In contrast, in choroid plexus carcinomas the DNA repair gene was often amplified in tumors and necessary for proliferation (Tong et al.) in murine models, suggesting an oncogenic phenotype.

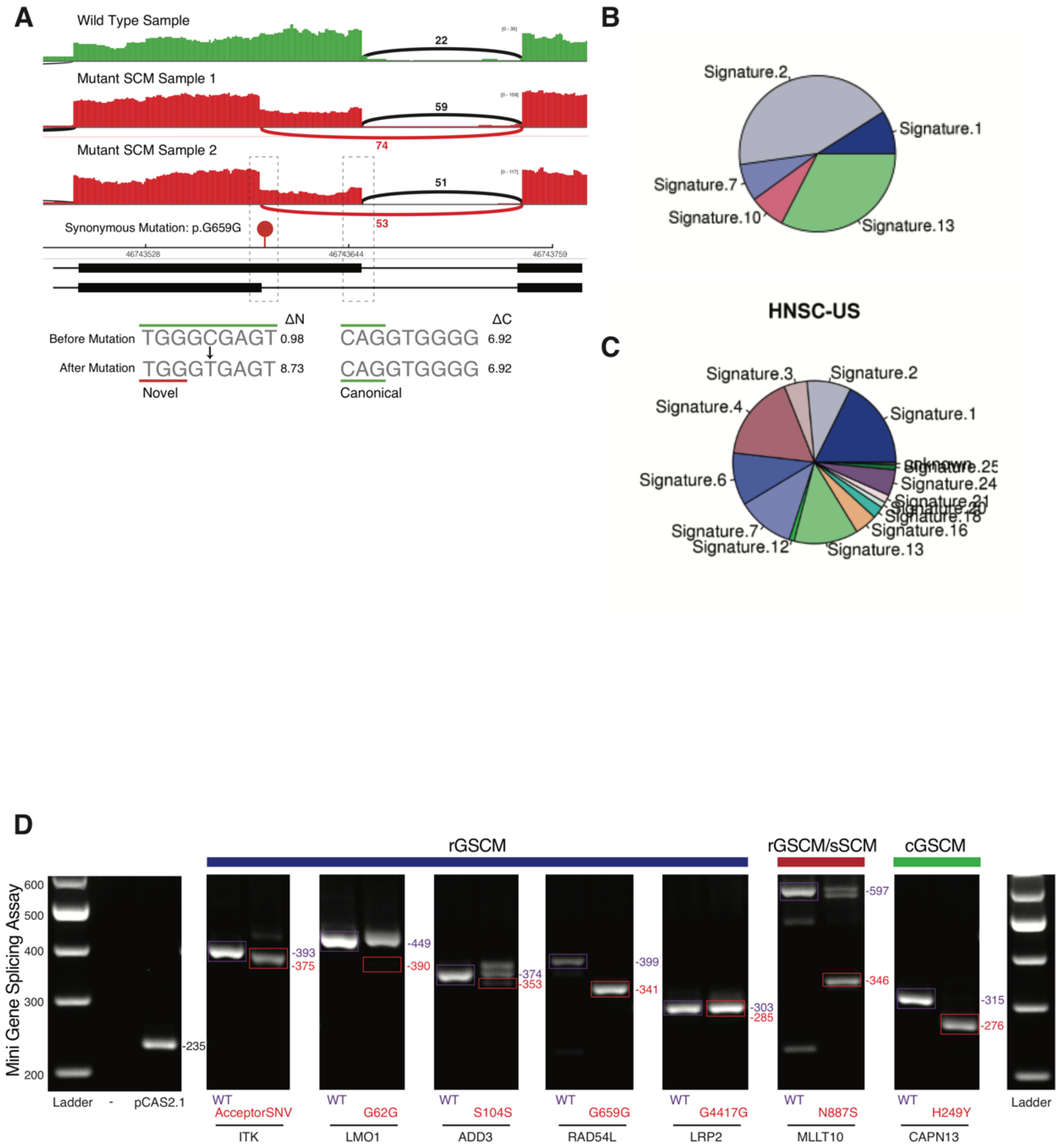


Figure 4.5. RAD54L and mini-gene splicing assay. (A) Splice junction quantification of te novel and canonical splice sites for two samples containing the RAD54L silent mutation and one normal sample. Genomic sequence of the novel and canonical splice site are highted below the exon pictogram. The novel change in novel and canonical splice score are indicated next to the genomic segments. (B) Mutational signature for RAD54L HNSC Mutant sample. (C) Mutational signature for overall HNSC cohort (D) Candidate splice creating variants validated by the mini-gene splicing assay. Exons of interest were

cloned into the pCAS2.1 vector and mutant (red) and wildtype (purple) plamids were transfected into 293T cells and total RNA was extracted to identify mutation induced alternatively spliced products.

In our sample set two conventionally annotated silent (p.G659G) mutations in *RAD54L* were identified more appropriately as an rgSCM and associated with a slightly higher overall expression relative to the head and neck squamous cell carcinoma (HNSC) cohort (wilcox p=0.099) but not in the testicular germ cell tumor (TGCT) (wilcox p=0.343). The rare germline mutation creates a new splice site (GC>GT) leading to a 58 bp exon shrinkage, frameshifting the remainder of the protein at exon 17 containing a Snf2 specific helical domain (HD2) domain and the C terminal Domain (CTD) (Heyer et al.) (Figure 4.5A). The novel splice junction is located within 55 bp of the last exon-exon junction and is thus predicted to escape NMD. The splicing score post-mutation increases dramatically from 0.98 to 8.73 and the variant allele fractions in the tumor and normal sample are comparable, 49% and 46% for the HNSC sample and 44% and 52% in the TGCT sample, respectively.

We next sought to evaluate whether the predicted rgSCM event in *RAD54L* can induce genomic changes relative to the associated cancer cohort. Both samples exhibit higher mutation rate relative to the rest of their cancer type cohort, with the TGCT sample being the highest mutated sample and the HNSC sample being one of the top 10 mutated samples. To evaluate whether the potentially pathogenic rare germline SCM mimics a hypermutator like phenotype, we evaluated the mutational signatures of each patient. After inspecting the mutational signature of the HNSC sample, signature 2 and signature 13 are the most prevalent followed by 1, 7 and 10 (Methods). In contrast, the HNSC cancer type cohort did not exhibit a similar mutation signature, suggesting this sample is under a different evolutionary pressure relative to the cancer cohort. The joint signatures of 2, 10 and 13 in the HNSC sample are commonly found in patients undergoing local hypermutator phenotypes and with this sample being one of the 5 top mutated samples in HNSC (lacking mutations in other DNA-repair related genes) (Alexandrov et al.) suggests the *RAD54L* rgSCM variant may be indicative of the observed phenotype.

Finally, to functionally confirm the predicted alternatively spliced product in the lab, we performed a mini-gene splicing assay to validate rgSCMs in 6 cancer-associated genes and 1 common germline SCM (Figure 4.5D). Briefly, we used a pCAS2.1 splicing reporter mini-gene functional assay by cloning mutant exons into the pCAS2.1 vector (Gaildrat et al.) and transiently transfected into HEK293T cells. RNA is extracted after 24 hours and after generating cDNA synthesis, RT-PCR is performed to evaluate the presence of the novel splice isoform relative to the wild-type vector. *RAD54L* rgSCM was one of the seven germline SCMs confirmed in the assay (Figure 4.5D). Taken together, MiSplice was able to predict a recurrent putative rare germline SCM event improperly annotated as a synonymous mutation. While mutations disrupting DNA repair have been identified in another member of the RAD family *RAD51D* (Pelttari et al.), to date no pathogenic mutations have been identified in *RAD54L*.

4.4 Methods

DATASET DESCRIPTION

Aligned RNA-Seq bam files were analyzed using the ISB google. These cancer types are Acute Myeloid Leukemia [LAML], Adrenocortical carcinoma [ACC], Bladder Urothelial Carcinoma [BLCA], Brain Lower Grade Glioma [LGG], Breast invasive carcinoma [BRCA], Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], Cholangiocarcinoma [CHOL], Colon adenocarcinoma [COAD], Esophageal carcinoma [ESCA], Glioblastoma multiforme [GBM], Head and Neck squamous cell carcinoma [HNSC], Kidney Chromophobe [KICH], Kidney renal clear cell carcinoma [KIRC], Kidney renal papillary cell carcinoma [KIRP], Liver hepatocellular carcinoma [LIHC], Lung adenocarcinoma [LUAD], Lung squamous cell carcinoma [LUSC], Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC], Mesothelioma [MESO], Ovarian serous cystadenocarcinoma [OV], Pancreatic adenocarcinoma [PAAD], Pheochromocytoma and Paraganglioma [PCPG], Prostate adenocarcinoma [PRAD], Rectum adenocarcinoma [READ], Sarcoma [SARC], Skin Cutaneous Melanoma [SKCM], Stomach adenocarcinoma [STAD], Testicular Germ Cell Tumors [TGCT], Thymoma [THYM], Thyroid carcinoma [THCA], Uterine Carcinosarcoma [UCS], Uterine Corpus Endometrial Carcinoma [UCEC], Uveal Melanoma [UVM].

MUTATION FILE DESCRIPTION

To evaluate somatic mutations we started with the mc3.v0.2.8.CONTROLLED.maf and filtered out mutations with the following MAF Filter Flags: StrandBias, pcadontuse, common_in_exac, oxog, contest, nonpreferredpair, ndp, badseq, broad_PoN_v2, leaving a total of 10,573,839 variants. With respect to mutation callers, single nucleotide polymorphisms and insertions and deletions were only maintained if there was agreement from two or more callers. Finally mutations with a variant allele fraction of less than 5%

were filtered out. Description of MC3 MAF file and filter flags found here: https://www.synapse.org/#!Synapse:syn7214402/wiki/405297. Final filtered somatic mutation file used for analysis found here: https://www.synapse.org/#!Synapse:syn12074532.

To evaluate germline mutations, individual VCFs were downloaded from Huang et al., 2018 (Huang et al.) encompassing TCGA cancer types listed in the Dataset Description. All variants were annotated using steps 6 and 7 of germline wrapper (https://github.com/ding-lab/germlinewrapper) against hg19 Homo Sapiens fasta file. After annotation, variants were maintained in a common germline MAF if allele frequency is greater than 0.01 and filtered into a rare MAF if allele frequency is less than 0.01 or if not reported. Common variants were only evaluated for BRCA. All rare variants were evaluated for splice-site-creating function using MiSplice. Furthermore, total reads supporting the alternatively spliced product were lowered for genes identified as significant in Huang et al., 2018 (Huang et al.) and re-run through MiSplice. After filtering, any variants with more than 10% of controls samples containing the alternatively spliced product were further filtered out leaving high confidence variants. For the 20,205,168 common variants in BRCA, in an initial screen only 72,245 unique variants were evaluated for splice-creating function. Of the 72,245 unique variants, 287 were shown to have splice-creating function in at least one sample. The variants with splice-creating function were then evaluated in the remaining samples. In total 74,383 variants were evaluated for splice-creating function. 29 variants with less than 10% of control samples exhibiting the alternatively spliced isoform were manually reviewed, resulting in 6 high confidence common germline splice altering variants.

Finally germline mutations were finally filtered as being rare if the allele frequency derived from the Genome Aggregation Database (gnomAD browser; http://gnomad.broadinstitute.org/; release 2.0.2) is less than 0.0005 AF (0.05%).

MISPLICE PIPELINE (SPLICE-SITE-CREATING MUTATION)

The MiSplice pipeline was developed to detect mutation-induced splicing events from RNA-Seq data. It is written in Perl and incorporates two standard tools, samtools and MaxEntScan. The pipeline is fully automated and can run multiple jobs in parallel on LSF cluster. It executes the following steps:

1) Splitting large maf file into multiple smaller files with less mutations (currently, the default setting is 200).

2) Discovering splicing junctions within 20bps of the mutation with at least 5 supporting reads with mapping quality Q20 and then filtering canonical junctions by using the Ensembl 37.75 database. We selected 20bp as a cut-off since it is the farthest distance from the splice junction in a splice region.

3) Computing the number of supporting reads of above cryptic splice sites for control samples without mutations.

4) Calculating the splicing scores for the cryptic splice sites via MaxEntScan.

5) Reporting the depth of each cryptic splice site via Samtools.

6) Filtering cryptic sites which fall in HLA loci or less than 5% of reads at the genomic location supporting the novel junction of interest.

MISPLICE GERMLINE FILTERING

All relevant scripts for filtering germline splice-site-creating mutations can be found: https://github.com/reykajayasinghe/MiSplice_Supplemental/.

1) Gene filter: Large exons with a high affinity for alternatively spliced products were filtered out including: AHNAK, AHNAK2, HLA family, FMN2, CRIPAK, IGH*, MUCIN family, RP11*, orf genes (Table X)

2) Remove events with less than 20 supporting controls.

3) Remove sites with greater than 5% of control samples having the same reported alternative splicing event.

4) Nearby Mutations and combining mutations in the same gene by cancer type: Combing the same splice-site-creating event into one entry when there are multiple mutations nearby one another for the same sample set. Combining the same SCM event across multiple patients into the same entry.

5) All sites are annotated with TransVar (Zhou et al.). Only variants with annotated variants in protein_coding transcripts that were classified by uniprot as TSL=1 were maintained for further analysis.

6) Finally, all single nucleotide variants were annotated with population frequencies derived from the genome aggregation database (gnomAD,http://gnomad.broadinstitute.org/).

SPLICE SITE SCORE ESTIMATION

For each cryptic splice site and nearby canonical splice site, the corresponding nucleotide sequences were first extracted for both the mutant and reference sequences (9-mer and 23-mer for donor and acceptor, respectively). Their splice scores as potential donor or acceptor sites were then estimated using MaxEntScan for single nucleotide polymorphisms.

MUTATIONAL SIGNATURE

For determining the mutation signatures of a the subset of samples, mutations were extracted from the MC3 public mutation file for the HNSC sample while mutations were extracted from the controlled mutation file for the TGCT sample (https://www.synapse.org/#!Synapse:syn7214402). Different mutation files were needed because all the mutations for the TGCT sample were removed from the public mutation file due to filter flags mostly considered “NonExonic.” All mutations were run through Mutational Signatures in Cancer (MuSiCa) to evaluate mutation signatures for each patient.

NUCLEOSOME POSITIONING ANALYSIS

For determining the nucleosome positioning of the alternatively spliced exons, all exons containing a rare germline or somatic SCM were collected and converted to bed format.

Coordinates were lifted over to hg38 from hg19 using ucsc hgLiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and coordinates were modified to match the input for NucMap (http://bigd.big.ac.cn/nucmap/Faq.php#q9) to extract enrichment scores of nucleosome binding sites across several cell lines including: hsNuc0270101, hsnuc0260501, hsNuc0390101, hsNuc0320101, hsNuc0070101, and hsNuc0020101. Using the browser analysis software normalized reads (RPM, reads per million) were extracted from the aforementioned cell lines and an enrichment score was calculated using the online software.

MANUAL REVIEW

All novel splice creating mutations were manually reviewed using the integrative genomics viewer (http://software.broadinstitute.org/software/igv/). Mutations were placed into one of three categories: Pass, Complex, and No Support. Mutations were classified as complex if more than one novel alternatively spliced product was observed for the mutated sample. After annotation, sites with large effect size (>100 bp) were manually reviewed for confirmation. Many events with very large new exons were recharacterized as ultra-short-introns.

TRANSLATION

All MiSplice Supplemental scripts are written in python (https://github.com/reykajayasinghe/MiSplice_Supplemental). Translation.py takes in the TransVar annotated matrix and defines important characteristics surrounding the SCM including: determining size of the upstream and downstream intron, size of the current exon, size of the novel splice isoform, overall change in size of the spliced product, mutation position relative to the start and end of the exon, determines the resulting wildtype and mutant amino acid predictions for the SCMs, and finally predicts degradation by NMD. Some sites may be lost during this translation script due to improper annotation by transvar to a protein coding transcript or the lack of a known transcriptional start site in the input coding sequence file.

CODE AVAILABILITY

MiSplice is written in Perl and is freely available from GitHub at https://github.com/ding-lab/misplice_gsSCM under the GNU general public license. MiSplice uses several independent tools and packages, including SamTools and MaxEntScan, all of which are likewise freely available, but which must be obtained from their respective developers. The MiSplice documentation contains complete instructions for obtaining and linking these applications into MiSplice.

MINI GENE SPLICING ASSAY

Exons of interest and approximately 150 bp of their flanking intron sequences were PCR amplified from HEK293T genomic DNA using primers carrying restriction enzyme sites for BamH1 and MluI. PCR products were cleaned up using NucleoSpin PCR Cleanup (Macherey-Nagel) or DNA Clean and Concentrator-5 Kit (Zymo Research) and digested with BamHI and MluI. The digested pCAS2.1 vector and PCR products were ligated using T4 DNA Ligase (NEB). Mutations were introduced via Q5 Site-Directed Mutagenesis (NEB). WT and MUT constructs were confirmed by sequencing of the insert region. The plasmids were transiently transfected into HEK293T cells using Lipofectamine 2000 (ThermoFisher Scientific). 24 hours post transfection, cDNA was synthesized using 2 ug of total RNA with the Superscript III First-Strand Synthesis System (ThermoFisher Scientific) or igScript Reverse Transcriptase (Intact Genomics) and priming with Oligo(dT)20. Finally, cDNA was amplified using pCAS-KO1-(5′-TGACGTCGCCGCCCATCAC-3′) and pCAS-R (5′-ATTGGTTGTTGAGTTGGTTGTC-3′) and the alternative splicing patterns were evaluated on a 2.5% agarose gel with ethidium bromide. Qiaquick Gel Extraction Kit (Qiagen) was used to purify bands for sequencing.

CELL CULTURE

HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with fetal bovine serum (FBS) and penicillin streptomycin.

SELECTION OF VARIANTS FOR VALIDATION

In selecting variants to validate, we focused on mutations in cancer associated genes and known cancer predisposition genes. After selected gSCMs from cancer predisposition genes, we then filtered sites in 1) low complexity regions to avoid creating non-specific primers; 2) exons containing variants with restriction enzyme cut sites in BamHI and MluI; 3) exons containing variants in the first or last exon of a gene. In these cases signals from the UTR regions may disrupt the mini-gene splicing assay; 4) complex splicing events that couldn’t be properly captured in a mini-gene splicing assay; 5) exons with small introns. For the mini-gene splicing assay, we selected exons of interest that were located far enough away from adjacent exons with specific enough primers to amplify 150 bp of the intronic region on either side of the exon of interest.

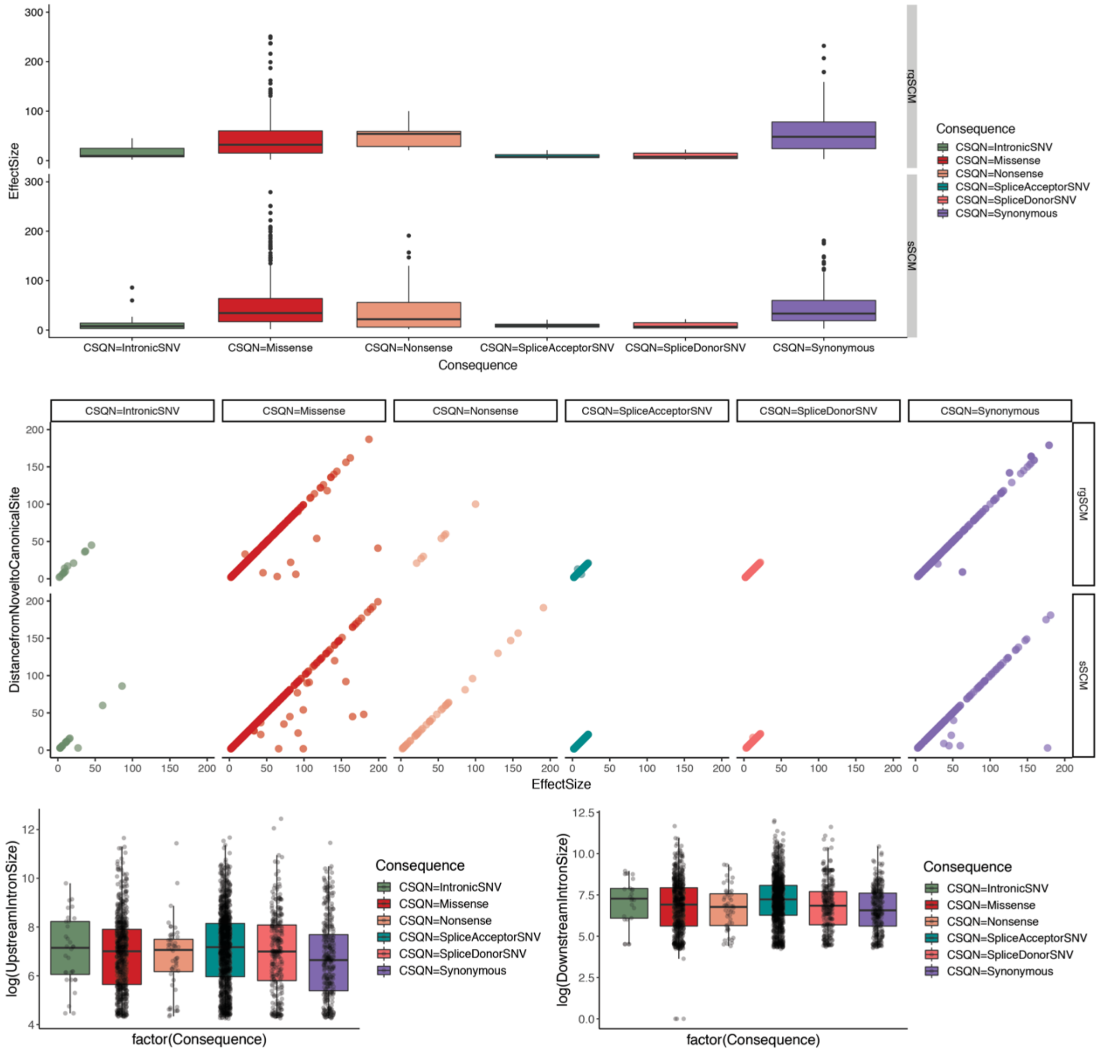
4.5 Discussion

Accurate and quick validation of somatic and germline variation is tantamount to personalized medicine. Our MiSplice pipeline has proven to properly classify and characterize splice-site-creating mutations and can help to improve our genomic annotation pipelines when RNA-sequencing and DNA-sequencing data is integrated to predict transcriptional consequences. Our findings also highlight the benefits of taking advantage of the selective mutagenesis that occurs in cancer genomes to evaluate splicing modulation across a given cohort. With our large and high confidence set of 2,888 SCMs we can effectively compare the landscape of rare and germline SCMs while grouping both sets together to evaluate overall trends in SCM+ exons. We were able to accurately determine that mutations overlapping the splice donor and splice acceptor splice site commonly undergo different selective pressures when mutated.

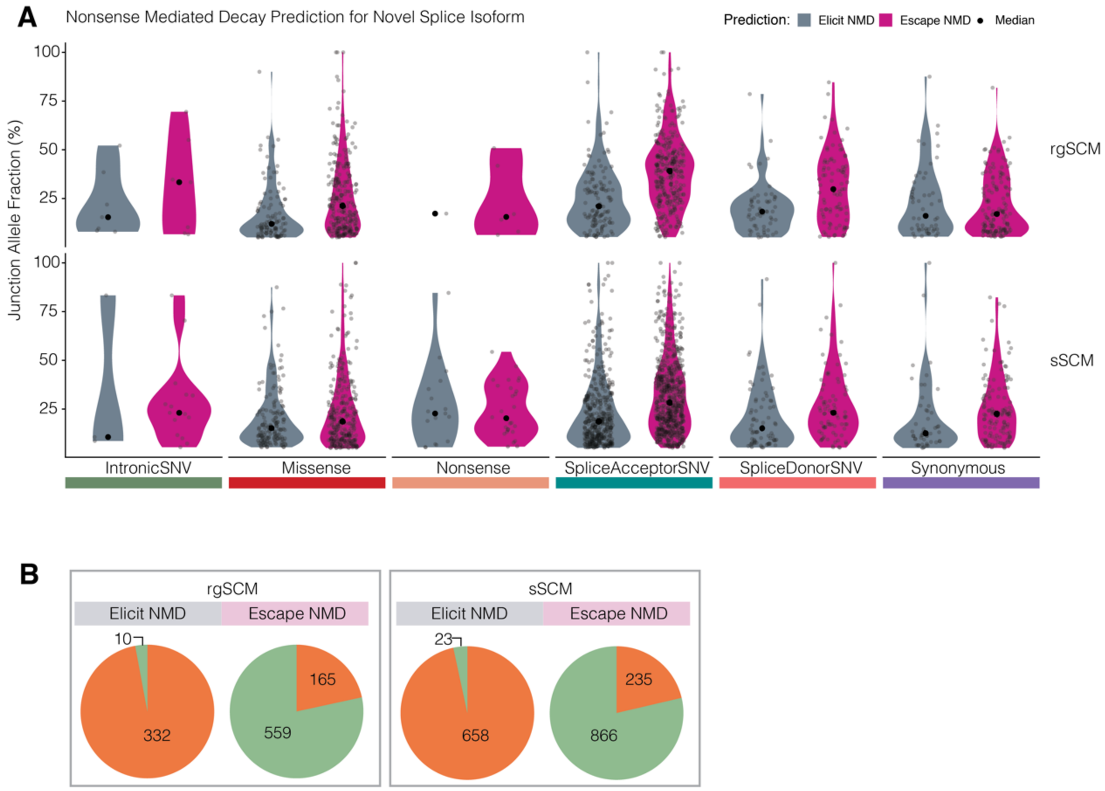
Specifically, mutations overlapping the splice donor site were sufficient to disrupt the canonical splice site usage but this phenomenon doesn't hold true for acceptor splice site mutations in our dataset. Alternatively acceptor SCMs needed to not only strengthen the novel splice site to facilitate the novel site usage, but also disrupt the canonical splice site. Furthermore, exons containing splice acceptor SCMs trended towards having a large upstream intron and downstream intron relative to other SCM+ exons. By surveying additional mutation induced alternative splicing events, we can continue to learn from the surrounding genomic context to tease apart the intricacies of the splicing code as it pertains to SCM+ exons. For example, we observed an enrichment of nucleosome occupancy overlapping SCM+ exons, suggesting active splicing of these exons across at least the 33 tissue types surveyed in our analysis. Overall, the size of the novel exon tended to decrease post mutation, mimicking a natural evolutionary selective pressure but exploited in the cancer genome to maintain proper alternative splicing. With respect to degradation of the novel isoform, regardless of whether or not the resulting protein product was expected to undergo degradation, we still actively see expression of the novel isoform, although significantly lower for the transcripts predicted to elicit NMD.

To date, this is the first analysis comparing rare germline SCMs and somatic SCMs revealing their comparable dysregulation to the splicing code in cancer. Evaluating mutation induced events separately from patient specific *de novo* events can provide a focused analysis on the genomic features selecting for SCM+ exons relative to leaky splicing or mutation independent cryptic splice site activation. As tissue type specific datasets continue to increase, developing novel tissue specific signatures will help inform the tissue type specific relevance of mutation induced SCMs. For example

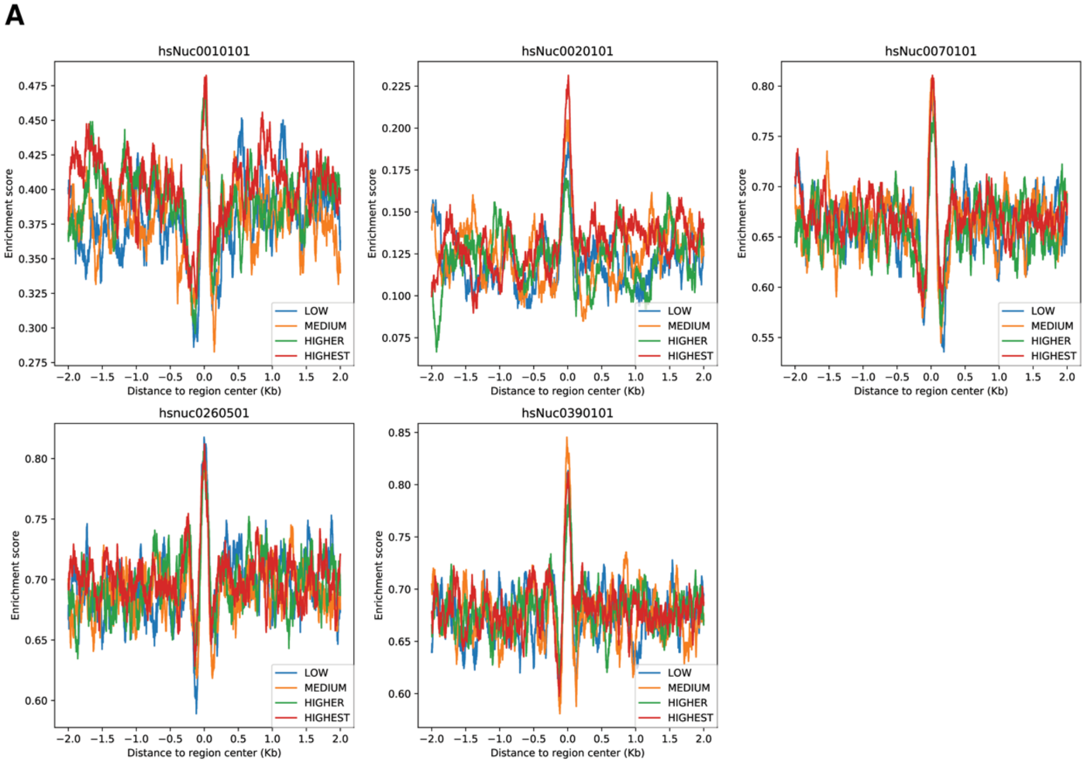
Understanding how splice-altering variants can lead to alternative isoforms in tumor samples and determining protein domains that are disrupted or created is still an open area of study. In order to better characterize events specific to a tissue type, developing modules of MiSplice that have known tissue specific transcripts will strengthen annotation and mutation calling. Furthermore, there is room to expand outside of the single nucleotide variant landscape to evaluate insertions, deletions and more complex events that are potentially also contributing to the disease state. Finally, MiSplice is a very stringent algorithm. There are many seemingly tissue type specific events that are present at a low level in some tissue types that have the potential to be exploited in other tissue types especially under direct mutagenesis. While we did not report on these cases in this analysis, the intermediate MiSplice outputs will provide very fruitful datasets for inferring novel biological mechanisms exploited in the cell.



Supplemental Figure S7. (A) Distribution of overall effect size of novel isoform broken down by conventional annotation and SCM type. (B) Comparing change in size of novel spliced isoform to the distance from the canonical splice site. Both values were generated from different steps in the pipeline, further confirming our annotation pipeline. (C) Distribution of upstream intron sizes by conventional annotation. (D) Distribution of upstream intron sizes by conventional annotation.



Supplemental Figure S8. NMD Prediction Overview . (A) Comparing overall JAF distribution by conventional annotation type and colored by NMD prediction. (B) Based on the location of the premature termination codon, each predicted protein product was classified as eliciting or escaping nonsense mediated decay based on the 50 bp rule.



Supplemental Figure S9. Nucleosome Occupancy by Canoncical Splice Site Score . (A) Exons containing SCMs were grouped into bins by strength of canonical splice site and nucleosome occupancy was evaluated by group using NucMap. Low [-17.58,6.39], Medium [6.40,8.21], Higher[8.22,9.77], Highest [9.78,15.07].

4.6 References

Uncategorized References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., *et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., *et al.* (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep *1*, 543-556.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., *et al.* (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385 e318.

Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nature structural & molecular biology *16*, 107-113.

Cieslik, M., and Chinnaiyan, A.M. (2018). Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet *19*, 93-109.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., *et al.* (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst *6*, 271-281 e277.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frebourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol Biol *653*, 249-257.

Heyer, W.D., Li, X., Rolfsmeier, M., and Zhang, X.P. (2006). Rad54: the Swiss Army knife of homologous recombination? Nucleic Acids Res *34*, 4115-4125.

Huang, K.L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N.*, et al.* (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355-370 e314.

Jayasinghe, R.G., Cao, S., Gao, Q., Wendl, M.C., Vo, N.S., Reynolds, S.M., Zhao, Y., Climente-Gonzalez, H., Chai, S., Wang, F.*, et al.* (2018). Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Rep *23*, 270-281 e273.

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. Mol Cells *40*, 1-9.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science (New York, NY) *302*, 2141-2144.

Kahles, A., Lehmann, K.V., Toussaint, N.C., Huser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research, N.*, et al.* (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211-224 e216.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A *100*, 189-192.

Li, Q., Xie, W., Wang, N., Li, C., and Wang, M. (2018). CDC7-dependent transcriptional regulation of RAD54L is essential for tumorigenicity and radio-resistance of glioblastoma. Transl Oncol *11*, 300-306.

Maquat, L.E., Tarn, W.Y., and Isken, O. (2010). The pioneer round of translation: features and functions. Cell *142*, 368-374.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol *6*, 386-398.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nature genetics *30*, 13-19.

Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends in biochemical sciences *23*, 198-199.

Nowacka-Zawisza, M., Brys, M., Hanna, R.M., Zadrozny, M., Kulig, A., and Krajewska, W.M. (2006). Loss of heterozygosity and microsatellite instability at RAD52 and RAD54 loci in breast cancer. Pol J Pathol *57*, 83-89.

Pelttari, L.M., Kiiski, J., Nurminen, R., Kallioniemi, A., Schleutker, J., Gylfe, A., Aaltonen, L.A., Leminen, A., Heikkila, P., Blomqvist, C.*, et al.* (2012). A Finnish founder mutation in RAD51D: analysis in breast, ovarian, prostate, and colorectal cancer. J Med Genet *49*, 429-432.

Popp, M.W., and Maquat, L.E. (2016). Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. Cell *165*, 1319-1322.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat Struct Mol Biol *16*, 990-995.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol Cell *36*, 245-254.

Sun, W., Duan, T., Ye, P., Chen, K., Zhang, G., Lai, M., and Zhang, H. (2018). TSVdb: a web-tool for TCGA splicing variants analysis. BMC Genomics *19*, 405.

Tang, L., Deng, L., Bai, H.X., Sun, J., Neale, N., Wu, J., Wang, Y., Chang, K., Huang, R.Y., Zhang, P.J.*, et al.* (2018). Reduced expression of DNA repair genes and chemosensitivity in 1p19q codeleted lower-grade gliomas. J Neurooncol.

Tong, Y., Merino, D., Nimmervoll, B., Gupta, K., Wang, Y.D., Finkelstein, D., Dalton, J., Ellison, D.W., Ma, X., Zhang, J.*, et al.* (2015). Cross-Species Genomics Identifies TAF12, NFYC, and RAD54L as Choroid Plexus Carcinoma Oncogenes. Cancer Cell *27*, 712-727.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J.S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B.T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol *13*, R35.

Wong, M.S., Kinney, J.B., and Krainer, A.R. (2018). Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. Mol Cell *71*, 1012-1026 e1013.

Zhao, Y., Wang, J., Liang, F., Liu, Y., Wang, Q., Zhang, H., Jiang, M., Zhang, Z., Zhao, W., Bao, Y.*, et al.* (2018). NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Res.

Zhou, W., Chen, T., Chong, Z., Rohrdanz, M.A., Melott, J.M., Wakefield, C., Zeng, J., Weinstein, J.N., Meric-Bernstam, F., Mills, G.B.*, et al.* (2015). TransVar: a multilevel variant annotator for precision genomics. Nat Methods *12*, 1002-1003.

Chapter 5: Conclusion and Future Directions

Cancer Specific Exon Definition

From studying differential alternative splicing across species, it is understood the appearance of alternatively spliced exons has manifested through three different mechanisms: (1) the transition of a constitutive exon to an alternatively spliced exon, (2) exon shuffling and (3) exonization of intronic sequences.

When a new or duplicated exon is inserted into an existing gene it is known as exon shuffling. This process is facilitated by the repetitive elements in introns leading to recombination events and thereby exon shuffling. The model of exon creation through exonization was initially thought to occur from nothing. While many exon creation events are due to repetitive elements (Alu elements) which can undergo exonization, there are two well accepted methods of exonization including: intron-mutation induced splice signal creation events and RNA editing. Alu elements belong to the short interspersed element family (SINE), consist of a 300 nucleotide sequence, and are inserted into various regions of the genome via retrotransposition. The regions of an Alu element strongly resemble a canonical splice site in that they contain a poly-pyrimidine tract and splice-site like signals nearby that can be inaccurately identified by the splicing machinery. The use of a new exon is further encouraged by strong alternatively spliced exons flanking the inserted sequence along with splicing enhancer and silencer elements. Interestingly, Alu exons have been found to be enriched for enhancer elements and depleted of silencer elements, encouraging the exonization event. In addition to the primary sequence, secondary structure can also affect exonization by means of adenosine deaminase acting on RNA (ADAR) editing. Adenosine to Inosine editing is common in Alu elements due to the secondary structure created when two inverted Alu elements are inserted near one another. During this process, the ADAR family of enzymes alters Adenosine residues to Inosine. Since Inosine is recognized by the biological machinery as Guanosine, a functional splice site can be created in an Alu element by altering AA to AG. There have been several cases of exon gain due to Alu insertions in exons (Sela et al.) and introns (unpublished).

Evolutionary conservation or depletion of a splice site provides strong evidence as to the biological function of the alternatively spliced product.

Understanding how exon creation or modification can generate a novel protein isoform is vital to understanding the evolution of canonical gene function. New exon events that arise in cancer or in natural selection in various species allows for the new exon event to be “tested”, since the new event is almost always alternatively spliced the wildtype form is also present in the organism as well (Sorek, 2007). If the alternative isoform is detrimental it will be selected against, and if beneficial we should see evidence of selection in the form of expression of the alternatively spliced product. It has been suggested that exonization events are ways to test out new gene modules with a generally small effect on the overall evolution of the species.

The findings from my thesis work suggests that evolutionary constraints on the tumor has selected for these splice-site-creating mutants and conferred some combination of the following criteria: (1) modify protein structure and/or (2) shift mRNA usage away from the wild type.

Although many of the predicted splice-site-creating events are predicted to escape nonsense mediated decay, about half of the sites should be degraded but are still present in the total RNA population, signifying a novel means of escaping degradation. In this thesis we propose an additional classification of “exonization” through the lens of cancer evolution. We have identified thousands of somatic and germline coding mutations, inducing the use of a novel or cryptic splice site functionally altering the reading frame of the mutant isoform and diversifying the landscape of the cancer genome. In the following sections, I will discuss additional analyses that can be performed on this compendium of splice-site-creating mutants to evaluate selection of SCMs across the TCGA dataset.

Allelic Imbalance of SCMs

Over evolutionary time during the birth of a new exon, while undergoing various selective pressures, newly created exons tend to exhibit increased expression and eventually become fixed over the previous wild-type form (Sorek, 2007). Similarly with our splice-creating mutants, we see varying levels of selection by the tumor, and in some cases the mutant induced splice form is selected at a higher frequency than the wild type suggesting preference for the alternatively spliced product (Sorek, 2007).

In thinking about cancer from an evolutionary standpoint, we have a unique opportunity to evaluate changes to overall exon definition and determine their potential for purifying selection over a relatively short time scale. When interrogating mutational signatures in cancer, we predict genomic changes will lead to changing expression of the resulting gene product. The clonal heterogeneity of a tumor exemplifies the theory of selective pressures on growing populations of cells in a microenvironment. Selective pressures (including treatment, microenviroment, tissue etc.) alter the clonal architecture and the clone with the most beneficial fitness is maintained until treated (Gerlinger et al.). By looking at the overall variant allele fraction of the DNA and RNA, we can evaluate the transcriptional dynamics that could be exploited by the cancer cell. When exons are newly created, some might fix over time and we can evaluate this by comparing expression of the newly created exons relative to the wildtype sequence (Sorek, 2007). When comparing the different Alu element insertions between mouse and humans, Sela et al., found increased levels of the newly created exons in humans relative to mouse (Sela et al.) suggesting selection for the Alu created exons over evolutionary time.

For approximately 80 rare germline splice-site-creating events in our dataset, we have access to DNA-Sequencing and RNA-Sequencing for both the tumor and normal samples. With these data we can assess the comparative Variant Allele Fractions (VAF) for DNA-seq and Junction Allele Fraction (JAF) in the RNA-Seq between the tumor and normal data. Understanding allelic imbalance will aid us in understanding the difference in expression between the mutant and wild type allele, a selection of one allele over another in the tumor would suggest either cis or trans factors that affect expression of the associated gene thereby altering the tumor biology.

Protein Support from Mass Spectrometry

Support of the novel splice-site-creating event could be significantly strengthened by peptide support. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) has performed mass spectrometry for a subset of samples from the cancer genome atlas project to further interrogate how mutations can manifest on the translational level. To evaluate the translational potential of the splice-site-creating mutants, for a handful of samples in our dataset we have mass spectrometry data to evaluate if the novel isoform is translated (Table 5.1). In Table 5.1 I have highlighted all the sample and gene pairs for which we have associated mass spectrometry data and can validate the novel splice isoform if it is expressed. Specifically, MSGF+ (https://omics.pnl.gov/software/ms-gf) can be utilized to identify peptides specific to our novel splice isoforms. Although we have evidence of the mutant isoforms in the RNA-Sequencing data, support by mass spec will provide very strong evidence that the mRNA is translated to a protein whose functions could be altered due to the alternatively spliced product.

Table 5.1. rgSCM and sSCM events with mass spectrometry data

Type SCM	**Result**	**Frame**	**NMD Classification**	**Gene**	**Cancer**	**JAF**	**Conventional Annotation**
germline	exon-extension	off-frame	ELICIT-NMD	CBWD5	BRCA	8.62069	SpliceDonorSNV
germline	exon-extension	off-frame	ESCAPE-NMD	C7	OV	39.2461	SpliceDonorSNV
germline	exon-shrinkage	in-frame	ESCAPE-NMD	SORBS3	BRCA	9.52381	Synonymous
germline	exon-shrinkage	in-frame	ELICIT-NMD	TUBGCP6	BRCA	41.3793	SpliceAcceptorSNV
germline	exon-shrinkage	off-frame	ELICIT-NMD	CCDC94	OV	9.40171	SpliceAcceptorSNV
germline	exon-shrinkage	off-frame	ELICIT-NMD	MTBP	BRCA	22.2222	Missense
germline	exon-shrinkage	in-frame	ESCAPE-NMD	SKA2	OV	21.4286	Missense
germline	exon-shrinkage	off-frame	ESCAPE-NMD	LMO1	OV	18.75	Synonymous
germline	exon-shrinkage	in-frame	ESCAPE-NMD	ZCCHC17	BRCA	14.8148	Missense

germline	exon-shrinkage	in-frame	ESCAPE-NMD	PFKL	BRCA	45.1128	SpliceAcceptorSNV
germline	exon-shrinkage	in-frame	ESCAPE-NMD	PIAS3	OV	34.9593	Missense
germline	exon-shrinkage	off-frame	ELICIT-NMD	ELMO3	OV	6.76692	Missense
germline	exon-shrinkage	off-frame	ESCAPE-NMD	HMGCR	BRCA	8.94309	SpliceAcceptorSNV
germline	exon-shrinkage	in-frame	ESCAPE-NMD	CHD8	OV	27.3092	Missense
germline	exon-shrinkage	off-frame	ELICIT-NMD	NEDD1	BRCA	5.9322	Missense
germline	exon-shrinkage	off-frame	ESCAPE-NMD	CUEDC1	BRCA	36.4985	SpliceDonorSNV
germline	exon-shrinkage	in-frame	ESCAPE-NMD	PTPRM	OV	49.1228	SpliceAcceptorSNV
somatic	exon-extension	off-frame	ELICIT-NMD	TP53	OV	46.6667	SpliceDonorSNV
somatic	exon-extension	off-frame	ESCAPE-NMD	ZDHHC16	OV	30.2013	Missense

somatic	exon-shrinkage	off-frame	ESCAPE-NMD	SNAPC1	BRCA	21.1538	SpliceAcceptor SNV
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	CDC37	OV	16.7965	Synonymous
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	NFYB	OV	19.5402	SpliceAcceptor SNV
somatic	exon-shrinkage	off-frame	ELICIT-NMD	LANCL2	OV	22.7273	SpliceAcceptor SNV
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	NOP9	BRCA	36.3636	Missense
somatic	exon-shrinkage	off-frame	ELICIT-NMD	EDC3	OV	7.5188	SpliceAcceptor SNV
somatic	exon-shrinkage	off-frame	ESCAPE-NMD	C1orf172	BRCA	12.8205	Missense
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	GTF2E2	BRCA	12.8713	SpliceAcceptor SNV
somatic	exon-shrinkage	off-frame	ESCAPE-NMD	ID3	OV	23.3051	SpliceAcceptor SNV
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	EIF1AX	OV	39.0476	SpliceAcceptor SNV

somatic	exon-shrinkage	in-frame	ESCAPE-NMD	FBXO18	OV	8.31889	Missense
somatic	exon-shrinkage	in-frame	ESCAPE-NMD	CRAMP1L	BRCA	39.2857	Missense
somatic	exon-shrinkage	off-frame	ELICIT-NMD	CDK12	OV	40.7407	SpliceAcceptor SNV

Oncogenic Function of SCMs

MiSplice identified two kidney renal clear cell carcinoma (KIRC) samples having the same conventionally annotated missense mutation (c.233A>G, p.N78S) in BAP1, a nuclear deubiquitinase, that created the same novel spliced-out alternative splicing product. Inactivation of BAP1 is prevalent among renal cell carcinomas (Wadt et al., 2012) and an annotated missense mutation (p.L570V) has been reported to create a cryptic splice site in melanoma. At the transcriptional level, the expression of the case and control samples are relatively comparable, but at the translational level, one case with available protein data (RPPA) showed significantly lower expression relative to the controls. This result suggests the conventionally annotated missense mutations in BAP1 likely create an alternatively spliced transcript that is not readily expressed at the protein level. With this evidence it would be interesting to determine the in vitro function of BAP1 splice-creating variant in a cell line of interest. Using the established cell lines with our splice-site-creating variants of interest, downstream assays can be performed to determine if the SCMs have the potential to contribute to oncogenesis.

Mutation Induced Ultra Short Introns

With the birth of a new intron, we see cases of intronization or the alternative splicing of an internal exon (Irimia et al., 2008). With certain events we also see the alternative splicing of internal introns that look to be facilitated by selected mutation induced splicing events. Small intron-creation events in exons were observed in approximately ~150 unique gene-cancer type events. While introns are commonly spliced out by the major and minor spliceosome, several alternative mechanisms can facilitate the splicing of very short introns, typically less than 65 nucleotides in length. For example a U2 snRNP SF3b excises short introns ranging from 43 to 56 bp in *NDOR1*, *HNRNPH1* and *ESRP2* (Sasaki-Haraguchi et al.). While introns in humans are quite long, introns in invertebrates are much shorter. With this knowledge, Lim et al. assessed genomic features of short introns across organisms to determine the minimum amount of information required to define an intron (Lim and Burge). While there is ample evidence in the literature supporting ultra short introns (USI), genomic deletions can arise due to repetitive sequences. To filter out common genomic deletions, we overlapped our predicted USI's with computationally predicted USI's in a recent publication by Abebrese et al (Abebrese et al.). In total, 29 USI's were filtered leaving less than 20 putative mutation induced USI's. Furthermore, 3 USI's in *KMT2D*, *CBLN3* and *PRRC2C* overlapped with Abebrese et al. and *PRRC2C* had evidence of splice junction usage across several cell lines after screening A549, Bj, H1 hESC, HeLa, HepG2, Hsmm, K562, Mcf7, Nhek, and Nhlf.

To date, mutation induced USI's have not been systematically evaluated in cancer. In general for all the USI's identified, all but one are in frame and generally do not change the splice site score but are recurrent and span cancer types. While the splicing score doesn't change, it has been reported that short introns generally have non-canonical splice sites such as those observed in IRE1alpha-dependent Xbp1 mRNA splicing known for splicing out 26 nucleotide sequences in various genes including itself (Bai et al.). The predicted mutation induced USIs identified in the TCGA dataset provides additional evidence regarding the mechanisms by which the splicing code can be dysregulated in the human population and potentially exploited in cancer.

Common Germline Splice-Site-Creating Variants

As a proof concept, we wanted to explore common germline splice-site-creating events in one cancer type to determine ancestry specific events. For the 20,205,168 common variants in Breast Invasive Carcinoma (BRCA), in an initial screen only 72,245 unique variants were evaluated for splice-site-creating function. Of the 72,245 unique variants, 287 had evidence of splice-creating function in at least one sample. The variants with splice-site-creating function were then evaluated in the remaining samples. In total 74,383 variants derived from 287 unique sites were evaluated for splice-site-creating function. 29 variants with less than 10% of control samples exhibiting the alternatively spliced isoform were manually reviewed, resulting in 6 high confidence common germline splice-site-creating variants (Table 5.2).

Table 5.2. Common SCMs

Gene	Mutation	# Cases	% of Controls with event	# reads in case samples
Neurolysin, mitochondrial (NLN)	5:65084101 A>G	22	0 %	15,14,18,11,8,7,12,13,5,7,11,13,6,5,10,13,6,7,18,27,9,10
Leukocyte Immunoglobulin Like Receptor A2 (LILRA2)	19:55098667 G>A	20	0.39 %	11,5,7,5,15,6,7,7,5,8,13,10,15,5,17,9,8,10,9,10
Glutaminyl-Peptide Cyclotransferase Like (QPCTL)	19:46206262 G>A	4	1.07 %	7,5,5,7
TatD DNase Domain Containing 3 (TATDN3)	1:212985592 G>A	17	2.25 %	6,6,6,6,6,5,8,6,6,7,5,8,5,8,7,6,12
Calpain 13 (CAPN13)	2:30985977 G>A	4	2.64 %	9,11,8,16

NAD kinase 2, mitochondrial (NADK2)	5:36219710 C>T	22	6.01 %	13,6,18,9,8,5,17,8,25,18,11,7,8,19,12,35,17,34,16,6,6,7

The NLN common germline variant has a population frequency of 0.19 in the african american community. 20 of the 22 samples with the germline variant are black or african american in the TCGA dataset and have a slightly higher expression of the NLN gene relative to non-mutated controls in breast cancer, although not significant (wilcox test = 0.09). Our initial findings of the NLN gene and other common variants suggest a subset could be mis-annotated when evaluating both RNA and DNA sequencing. Interestingly a few variants created ultra-short-intron events described in the previous section.

Non-Canonical Splice Sites and Minor Spliceosome Alterations

For both splice-site-creating papers, we only focused on sites creating canonical GT or AG splice sites. Several interesting SCMs created non-canonical splice sites (deviating from GT and AG genomic context) suggesting usage of non-canonical splice sites by the major spliceosome or minor spliceosome.

The splicing code is made up of cis-acting elements that help the splicing complex distinguish between non-coding (intron) and coding (exon) regions. There are two known distinct ribonucleoprotein complexes known as the major and minor spliceosomes that are responsible for joining exons and splicing out two types of introns, U2 and U12. For

U2 type introns, the consensus intronic dinucleotide GT splice donor and AG splice acceptor flank the exons at the 3’ and 5’ ends respectively, and can be found in 99% of all introns. The remaining 1% are U12 type introns and were initially recognized due to their non-consensus splice site sequences AT and AC at the 3’ and 5’ splice sites, respectively (Hall and Padgett; Jackson). Interestingly, some introns with the canonical GT-AG splice sites are also spliced by the minor spliceosome. While the consensus sequence alone cannot differentiate between the two groups, several other cis-sequences including lacking a polyprimidine tract upstream of the 3’ splice site, polypyrimidine tract and branch point (Dietrich et al.) can help differentiate between the two groups. The major and minor spliceosome share the U5 snRNP and many spliceosomal proteins.

While only 1% of U12 type introns are present in the genome, they are highly conserved across eukaryotes suggesting a potential functional role in the genome. For example, U12 type introns are shown to be removed at a much slower rate than U2 type introns (Patel et al.), and the minor spliceosomal proteins are present at a much lower abundance than the major spliceosome (Tarn and Steitz). This evidence suggests that minor introns are likely a rate limiting step in the expression of various mRNAs that contain U12 type introns. Despite this evidence, the current compendium of minor introns is still expanding.

The benefit of coupling U12 intron containing genes and U12 spliceosome abundance allows the cell a regulatory mechanism by which only a subset of genes can be differentially regulated in response to stimuli. Younis et al., determined the U6atac snRNA minor spliceosomal gene expression could be increased by cell stress activated kinases leading to increased expression of U12 intron containing genes that regulate cell stress physiology (Younis et al.). RNA-seq analysis revealed a down-regulation of 2,088 genes including 429 minor intron genes. Together, these results show that altering expression of minor spliceosomal genes can alter a distinct subset of genes with overarching changes at the transcriptome level.

With the large dataset provided by The Cancer Genome Atlas mechanisms of U12 dysregulation across cancer genomes can be more effectively evaluated. Furthermore, our analysis of splice-site-creating mutants identified the usage of many non-canonical GT-AG splice sites, suggesting a potential dysregulation of spliceosomal usage.

As with aberrations in major spliceosome related genes, mutations in minor spliceosomal genes lead to different disease phenotypes ranging from brain to skeletal irregularities. Recently, Madan et al., evaluated alternatively spliced junctions in 8 Myelodysplastic syndrome (MDS) patients with *ZRSR2* mutations and identified 689 mis-spliced junctions in all 8 cases (Madan et al.). The mispliced junctions are specific to *ZRSR2* mutants. Additionally, a conditional knockdown of *Rnu11* in mice resulted in microcephaly and upregulated intron retention of minor intron containing genes (Baumgartner et al.). Many of the minor intron containing genes were not differentially expressed but 178 introns were detained at higher levels in the mutant relative to the wildtype while one was downregulated. Finally a family of three sisters were found to have biallelic mutations in *RNPC3* and when downstream U12 intron retention rates were compared, 21 genes out of 522 tested had decreased U12/U2 ratios in the patient cells(Argente et al.). Interestingly the subset of aberrant U12 introns were relevant to the disease phenotype and some U2 cryptic splice sites were activated instead of the U12 intron. Mutations in U12 spliceosomal proteins have been linked to several other diseases including RNU12 (Elsaid et al.), RNU4ATAC (Edery et al.; He et al.; Heremans et al.; Merico et al.), TRAPPC2 (Shaw et al.), FUS (Reber et al.), and SMN1 (Zhang et al.). The previous studies provide strong evidence that minor splicing factors regulate subsets of minor introns. By evaluating differential expression of related minor introns, novel functional mutations in minor spliceosome genes or U12 minor introns themselves can be identified to determine if minor intron splicing is altered globally in cancer.

As with mutations in spliceosomal genes, mutations in cis can also alter minor intron splicing. *LKB1* encodes a serine threonine protein kinase involved in major cellular processes including cell cycle arrest, p53 mediated apoptosis among others. The second intron of *LKB1* is a minor intron and a mutation that affects the 5’ splice site causes Peutz-Jeghers syndrome (Hastings et al.). Interestingly, the 5’ splice site mutation changes the splice sites of intron 2 from AT-AC to GT-AC. One would expect that a 5’ splice site mutation would cause skipping of the subsequent exon or use of an alternative 5’ splice site, but the authors found instead that a cryptic 3’ splice site was utilized, thereby altering the frame of the final transcript, which induced a premature termination codon and degraded by nonsense mediated decay.

In a related manner, a mutation in *SCN8A* has been shown to inactivate the 5’ splice site thereby resulting in an exon skipping event associated with an X-linked recessive disorder (Shaw et al.). While this exon skipping event doesn’t occur in the minor intron itself, it does lead to altering the usage of minor intron contained within the gene, suggesting a strong interplay between splice site usage.

In 2018, our lab published a comprehensive analysis of gene fusions identified in RNA-Seq across the TCGA compendium (Gao et al.). From this publication, several gene fusions were identified involving minor spliceosomal related genes (Table 5.3). All of these resulting fusion products in the associated samples have the potential to dysregulate a large subset of downstream minor introns, but their current consequence is unknown. A case control comparison evaluating overall expression of the fusion genes with wild type counterparts will determine if the expression of the fusion itself or downstream targets are affected due to the predicted fusion product.

Table 5.3. U12 and U12 associated genes with detected Fusions in TCGA

Cancer	**Sample**	**Fusion**	**Junction**	**Spanning**	**Breakpoint1**	**Breakpoint2**
HNSC	TCGA-CV-7090-01A-11R-2016-07	DHX15--RBPJ	6	21	chr4:24548855:-	chr4:26386353:+
PRAD	TCGA-G9-6339-01A-12R-A311-07	DHX15--ETV1	9	35	chr4:24584323:-	chr7:13935896:-
SARC	TCGA-DX-A3LU-01A-11R-A21T-07	DHX15--CCND3	1	12	chr4:24541873:-	chr6:41940585:-

COAD	TCGA-F4-6809-01A-11R-1839-07	FUS--PYCARD	2	116	chr16:31182422:+	chr16:31202203:-
ESCA	TCGA-IG-A51D-01A-11R-A36D-31	FUS--IL9R	6	7	chr16:31182664:+	chrY:57189426:+
HNSC	TCGA-CV-6941-01A-11R-1915-07	FUS--PRSS36	1	11	chr16:31184396:+	chr16:31149731:-
OV	TCGA-13-1481-01A-01R-1565-13	FUS--KAT8	22	22	chr16:31180227:+	chr16:31127035:+
STAD	TCGA-HU-A4HB-01A-12R-A251-31	FUS--TMEM114	71	35	chr16:31185179:+	chr16:8572224:-
BRCA	TCGA-LD-A7W5-01A-22R-A352-07	PDCD7--TAF15	3	16	chr15:65129032:-	chr17:35817716:+
BRCA	TCGA-E2-A14Z-01A-11R-A115-07	RNPC3--KIAA0825	3	5	chr1:103526262:+	chr5:94417380:-
BRCA	TCGA-E2-A10A-01A-21R-A115-07	SNRNP48--NISCH	2	1	chr6:7595101:+	chr3:52484513:+
SARC	TCGA-IW-A3M4-01A-11R-A21T-07	ZCRB1--YAF2	1	8	chr12:42325924:-	chr12:42210619:-
SKCM	TCGA-D9-A6E9-06A-12R-A311-07	ZCRB1--LIMA1	18	64	chr12:42324019:-	chr12:50222485:-
BLCA	TCGA-G2-AA3B-01A-11R-A39I-07	ZMAT5--ASCC2	4	15	chr22:29766949:-	chr22:29793676:-
GBM	TCGA-41-2572-01A-01R-1850-01	ZMAT5--ASCC2	110	32	chr22:29766872:-	chr22:29790548:-
HNSC	TCGA-T3-A92M-01A-31R-A39I-07	ZMAT5--NIPSNAP1	5	11	chr22:29766872:-	chr22:29561862:-

LUAD	TCGA-05-4426-01A-01R-1206-07	ZMAT5--KIAA1671	1	3	chr22:29766949:-	chr22:25014430:+
LUSC	TCGA-85-8048-01A-11R-2247-07	ZMAT5--HIRA	2	7	chr22:29738330:-	chr22:19331556:-
SARC	TCGA-Z4-A9VC-01A-11R-A37L-07	DDIT3--FUS	3	23	chr12:57520418:-	chr16:31189665:+
BRCA	TCGA-E2-A1IO-01A-11R-A144-07	CLPX--PDCD7	1000	1000	chr15:65166642:-	chr15:65133183:-
LUAD	TCGA-97-A4M7-01A-11R-A24X-07	MPG--SNRNP25	32	116	chr16:83256:+	chr16:57086:+
LUSC	TCGA-21-1070-01A-01R-0692-07	ARHGAP17--SNRNP25	1	7	chr16:24964197:-	chr16:56539:+
SKCM	TCGA-D3-A8GC-06A-11R-A37K-07	FAM129B--SNRNP25	6	35	chr9:127516857:-	chr16:55777:+
SKCM	TCGA-EB-A44Q-06A-11R-A266-07	KDM2B--SNRNP35	5	24	chr12:121532806:-	chr12:123458021:+
BRCA	TCGA-E2-A14T-01A-11R-A115-07	TMEM117--ZCRB1	18	26	chr12:43844928:+	chr12:42324104:-
STAD	TCGA-BR-8077-01A-11R-2343-13	YAF2--ZCRB1	9	8	chr12:42210355:-	chr12:42313986:-
OV	TCGA-24-1425-01A-02R-1566-13	ST6GALNAC1--ZMAT5	7	8	chr17:76643508:-	chr22:29742480:-
PAAD	TCGA-3A-A9IH-01A-12R-A39D-07	NF2--ZMAT5	3	5	chr22:29678323:+	chr22:29742480:-

In evaluating disruptions of minor introns, several experiments can validate the *in silico* predictions. If the minor intron is retained due to mutation or disruption of U12 or U12-

associated splicing factor we would expect to see either expression of the retained intron in RNA-Seq or decreased expression of the associated gene. If the minor intron has an associated premature termination codon, one would predict the resulting transcript would be degraded by nonsense mediated decay. If effectively degraded, no reads supporting the transcript would be identified in the RNA-Seq. To confirm this, the transcript containing the mutation can be transfected into the cell line of interest and puromycin treated to inactivate the NMD pathway thereby allowing us to evaluate the full repertoire of transcripts produced by the mutant transcript. If the transcript is alternatively spliced due to the mutation in the minor intron and is normally degraded by NMD, in this experiment we should see evidence of the minor intron retained due to the knockdown of NMD factors by puromycin. Alternatively, if we do not see the transcript even after NMD knockdown, then the mutation in the minor intron is likely not disrupting this particular intron retention event.

If the minor intron is retained as suggested by evidence in the RNA-Sequencing data and it isn't degraded efficiently, we can next try and determine if the major and minor spilceosomal machinery is responsible for removal of the minor intron of interest. Since the U2 spliceosome is known to be degenerate and can identify multiple sequences, it is possible that the U12 spliceosome isn't playing a role in removing the minor intron of interest. To test this experimentally(Hastings et al.), a mini-gene derived with our exon, intron and exon of interest can be can be incubated in nuclear extract. When incubated in the presence of ATP, we can evaluate mRNA products that are derived from ATP dependent splicing reactions. Then in this same experimental condition, U2 and U12 dependent spliceosomes can be inactivated using antisense oligonucleotides directed at the U2 small nuclear RNA, U12 snRNA or both. Finally, RNA is extracted from all conditions and alternatively spliced products can be evaluated via RT-PCR. Outcome 1: If U12 spliceosome is important for this particular minor intron splicing then you will see a decrease or absence in the spliced product when associated spliceosomal RNA is knocked down. And when U2 spliceosome is knocked down there should be no change in spliced product. Outcome 2: If the U2 spliceosome is important for the splicing of this transcript then you will see a decrease or absence in the spliced product when associated spliceosomal RNA is knocked down. And when U12 is knocked down there should be no change in spliced product. Outcome 3: Both spliceosomes may be important for this process thereby resulting in differences in the spliced products under both knockdown conditions.

5.1 References

Uncategorized References

Abebrese, E.L., Ali, S.H., Arnold, Z.R., Andrews, V.M., Armstrong, K., Burns, L., Crowder, H.R., Day, R.T., Jr., Hsu, D.G., Jarrell, K., *et al.* (2017). Identification of human short introns. PLoS One *12*, e0175393.

Argente, J., Flores, R., Gutierrez-Arumi, A., Verma, B., Martos-Moreno, G.A., Cusco, I., Oghabian, A., Chowen, J.A., Frilander, M.J., and Perez-Jurado, L.A. (2014). Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. EMBO Mol Med *6*, 299-306.

Bai, Y., Hassler, J., Ziyar, A., Li, P., Wright, Z., Menon, R., Omenn, G.S., Cavalcoli, J.D., Kaufman, R.J., and Sartor, M.A. (2014). Novel bioinformatics method for identification of genome-wide non-canonical spliced regions using RNA-Seq data. PLoS One *9*, e100864.

Baumgartner, M., Olthof, A.M., Aquino, G.S., Hyatt, K.C., Lemoine, C., Drake, K., Sturrock, N., Nguyen, N., Al Seesi, S., and Kanadia, R.N. (2018). Minor spliceosome inactivation causes microcephaly, owing to cell cycle defects and death of self-amplifying radial glial cells. Development *145*.

Dietrich, R.C., Incorvaia, R., and Padgett, R.A. (1997). Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell *1*, 151-160.

Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M.B., Nampoothiri, S.*, et al.* (2011). Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. Science *332*, 240-243.

Elsaid, M.F., Chalhoub, N., Ben-Omran, T., Kumar, P., Kamel, H., Ibrahim, K., Mohamoud, Y., Al-Dous, E., Al-Azwani, I., Malek, J.A.*, et al.* (2017). Mutation in noncoding RNA RNU12 causes early onset cerebellar ataxia. Ann Neurol *81*, 68-78.

Gao, Q., Liang, W.W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L.*, et al.* (2018). Driver Fusions and

Their Implications in the Development and Treatment of Human Cancers. Cell Rep *23*, 227-238 e223.

Gerlinger, M., McGranahan, N., Dewhurst, S.M., Burrell, R.A., Tomlinson, I., and Swanton, C. (2014). Cancer: evolution within a lifetime. Annu Rev Genet *48*, 215-236.

Hall, S.L., and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J Mol Biol *239*, 357-365.

Hastings, M.L., Resta, N., Traum, D., Stella, A., Guanti, G., and Krainer, A.R. (2005). An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. Nat Struct Mol Biol *12*, 54-59.

He, H., Liyanarachchi, S., Akagi, K., Nagy, R., Li, J., Dietrich, R.C., Li, W., Sebastian, N., Wen, B., Xin, B.*, et al.* (2011). Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. Science *332*, 238-240.

Heremans, J., Garcia-Perez, J.E., Turro, E., Schlenner, S.M., Casteels, I., Collin, R., de Zegher, F., Greene, D., Humblet-Baron, S., Lesage, S.*, et al.* (2018). Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. J Allergy Clin Immunol *142*, 630-646.

Irimia, M., Rukov, J.L., Penny, D., Vinther, J., Garcia-Fernandez, J., and Roy, S.W. (2008). Origin of introns by 'intronization' of exonic sequences. Trends Genet *24*, 378-381.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res *19*, 3795-3798.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A *98*, 11193-11198.

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., Sanada, M., Grossmann, V., Sundaresan, J., Shiraishi, Y.*, et al.* (2015). Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. Nat Commun *6*, 6042.

Merico, D., Roifman, M., Braunschweig, U., Yuen, R.K., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B.*, et al.* (2015). Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nat Commun *6*, 8718.

Patel, A.A., McCarthy, M., and Steitz, J.A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. EMBO J *21*, 3804-3815.

Reber, S., Stettler, J., Filosa, G., Colombo, M., Jutzi, D., Lenzken, S.C., Schweingruber, C., Bruggmann, R., Bachi, A., Barabino, S.M.*, et al.* (2016). Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. EMBO J *35*, 1504-1521.

Sasaki-Haraguchi, N., Shimada, M.K., Taniguchi, I., Ohno, M., and Mayeda, A. (2012). Mechanistic insights into human pre-mRNA splicing of human ultra-short introns: potential unusual mechanism identifies G-rich introns. Biochem Biophys Res Commun *423*, 289-294.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A., and Ast, G. (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol *8*, R127.

Shaw, M.A., Brunetti-Pierri, N., Kadasi, L., Kovacova, V., Van, M.L., De, B.D., Salerno, M., and Gecz, J. (2003). Identification of three novel SEDL mutations, including mutation in the rare, non-canonical splice site of exon 4. ClinGenet *64*, 235-242.

Sorek, R. (2007). The birth of new exons: mechanisms and evolutionary consequences. RNA *13*, 1603-1608.

Tarn, W.Y., and Steitz, J.A. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801-811.

Wadt, K., Choi, J., Chung, J.Y., Kiilgaard, J., Heegaard, S., Drzewiecki, K.T., Trent, J.M., Hewitt, S.M., Hayward, N.K., Gerdes, A.M.*, et al.* (2012). A cryptic BAP1 splice mutation in a family with uveal and cutaneous melanoma, and paraganglioma. Pigment Cell Melanoma Res *25*, 815-818.

Younis, I., Dittmar, K., Wang, W., Foley, S.W., Berg, M.G., Hu, K.Y., Wei, Z., Wan, L., and Dreyfuss, G. (2013). Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. Elife *2*, e00780.

Zhang, Z., Lotti, F., Dittmar, K., Younis, I., Wan, L., Kasim, M., and Dreyfuss, G. (2008). SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing. Cell *133*, 585-600.

Appendix: Additional Projects

5.1 Complex Insertions and Deletions in Cancer Genomes

Systematic Discovery of Complex Indels in Human Cancers (Nature Medicine, 2016)
(Ye et al., 2016)

Contribution: I lead the development and scripted several scripts that were used to develop a variant quality control pipeline for filtering complex insertions and deletions. I manually reviewed many complex indels and helped with writing and submission. I helped to validate complex insertions and deletions identified in a COLO829 cell line by sanger sequencing.

Refer to Supplementary Note 2: Sanger Sequencing of Validated Complex Indels from COLO829.

5.2 Analysis of Somatic Complex Insertions and Deletions in Pediatric Hematopoietic Malignancies

Identification of Complex Indels in Pediatric Cancer Genomes

Complex indels are formed by deleting and inserting DNA fragments of different sizes at a common genomic location. Application of the publicly available Pindel-C suite (https://github.com/genome/pindel) to 545 whole genome Pediatric Cancer Genome Project (PCGP) samples uncovered 21 somatic complex insertions and deletions among 20 pediatric hematopoietic cancer samples listed in Table below. Our analysis was restricted to the following genes commonly mutated in hematopoietic diseases including: *TP53, CBL, CREBBP, FLT3, KMT2D (MLL2), PAX5, SETD2, IKZF1, PMS2, RAD51, NCOR1, TBL1XR1 and KMT2C (MLL3)*. Nine of the 21 complex insertions and deletions identified differed from those reported in the original studies as follows:

- 2 complex events in *TBL1XR1* and *SETD2* were not reported;
- 6 complex events in *MLL2*, *SETD2*, *IKZF1*, and *CREBBP* were classified as simple indels;
- 1 complex event in *TP53* was classified as both a SNV and an INDEL separately.

The remaining 12 events occurring among several key genes are found among samples that are currently not publicly available. These samples are among the following cancer types: Hyperdiploid Acute Lymphoblastic Leukemia (SJHYPER), E2A-PBX Acute Lymphoblastic Leukemia (SJE2A), and ETS-Related Gene Associated Acute

Lymphoblastic Leukemia (SJERG). A handful of somatic and germline complex indels have also been identified in non-coding regions of the genome, but our initial analysis focuses on coding somatic mutations in cancer genes.

This preliminary analysis demonstrates the importance of using the most updated form of Pindel-C to identify complex events that may be absent or classified as other variant types in current cancer genomics studies.

Table A.1. Complex indels identified for PCGP Samples

Sample	**Gene**	**Chromosome**	**Start**	**End**
SJE2A043	*PAX5*	9	37002647	37002649
SJERG031	*MLL2*	12	49444452	49444454
SJERG020052	*MLL3*	7	151842335	151842365
SJERG020306	*SETD2*	3	47059133	47059166
SJERG020307	*TBL1XR1*	3	176756153	176756179
SJERG020309	*MLL2*	12	49444958	49444960
SJETV001	*TBL1XR1*	3	176769294	176769299
SJETV024	*SETD2*	3	47059194	47059204
SJHYPER005	*NCOR1*	17	15952238	15952243
SJHYPER007	*FLT3*	13	28608280	28608286

SJHYPER051	*TP53*	17	7577595	7577601
SJHYPER097	*SETD2*	3	47162194	47162195
SJHYPER104	*PAX5*	9	37002675	37002676
SJHYPER108	*MLL2*	12	49443861	49443863
SJHYPER111	*CBL*	11	119148875	119148900
SJHYPER111	*FLT3*	13	28608305	28608309
SJHYPER227	*IKZF1*	7	50367310	50367314
SJHYPO002	*CREBBP*	16	3843416	3843439
SJHYPO040	*CREBBP*	16	3799606	3799632
SJINF013	*TP53*	17	7577093	7577111
SJPHALL020043	*SETD2*	3	47147487	47147488

Complex indel discovery

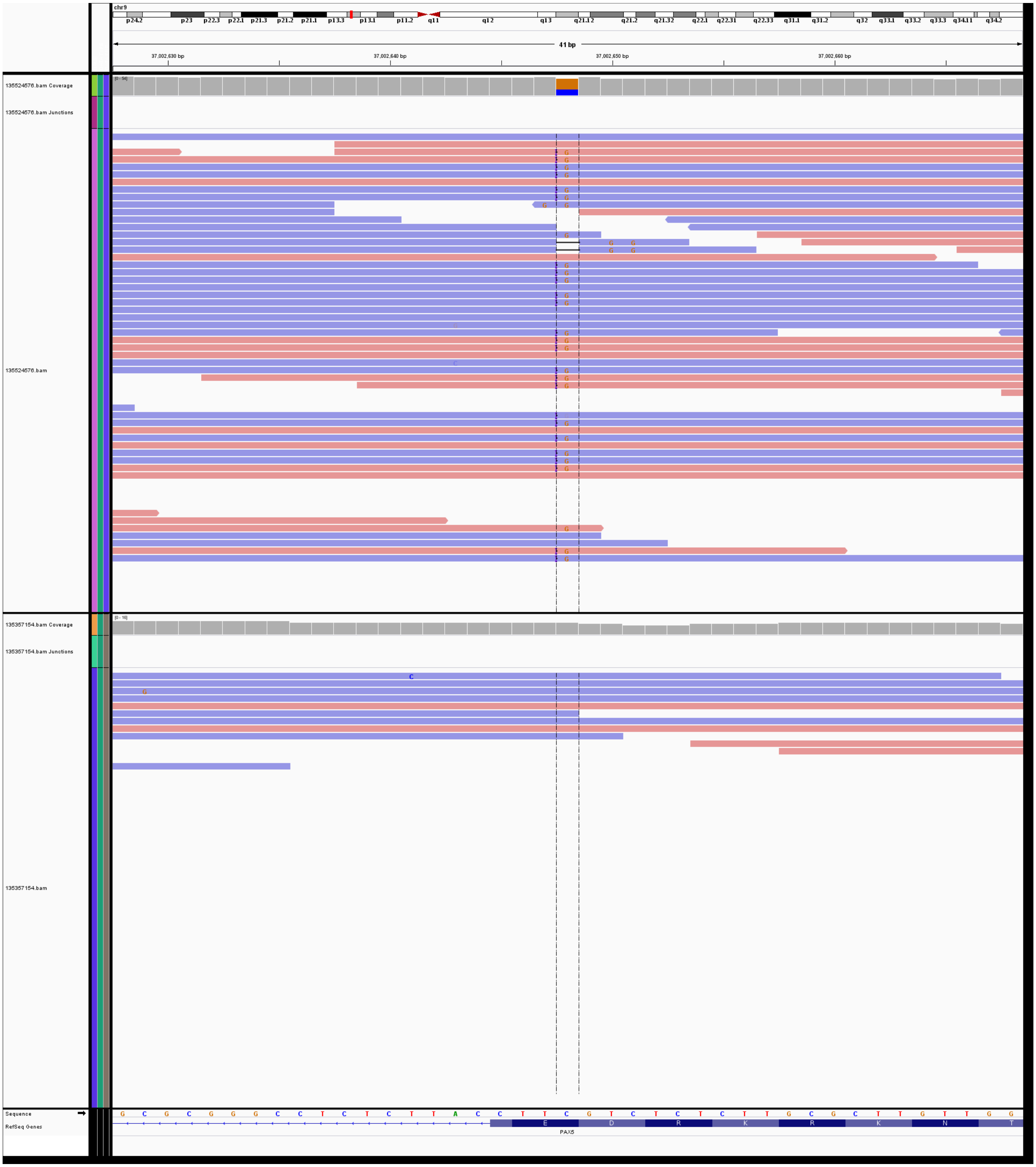
1. Read Extraction
2. Pattern Growth-Based Alignment
3. Distinguishing Complex Indels from Simple Indels
4. Remove False Predictions and Variant Allele Frequency Analysis

Filtering procedure

1. Coverage: All sites with at least 20 reads supporting the site of interest in both the tumor and normal sample were maintained for further analysis.
2. Coding Region Selection: Identified all complex indels overlapping an exon as defined by ensembl 75 including flanking 2 bp splice sites.
3. Repetitive Regions: Removed any variants falling in repetitive regions as defined by MSI Sensor. Reptitive regions are defined as areas having more than 6 or more repeat unit bases in a segment of the genome.
4. Manual Review: Finally all sites were manually reviewed in the integrative genomics viewer to finally confirm all complex indels.

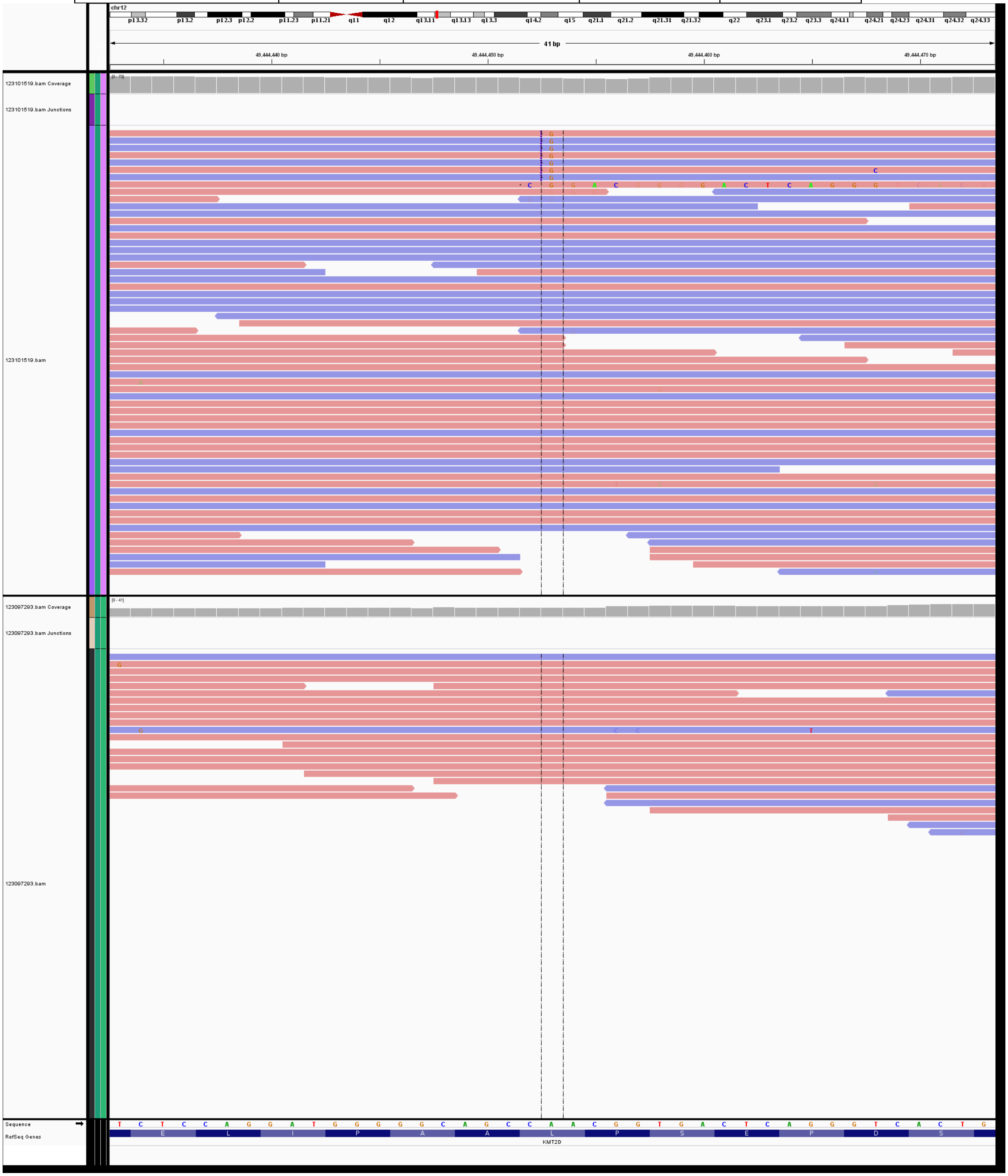
The screenshots below from the Integrative Genomics Viewer show the events described in the table above (tumor sample on top, normal sample on bottom).

Sample	**Gene**	**Chromosome**	**Start**	**End**
SJE2A043	*PAX5*	9	37002647	37002649

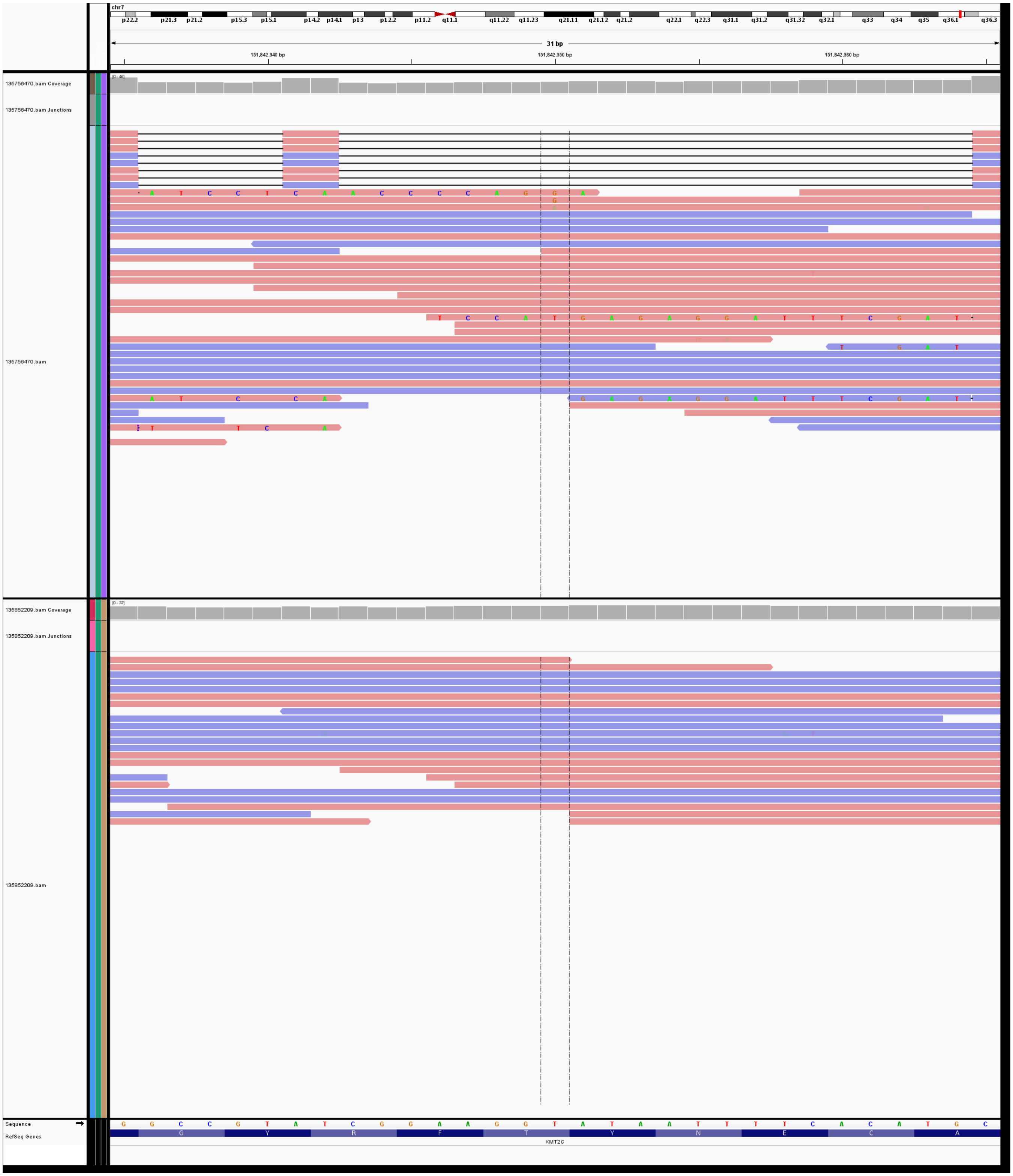


Sample	Gene	Chromosome	Start	End

SJERG031	*MLL2*	12	49444452	49444454

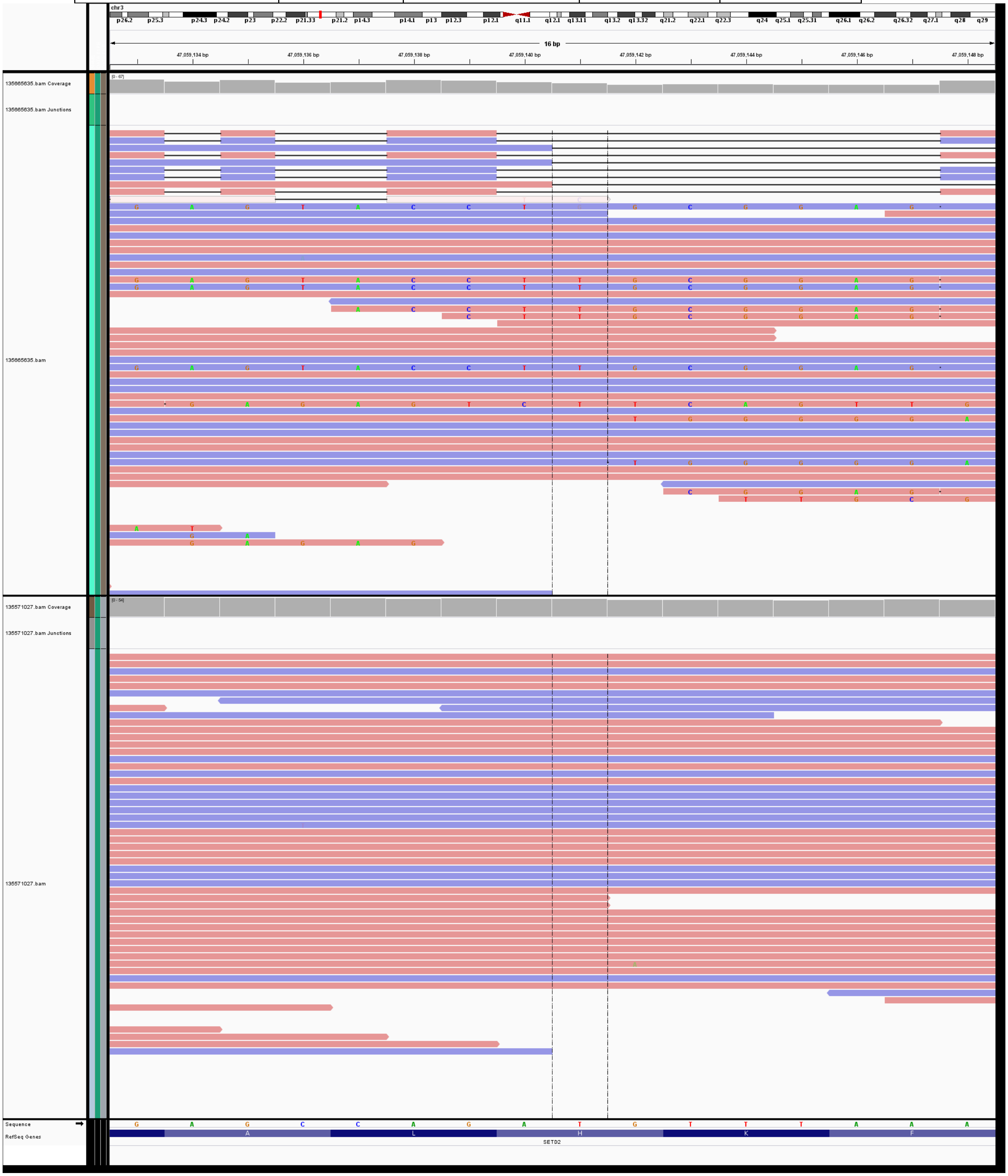


Sample	Gene	Chromosome	Start	End
SJERG020052	*MLL3*	7	151842335	151842365

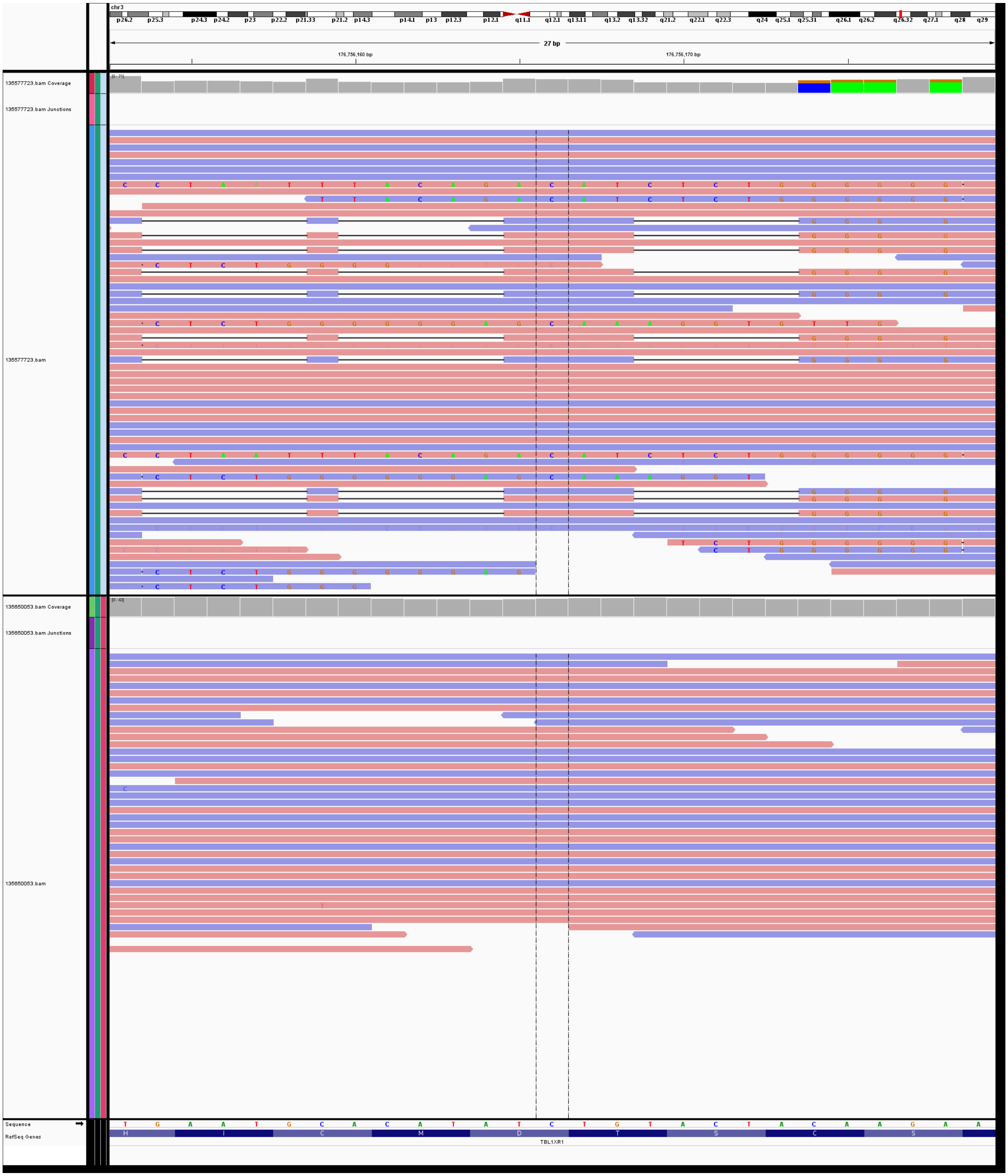


Sample	Gene	Chromosome	Start	End

SJERG020306	*SETD2*	3	47059133	47059166

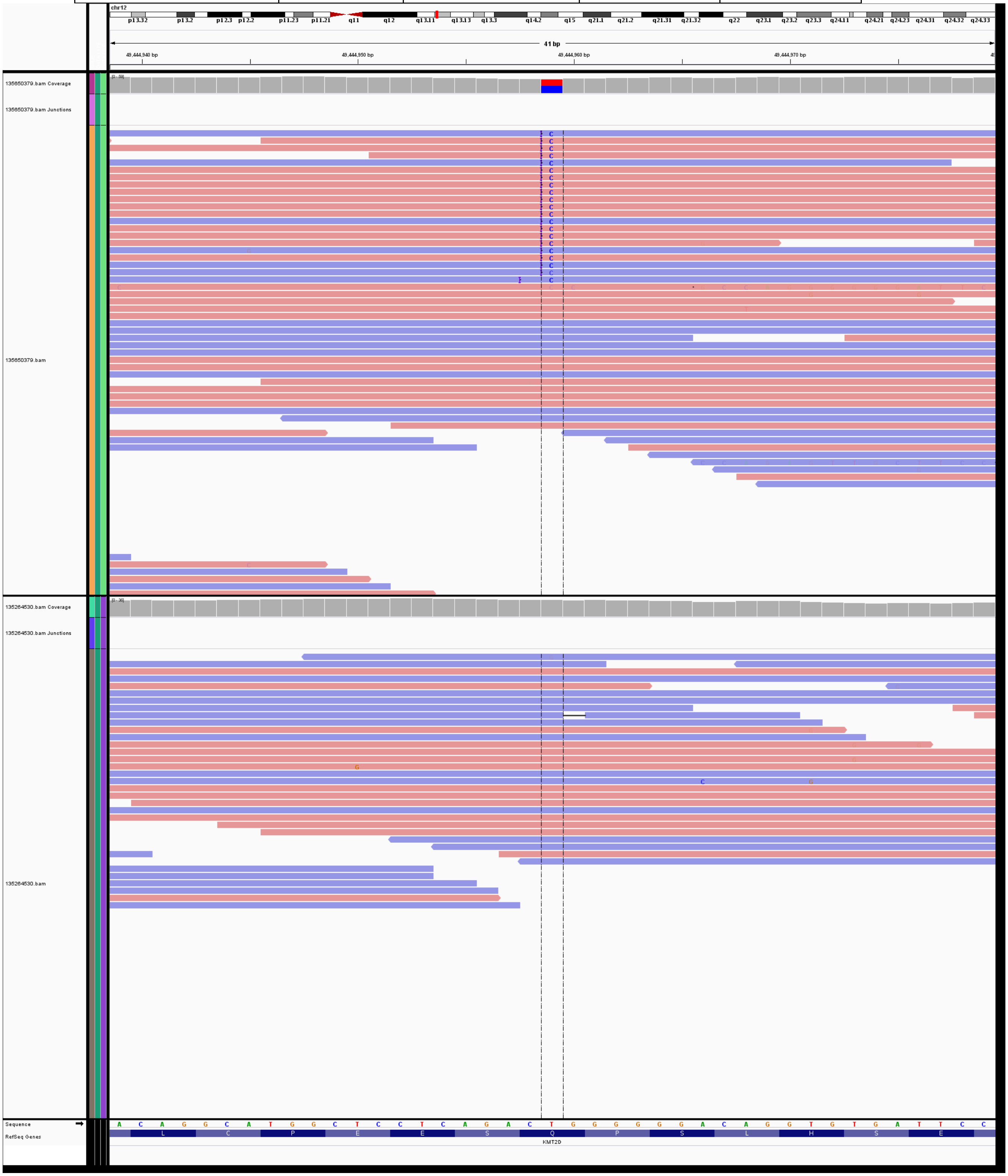


Sample	Gene	Chromosome	Start	End
SJERG020307	*TBL1XR1*	3	176756153	176756179

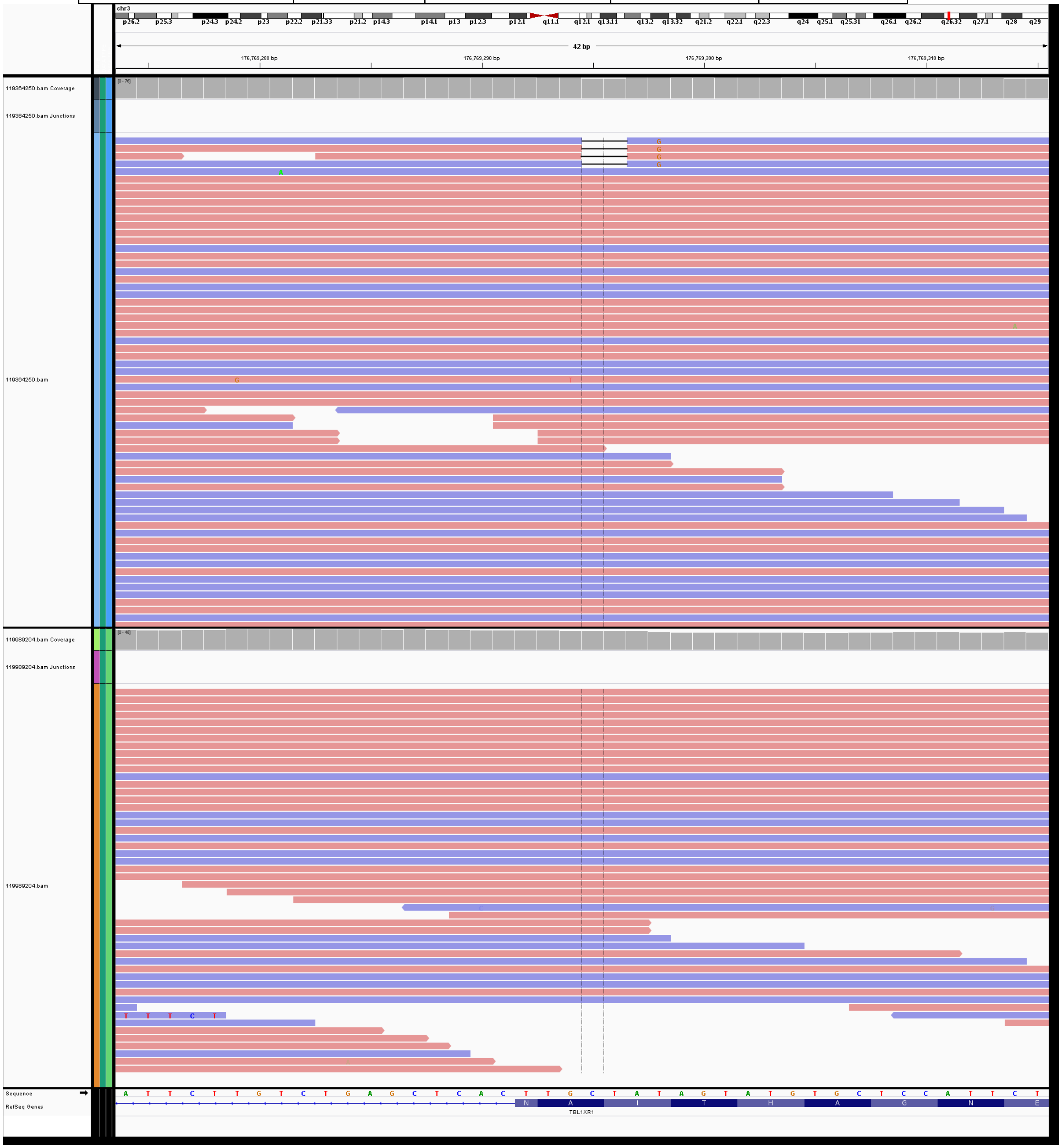


Sample	Gene	Chromosome	Start	End

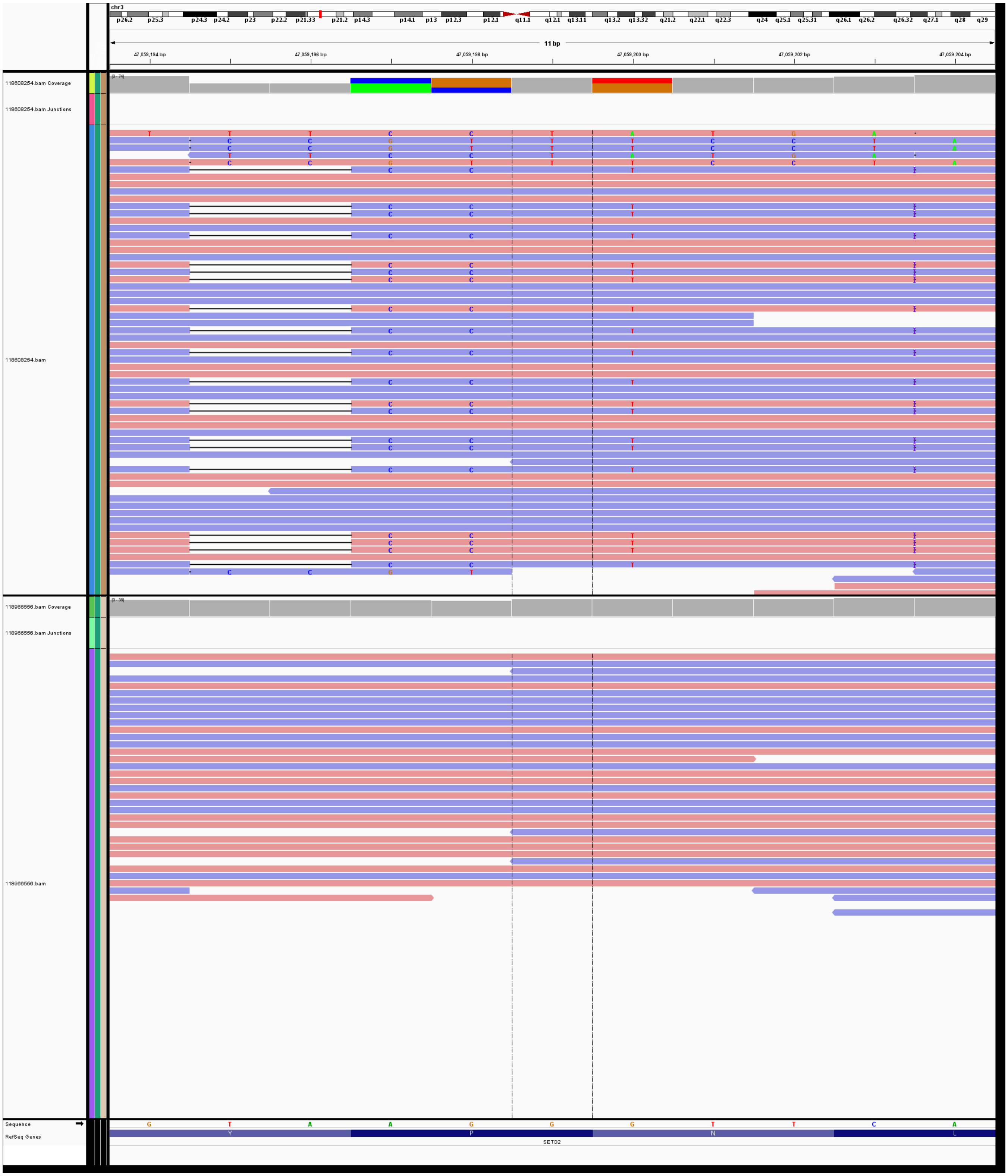
SJERG020309	*MLL2*	12	49444958	49444960



Sample	Gene	Chromosome	Start	End
SJETV001	*TBL1XR1*	3	176769294	176769299

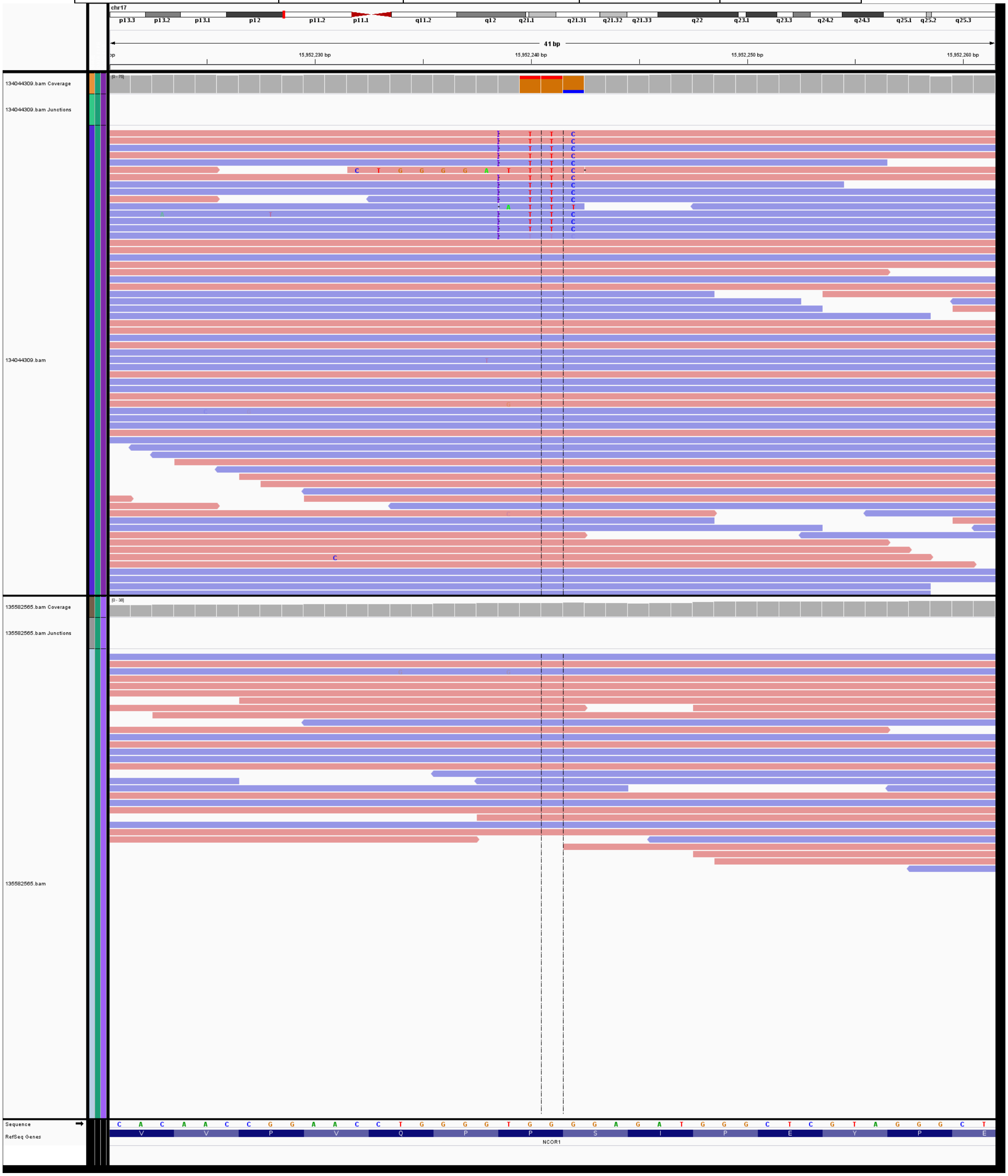


Sample	Gene	Chromosome	Start	End
SJETV024	*SETD2*	3	47059194	47059204

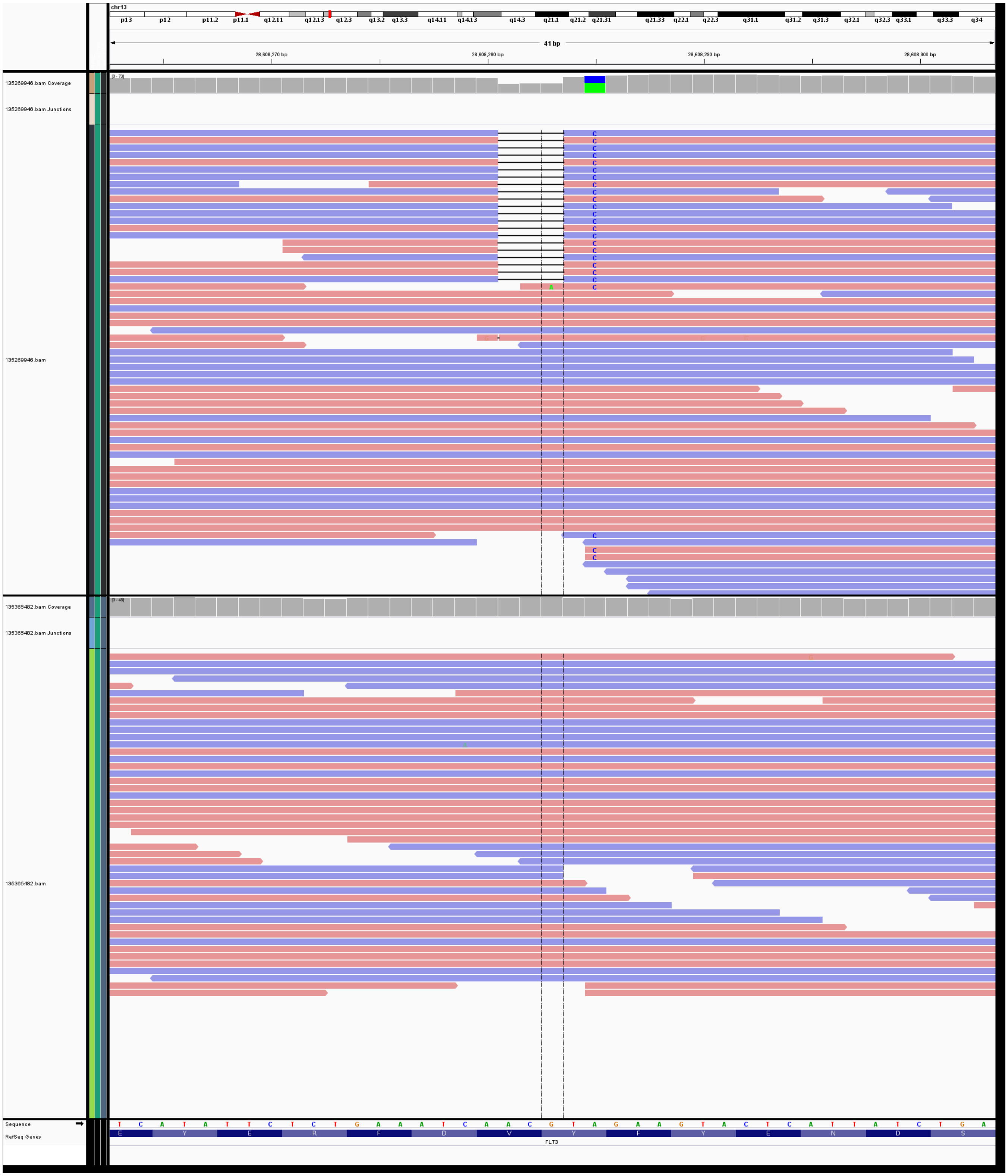


Sample	Gene	Chromosome	Start	End

SJHYPER005	*NCOR1*	17	15952238	15952243

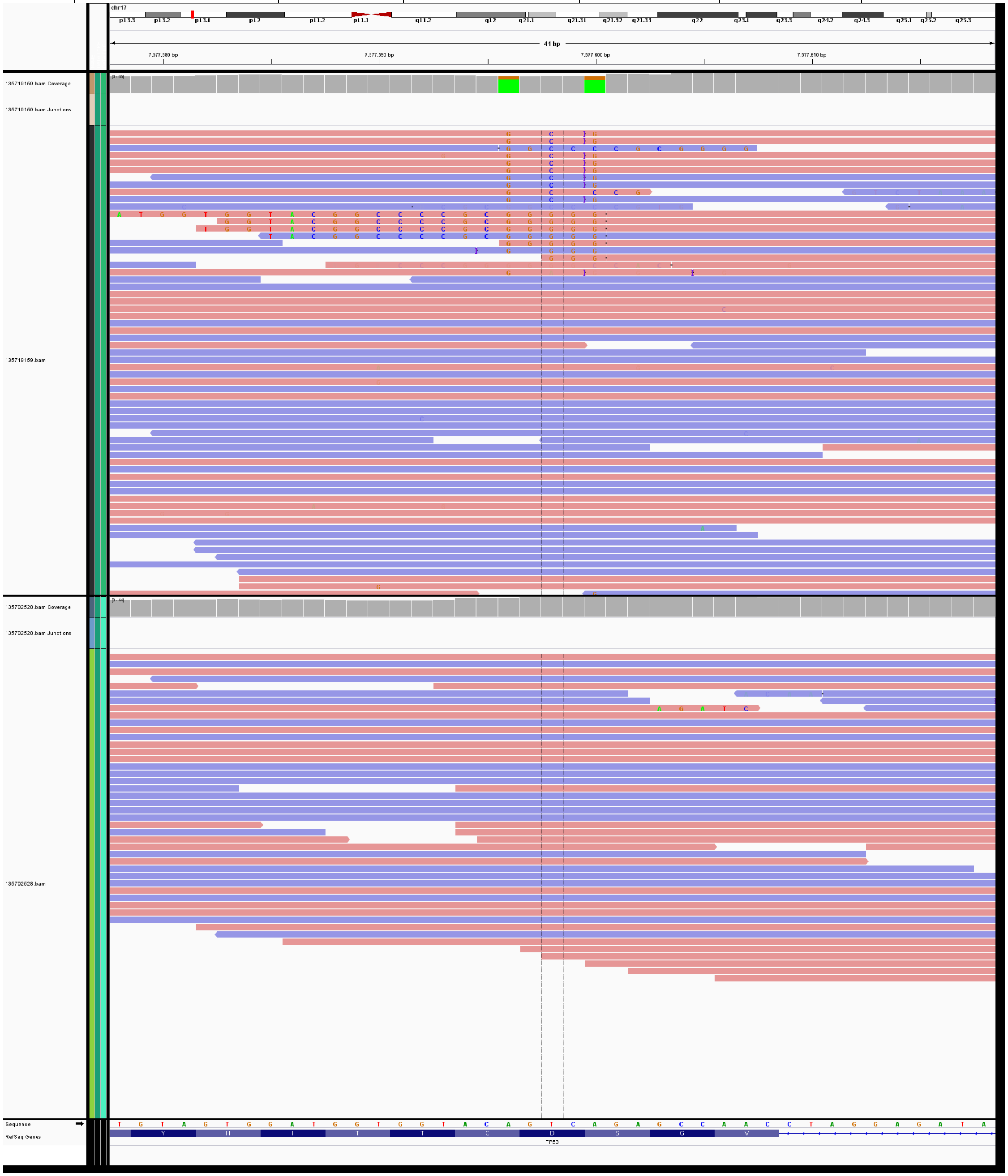


Sample	Gene	Chromosome	Start	End
SJHYPER007	*FLT3*	13	28608280	28608286

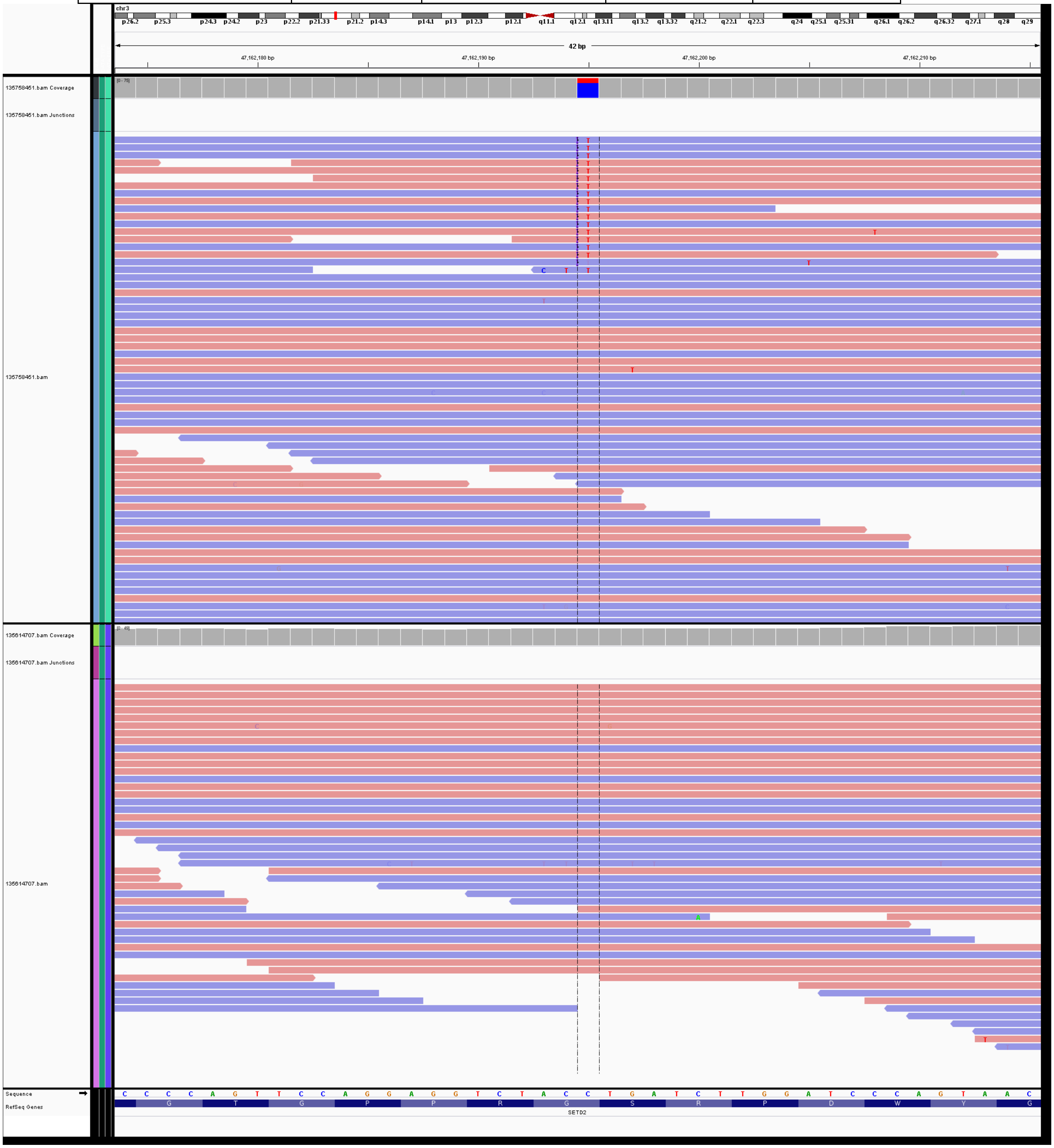


Sample	Gene	Chromosome	Start	End

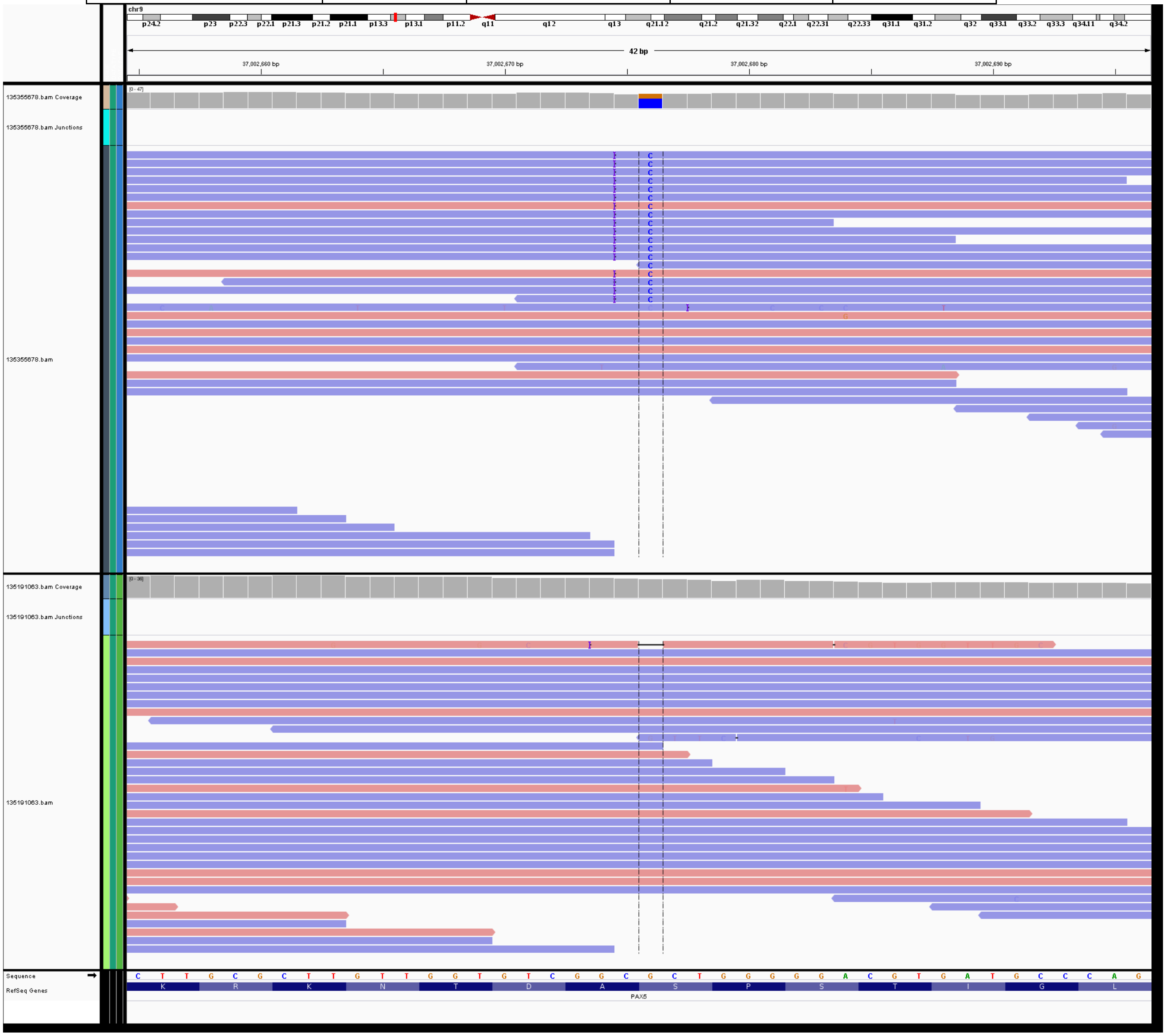
SJHYPER051	*TP53*	17	7577595	7577601



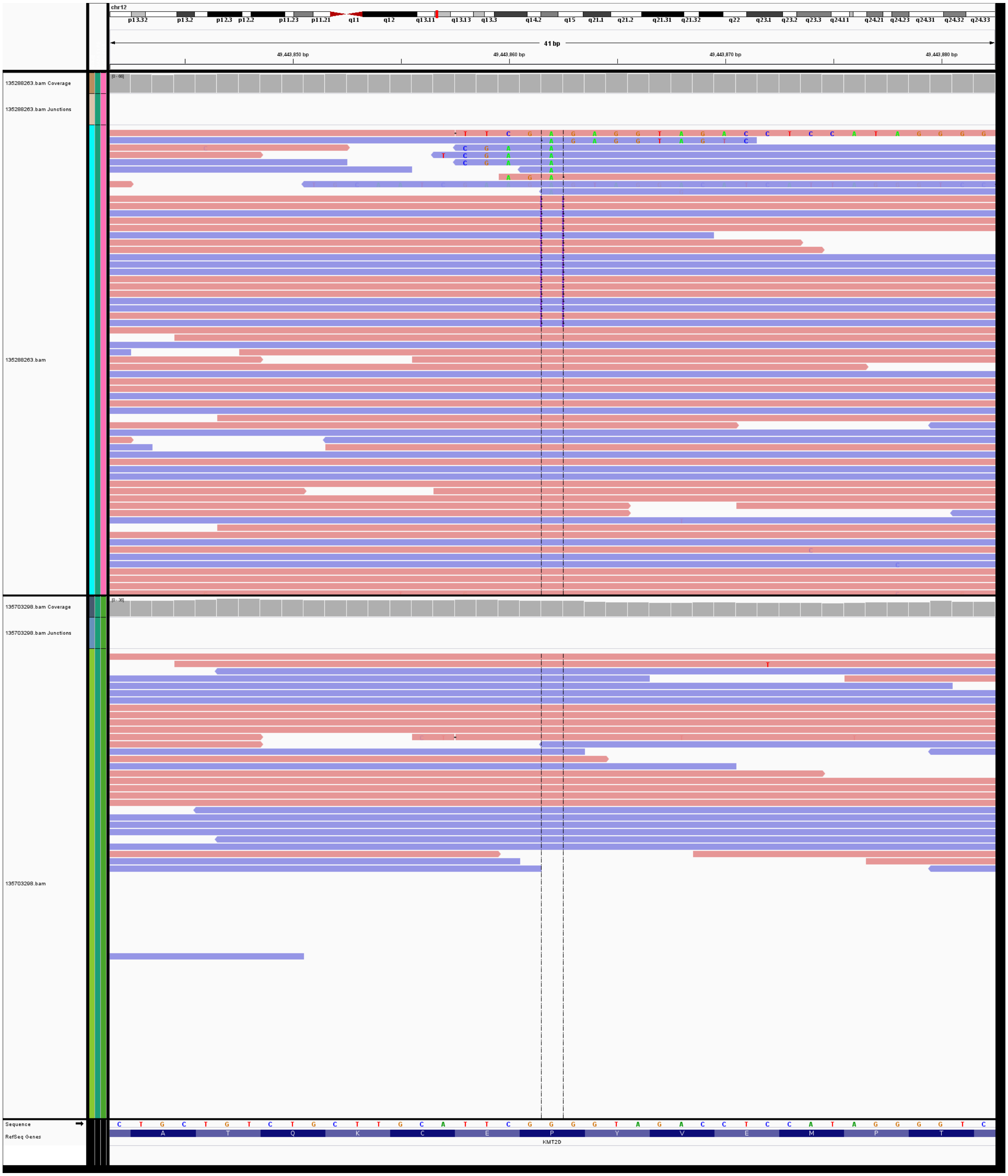
Sample	Gene	Chromosome	Start	End
SJHYPER097	*SETD2*	3	47162194	47162195


chr3
p26.2 p25.3 p24.3 p24.2 p23 p22.2 p21.33 p21.2 p14.3 p14.1 p13 p12.3 p12.1 q11.1 q12.1 q13.11 q13.2 q13.32 q21.2 q22.1 q22.3 q24 q25.1 q25.31 q26.1 q26.2 q26.32 q27.1 q28 q29
42 bp
47,162,180 bp
47,162,190 bp
47,162,200 bp
47,162,210 bp
135758451.bam Coverage
[0 - 75]
135758451.bam Junctions
135758451.bam
135614707.bam Coverage
135614707.bam Junctions
135614707.bam
Sequence
C C C C A G T T C C A G G A G G T C T A C C T G A T C T T G G A T C C C A G T A A C
RefSeq Genes
G T G P P R G S R P D W Y G
SETD2

Sample	Gene	Chromosome	Start	End
SJHYPER104	*PAX5*	9	37002675	37002676

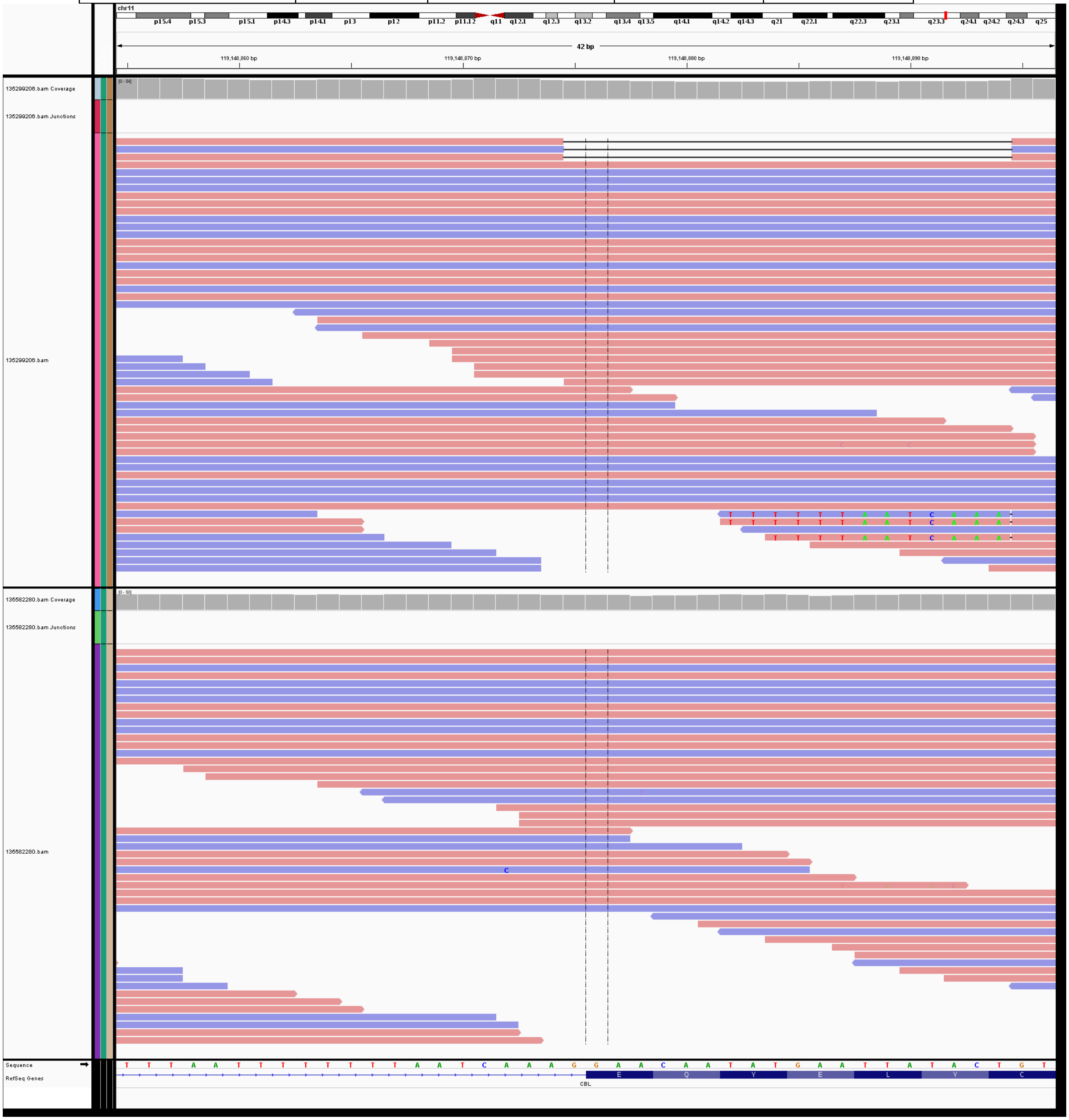


Sample	**Gene**	**Chromosome**	**Start**	**End**
SJHYPER108	*MLL2*	12	49443861	49443863

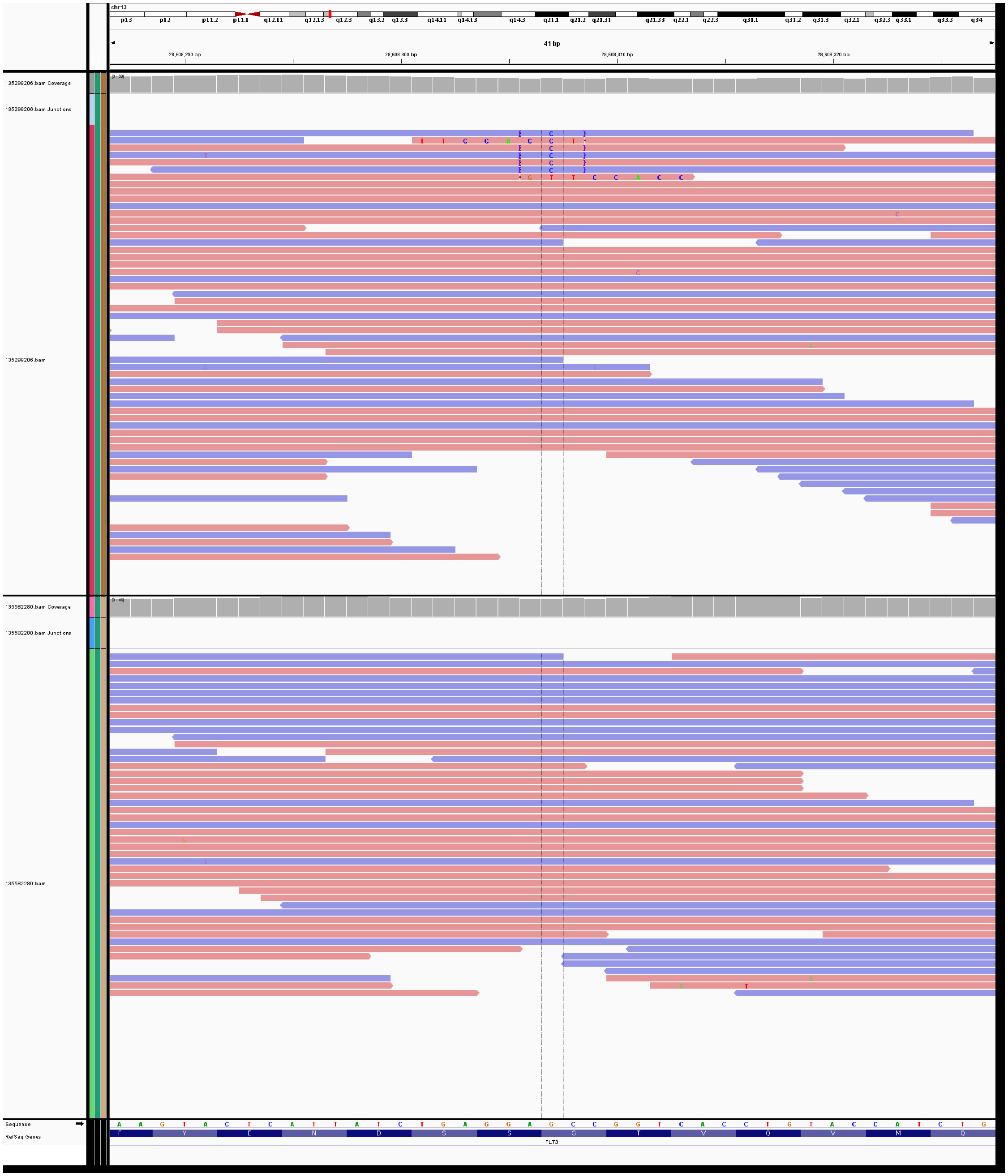


Sample	Gene	Chromosome	Start	End

SJHYPER111	*CBL*	11	119148875	119148900

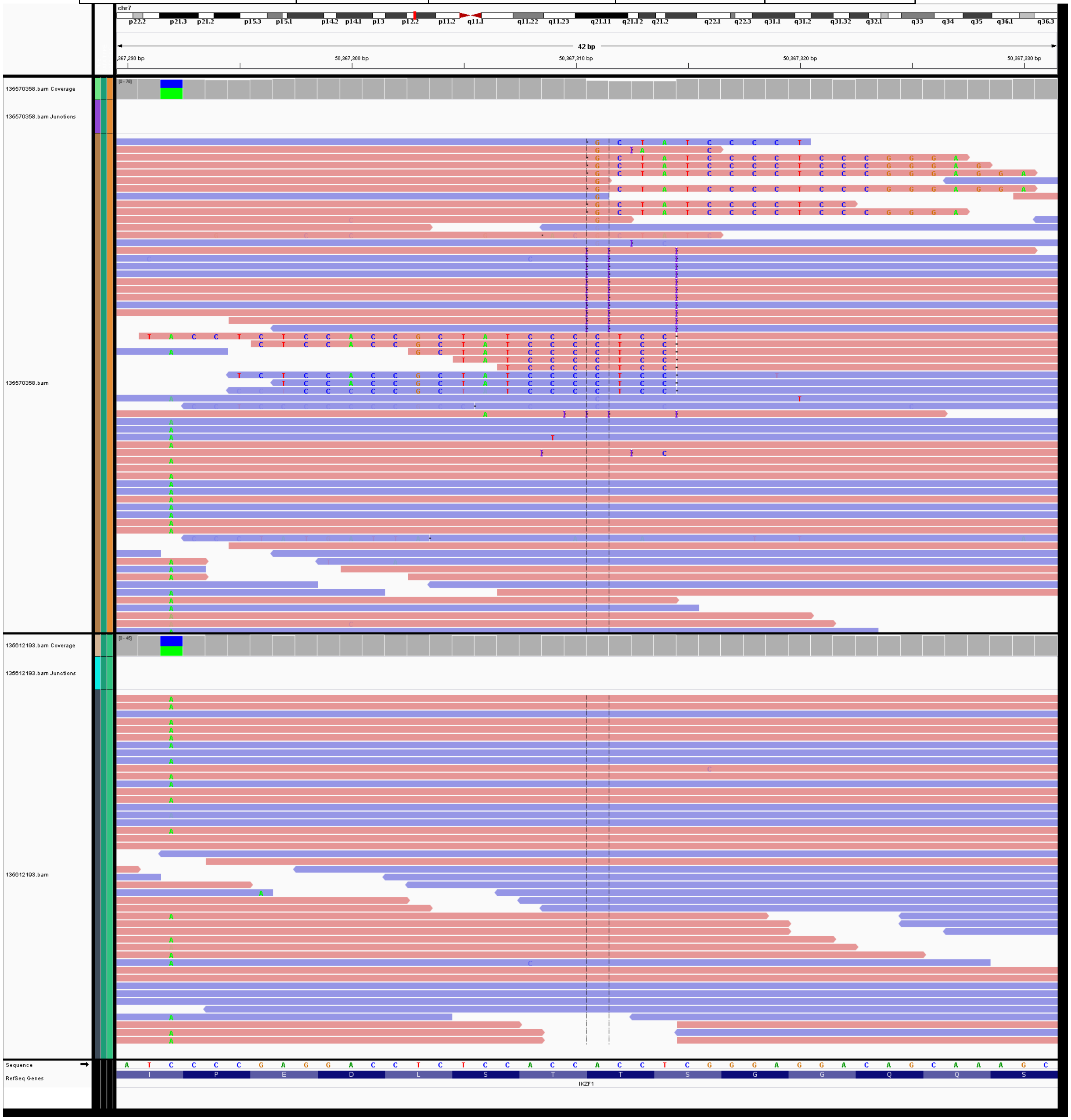


Sample	Gene	Chromosome	Start	End
SJHYPER111	*FLT3*	13	28608305	28608309

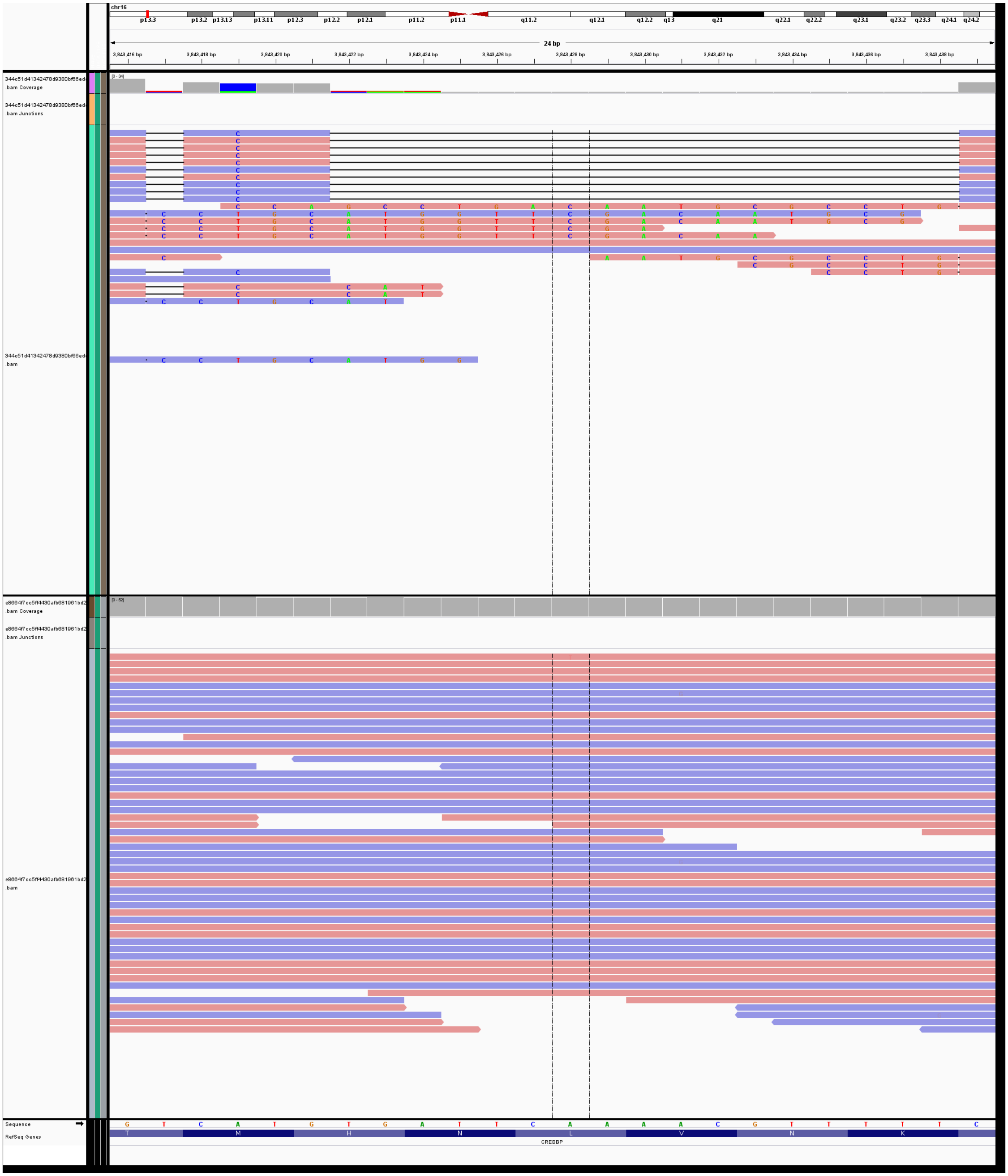


Sample	Gene	Chromosome	Start	End

SJHYPER227	*IKZF1*	7	50367310	50367314

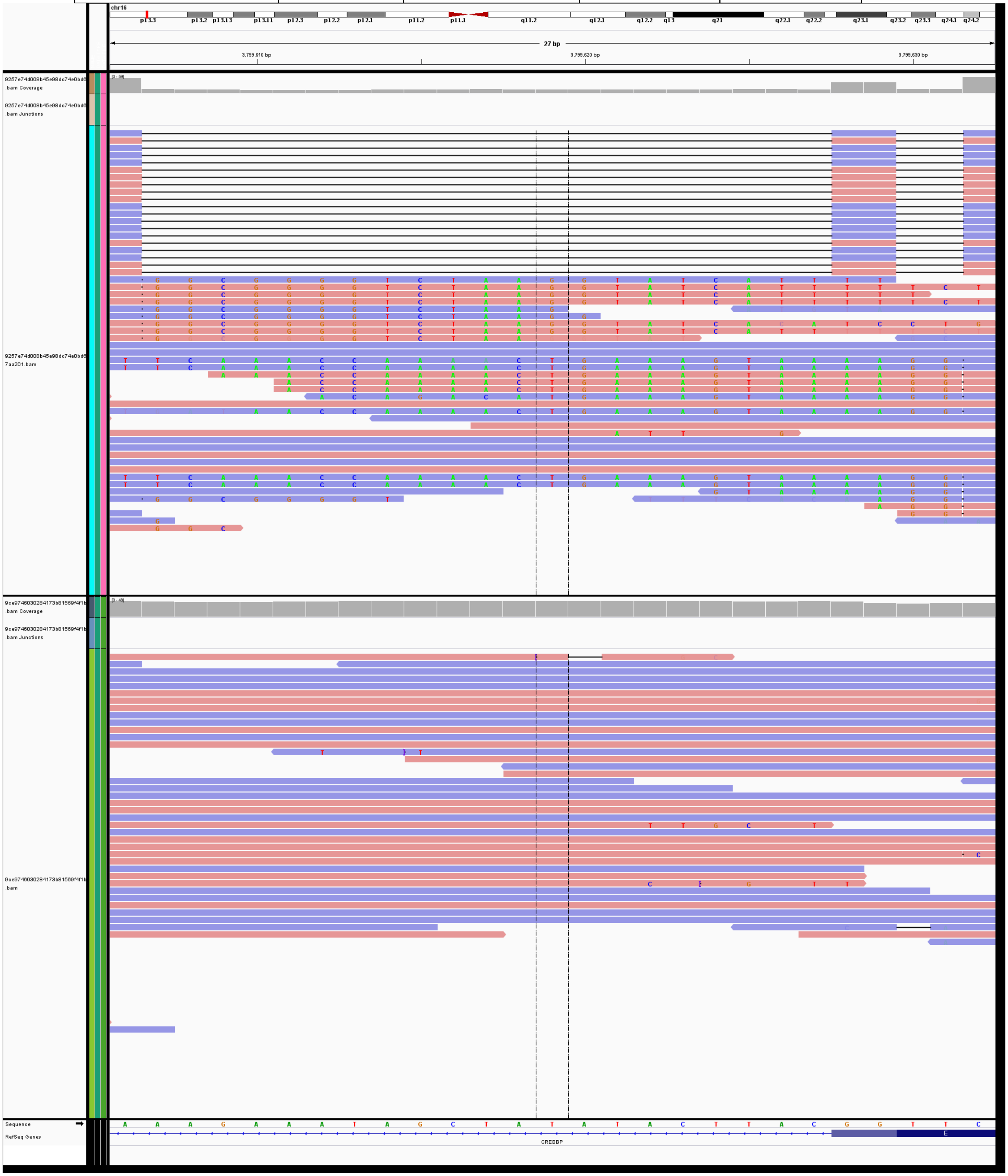


Sample	Gene	Chromosome	Start	End
SJHYPO002	*CREBBP*	16	3843416	3843439

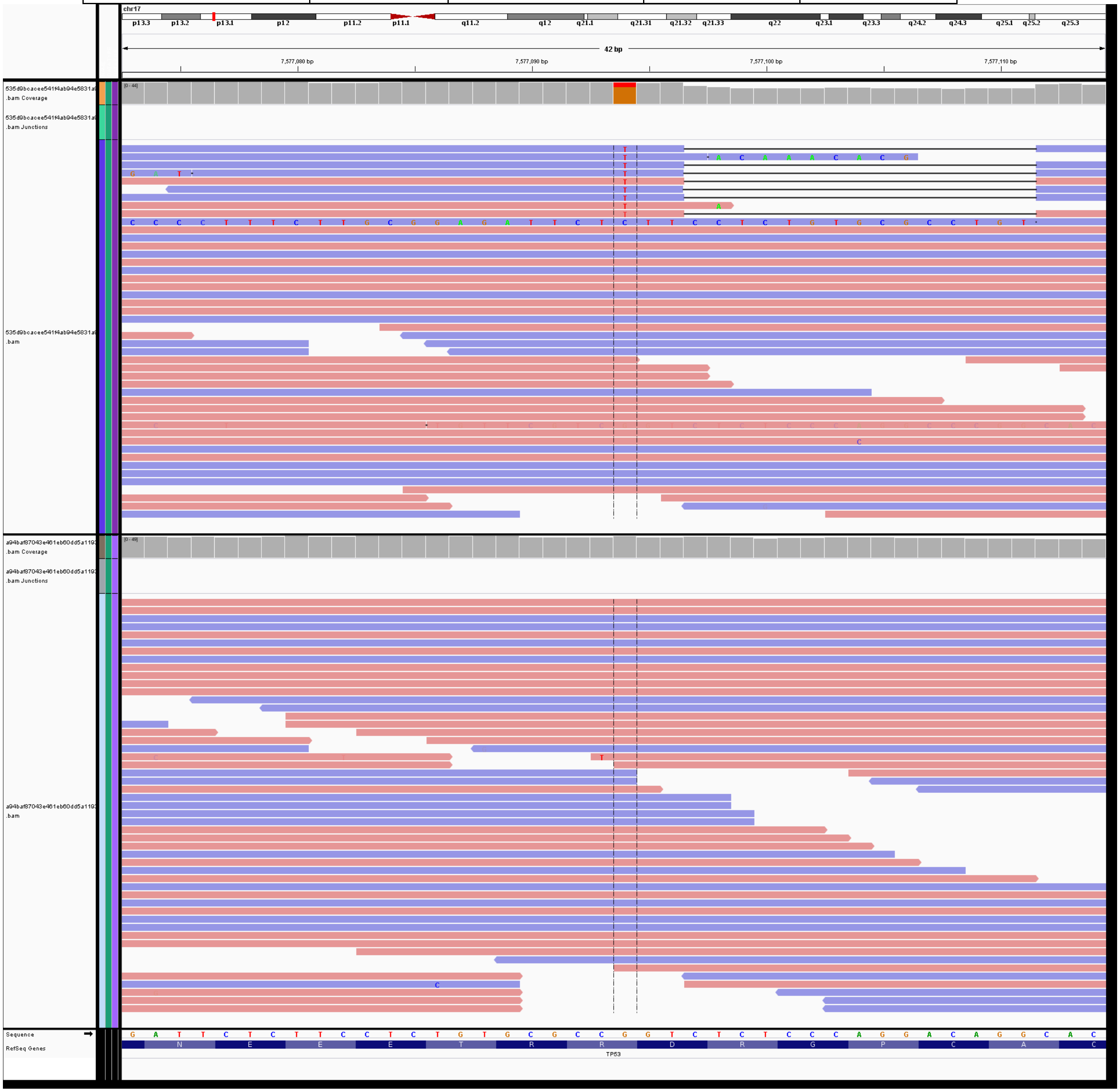


Sample	Gene	Chromosome	Start	End

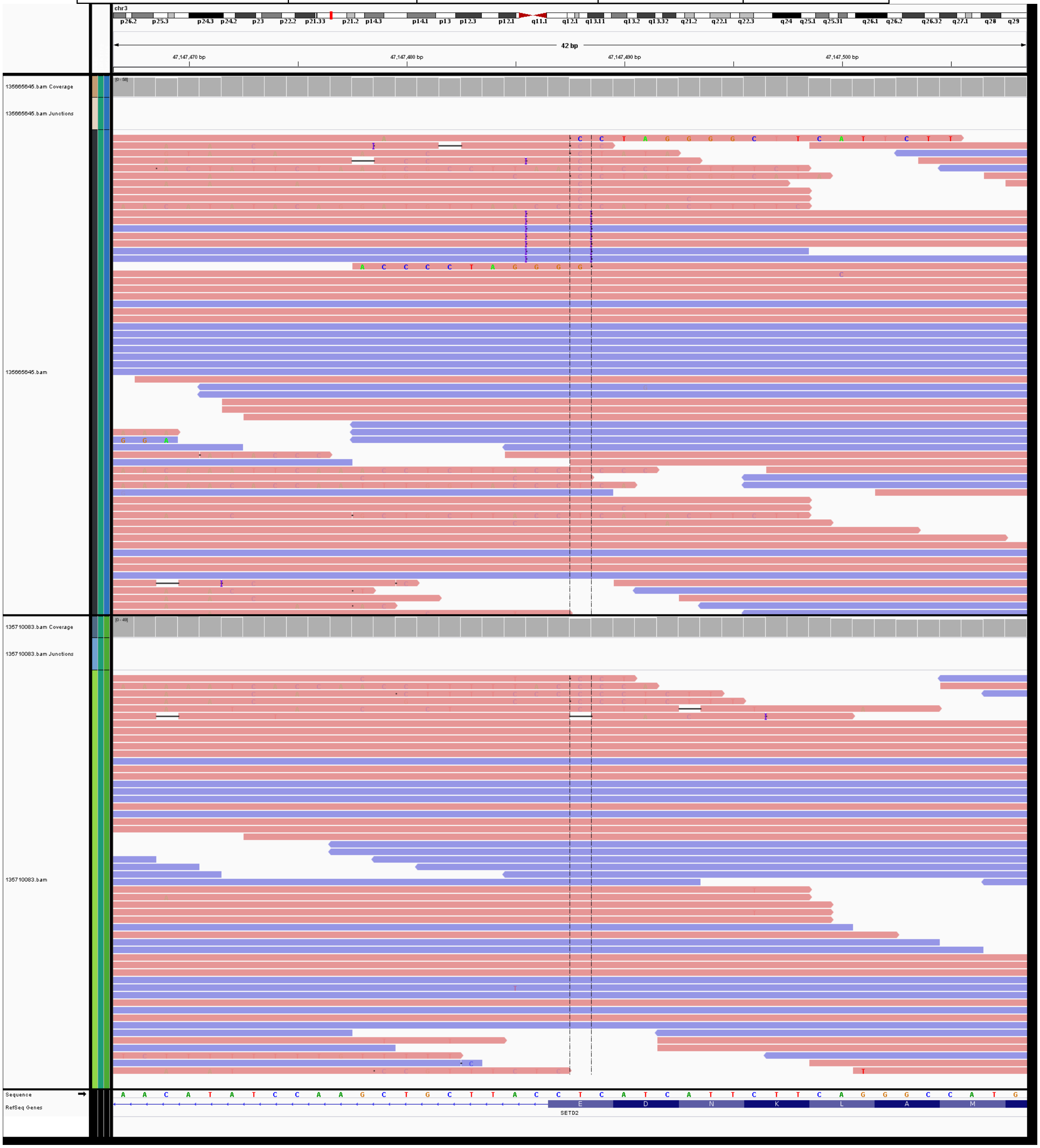
SJHYPO040	*CREBBP*	16	3799606	3799632



Sample	Gene	Chromosome	Start	End
SJINF013	*TP53*	17	7577093	7577111


chr17
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
42 bp
7,577,080 bp
7,577,090 bp
7,577,100 bp
7,577,110 bp
535d9bcacee541f4ab94e5831a9 .bam Coverage
[0 - 44]
535d9bcacee541f4ab94e5831a9 .bam Junctions
535d9bcacee541f4ab94e5831a9 .bam
a94baf87043e461eb60dd5a1193 .bam Coverage
[0 - 49]
a94baf87043e461eb60dd5a1193 .bam Junctions
a94baf87043e461eb60dd5a1193 .bam
Sequence
RefSeq Genes
G A T T C T C T T C C T C T G T G C G C C G G T C T C T C C C A G G A C A G G C A C
N E E E T R R D R G P C A C
TP53

Sample	Gene	Chromosome	Start	End
SJPHALL020043	*SETD2*	3	47147487	47147488



5.3 Children’s Discovery Institute: Utilizing CharGer to determine novel variants in undiagnosed pediatric cases

There is an urgent need for improved and quick diagnosis of birth defects. To address this concern our lab has developed CharGer (unpublished) to assess the pathogenicity of variants using and automated query of public data sources (example: Clinvar). Using patient mutation files (example: VCFs) derived from the mother, father and child, we can derive both de novo and compound heterozygous mutations.

When focusing on de novo variants we are interested in identifying spontaneous mutations in developmental genes compared to inherited mutations that are derived from both parents. We also require that mutations have a low frequency in the population (PM2). In this partial blind study we ran mutation files through the CharGer pipeline without knowing the associated phenotype. Variants of interest reported through the pipeline were reported to a physician at Washington University to determine the overlap between mutations called by an orthogonal gene sequencing service, GeneDx, and those that were missed.

To validate our automatic classification, we applied CharGer on a data set of 7 families with various birth defects in infants and children. For 6 families, CharGer confirmed a predicted “likely pathogenic” variant reported from GeneDx or followup functional studies from collaborators. Using variant call format (VCF) from GeneDx we derived mutations inherited from the mother, father and de novo variants in the child. Variants were filtered by high population allele frequency (>0.05) derived from 1000 genomes (1KG), Exome Aggregation Consortium (ExAC) and Exome Sequencing Project (ESP) and reported variants in our predefined gene list made up of 597 developmental genes (Saunders et al.,), 625 cancer genes and 11 additional genes identified in GeneDx positive analyses. CharGer identified 3/6 families as having 'Likely Pathogenic 'candidate variants and the remaining 3/6 as having 'Unknown Significance’. Furthermore, CharGer identified additional candidate variants for all 6 families, along with candidate variants for the 7th GeneDx negative family.

5.4 Discovery of Novel Fusions in Cancer Genomes

Contribution: I helped oversee the overall development of this project including figure development and writing.

Cell Rep. 2018 Apr 3;23(1):227-238.e3. doi: 10.1016/j.celrep.2018.03.050.

Driver Fusions and Their Implications in the Development and Treatment of Human Cancers {(Gao et al., 2018)}.

5.5 Evaluating Novel Germline Variants in Cancer Genomes

Contribution: I helped to manually review many germline variants and helped significantly with the BRCA1 homologous recombination assay for the below publication.

Patterns and functional implications of rare germline variants across 12 cancer types (Lu et al., 2015).

5.6 Effect of Blood Somatic Mutations in PPM1D on TP53 and Cell Cycle Arrest

In 2014, our lab was one of the first to identify blood specific somatic mutations in elderly patients without any adverse hematopoietic diseases (Xie et al., 2014). After, performing a similar analysis on additional samples, we identified one novel recurrently mutated gene with a significant enrichment of blood-somatic mutations. PPM1D is a phosphatase involved in the dephosphorylation of DNA damage response genes including TP53. While this gene has been associated with breast and ovarian cancer predisposition, it has still not yet been reported in hematological malignancies. C-terminal truncating mutations are predicted to hinder the activation of DNA damage response proteins by enhancing the dephosphorylation activity of *PPM1D*. Figure X highlights blood specific mutations identified across 2,278 patients.

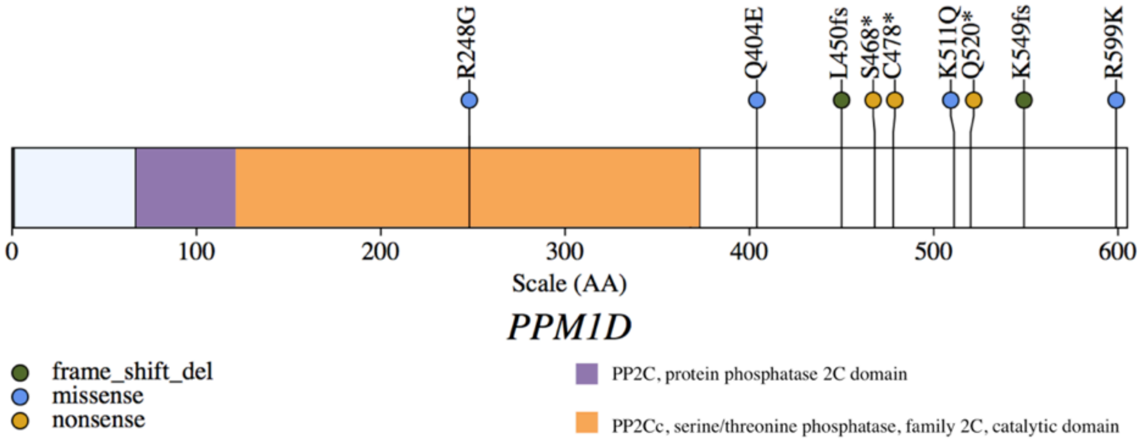


Figure A.1: Blood specific somatic mutations identified in *PPM1D*

We hypothesized blood specific mutations in *PPM1D* disrupt *TP53* phosphorylation. To evaluate this hypothesis, we determined the level of p53 Ser15 phosphorylation of PPM1D Mutants after radiation treatment (Methods). Under conditions of DNA damage (ex. Radiation) TP53 is phosphorylated and will activate downstream genes to initiate the DNA damage response pathway. Normally PPM1D will dephosphorylate TP53, thereby inactivating the DNA damage response pathway. Without UV treatment you would expect to see low levels of TP53 phosphorylation. As shown previously in our lab, PPM1D mutants further decreased Ser15 phosphorylation levels (stronger than WT), suggesting PPM1D mutants are more “active” than the wild type. Under wild type conditions, we expect to see a larger number of cells undergoing apoptosis after UV treatment. Under normal conditions TP53 should be phosphorylated and activated thereby halting the cell cycle process in response to DNA damage. To evaluate this, we can use propidium iodide (PI) staining to evaluate cells in various stages of the cell cycle. PI dyes DNA and binds to DNA in proportion to the amount of DNA present in the cell. Cells in S phase will therefore have more DNA than cells in G1 and will take up proportionally more dye (more fluorescence) with a higher amount of DNA content. Cells in G2 should be twice as bright as cells in G1 phase. Using FACS and the PI dye, we can conditionally sort these cells using fluorescence.

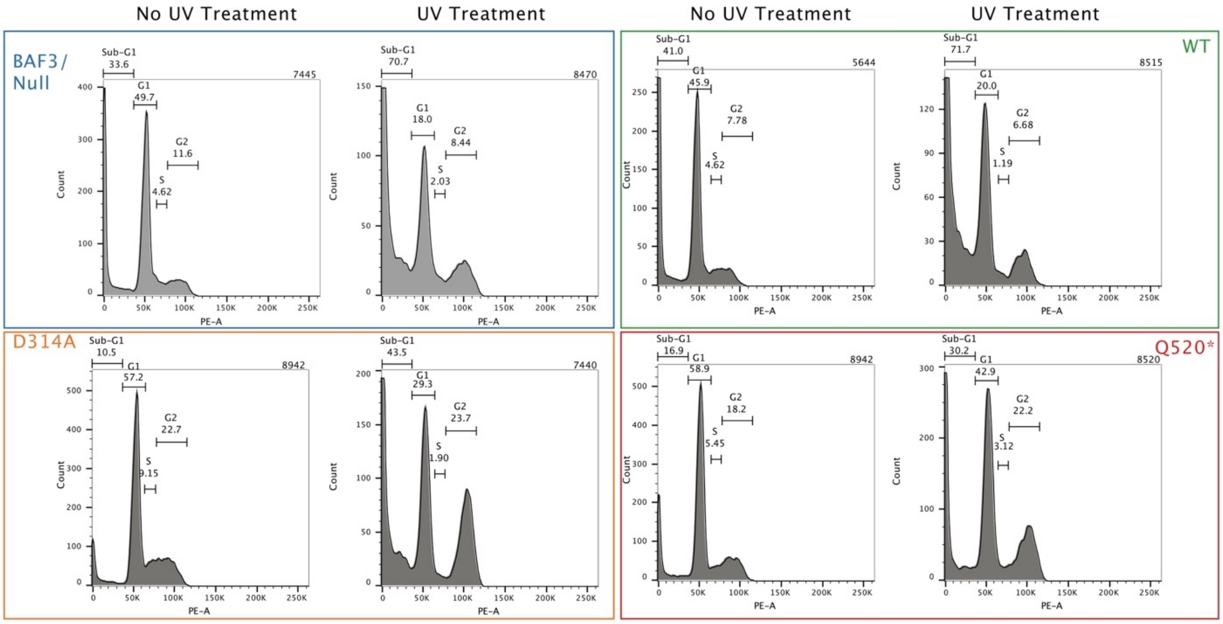


Figure A.2. FACS sorted cells from mutant and wildtype treatments after UV induction.

Table A.2. Predicted results for assay

Strain	TP53 Activity	Cell cycle changes after UV treatment
BAF3/Null	ON	More cells undergoing apoptosis
BAF3/WT-PPM1D	On-ish/Off-ish	Less cells undergoing apoptosis (less dividing cells)
BAF3/PPM1D-D314A (phosphate dead)	ON	More cells undergoing apoptosis

BAF3/PPM1D-Q520* (activating)	More OFF-ish	Less cells undergoing apoptosis

The results from Figure A.2 suggest that blood-specific PPM1D mutants tend to have higher phosphatase activity against TP53, and prevented the cells from apoptosis after DNA damage, suggesting that PPM1D mutations were potentially involved in cell proliferation regulation. Under conditions of BAF3/Null and the D314 phosphatase dead cell line, we would expect to always see active TP53 which should successfully halt the cell cycle process therefore more cells in S/G1. With the Q520 mutant, we expect to see higher expression of PPM1D thereby inactivating TP53 and more cells in G2 phase (mitotic cells-dividing). TP53 is unable to stop cells from entering the cell cycle even after DNA damage.

Protocol for Cell Cycle Analysis

Day 1

Prior to starting make sure cells are confluent. Move media to water bath.

1. Transfer cells from flasks to 15 mL tubes. BAF3 cells are a cell suspension cell line (so most cells are in the media floating around).
2. Spin down 200 rpm for 5 minutes
3. Remove supernatant.
4. Resuspend pellet in 1 mL of appropriate media.
5. Count Cells
    a. 1:5 dilution

   b. Add 80 ul of Trypan blue to small tube.

   c. Add 20 ul of sample to small tube.

   d. Calculate # of cells using hematocytometer. Need to have 10^6 cells

6. Spin down resuspension one more time.
7. Remove supernatant.
8. Add 1000 ul media.
9. Label 6 well plates. One should be UV treated and the other should be no UV.

   Set up experiments in duplicates or triplicates if possible.

   a. BAF3

   b. WT

   c. D314A

   d. Q520

10. Add 2 mL of media to each well.
11. Add calculated amount of each sample to the appropriate well.
12. Transfer remaining cells back to cell culture flask with 6 mL of media.
13. Incubate Errbody!

Day 2

Wait 24 hours and irradiate the cells with 6 Gy UV

Day 4

Wait 48 hours and harvest cells for fixation

Fixation

1. Remove cells from 6 well plate and move to 15 mL tubes.
2. Centrifuge at 500 rpm for 5 min

3. Remove supernatant and resuspend cells in 1 mL of media.
4. Count cells.
5. Resuspend at 2x10^6 cells in 1 mL ice cold buffer.
6. Vortex gently, slowly adding the cell suspension dropwise to 9 mL of 70% ethanol in a 15 mL polypropylene centrifuge tube.
7. Store at 4 degrees C for 24 hours.

Day 5

8. Centrifuge cells at 200 x g, 10 min, 4 degrees Celsius
9. Resuspend pellet in 3 mL cold PBS and transfer to tubes.
10. Wash cells with cold PBS.
11. Resuspend cells in 300-500 ul PI triton staining solution to 10 mL of 0.1 % (v/v) Triton X-100 (Sigma) in PBS add 2 mg DNAse-free RNAse A.
12. Incubate 37 degrees Celsius for 15 min or for 30 min at 20 degrees Celsius.
13. Transfer tubes to ice or store at 4 degrees Celsius protected from light.
14. Acquire data on flow cytometer within 48 hours.
15. Perform FACS on department of pathology and immunology

5.7 Discovery of Truncation Mutations Leading to Protein Alterations

Studying the allele specific expression of nonsense mutations across genes genome wide will broaden our understanding of rules by which nonsense mediated decay can effectively degrade a transcript. To study mutation enrichment and degradation efficiency across cancer types, we propose a novel scoring method using RNA-Sequencing and whole exome sequencing variant allele fraction ratio as a proxy for allele specific expression. By comparing case expression to a control dataset in each cancer type, our analysis has uncovered an enrichment of nonsense mutations predicted to escape nonsense mediated decay (NMD) and a subset showing evidence of N and C terminally truncated proteins. Identifying variants that target a transcript for degradation via NMD or produce a transcript that could be translated to a truncated protein are both clinically pathogenic. We further expanded our investigation to consider enriched mutations across mutation types, including missense, silent and splice site variants, to identify activating and inactivating truncation mutations.

Of note our analysis identified a subset of mutations in a tumor suppressor, STK11, utilize a downstream start codon to create N terminally truncated proteins lacking the N terminal localization sequence in lung cancer samples. We propose a novel scoring method using RNA and DNA variant allele fraction ratio as a proxy for allele specific expression to study mutation enrichment and degradation efficiency across cancer types.

Recurrent protein isoforms are shown to be present in prostate, lung, hepatocellular and tumor samples, but only a small percentage of samples in a few cancer types are shown to harbor mutations in splicing factors(Brooks et al.; Rajan et al.; Zhang et al.). Since the dominant isoform present in many tumors are not attributed to mutations in splicing factors, this directs us to focus on introduced genomic variants that are contributing to altering splicing patterns. Furthermore, since one-third of alternative splicing events in human genes are thought to cause NMD(Brogna and Wen; Lewis et al.; Ni et al.; Pan et al.; Venables; Weischenfeldt et al.; Weischenfeldt et al.), it is vital that we can predict nonsense mutations that will and won't be degraded by the introduced mutation.

One-third of SAVs are estimated to introduce a premature termination codon (PTC), which could lead to dominant negative or gain of function effects. Annotated nonsense mutations are predicted to create premature termination codons (PTC) in the resulting transcripts, which are predominantly degraded by the Nonsense Mediated Decay (NMD) pathway. The general rule of thumb is that PTC's located at least 50-55 bp upstream of the last exon-exon junction drive strong NMD, whereas those outside of this criteria are predicted to escape the degradation process. Finally, as noted in previous studies, N terminally truncated proteins can lead to translation re-initiation at a downstream codon, thereby bypassing NMD and creating a product with residual function. This exception to NMD is known for few genes, but has yet to be characterized in cancer and genome wide.

Understanding how SAVs can lead to alternative isoforms in tumor samples and determining if the new transcripts become a target of NMD is still an open area of study.

Identifying variants that target a transcript for degradation via NMD or produce a splicing product that could be translated to a truncated protein are both clinically pathogenic. A pan-cancer analysis of SAVs can expand the current paradigm about the exceptions and rules by which NMD and splicing are altered in cancer.

Dataset and mutation distribution

Stringent filters (Supplementary Methods) were used to create high quality mutation calls for 18 cancer types: breast adenocarcinoma (BRCA), bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon and rectum adenocarcinoma (COADREAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), low grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), Stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), uterine carcinosarcoma (UCS) and uterine corpus endometrioid carcinoma (UCEC).

Filters: We collected nonsense mutations predicted to introduce termination codons into a transcript of interest. We performed RNA-Seq and DNA-Seq readcount analysis to determine sites that had at least 20X coverage. From the high coverage nonsense mutation gene set, we collected silent and missense mutations, performed readcount analysis and filtered out sites with less than 20X coverage. Finally sites with both copy number and gene expression data available through the Broad GDAC Firehose were collected for further analysis.

Our final mutation dataset consisted of 88,845 silent, nonsense and missense mutations in 5,023 genes across 16 cancer types from 3,129 samples.

Dataset and mutation distribution

Stringent filters (Supplementary Methods) were used to create high quality mutation calls for 18 cancer types: breast adenocarcinoma (BRCA), bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon and rectum adenocarcinoma (COADREAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), low grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), Stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), uterine carcinosarcoma (UCS) and uterine corpus endometrioid carcinoma (UCEC). We collected 11,952 nonsense mutations predicted to introduce a premature termination codon into the transcript of interest. We additionally collected silent and missense mutations in all genes with nonsense mutations for comparison. Our final mutation dataset consisted of silent, nonsense and missense mutations in 6,107 genes across 18 cancer types in 3,863 samples. All mutations were collected in copy number neutral sites with at least 20 reads spanning the site of interest in RNA-Sequencing data and DNA-Sequencing.

Identification of activating nonsense mutations in copy number neutral data

To evaluate the effect of nonsense mutations on stability, we used a novel method to differentiate between nonsense mutations degraded and escaping degradation as measured by next generation sequencing data. We chose to leverage variant allele fractions (VAFs) measured by DNA and RNA sequencing of the tumor samples to define the relationship between the genomic position and transcriptome effect. One would expect a comparative relationship between RNA and DNA VAF if allelic content and transcription were directly proportional. But in practice there are a number of variables that can cause a deviation from this proportional relationship including post transcriptional modifications such as RNA degradation.

By comparing the RNA to DNA VAF ratio across exons for nonsense, missense and silent mutations in copy number neutral regions we can evaluate the positional effect of nonsense mutations. This comparison between mutation types across exons and genes will highlight exons of interest that have a higher than expected RNA VAF when compared to DNA VAF and can potentially be used as a proxy to measure sites that escape nonsense mediated decay.

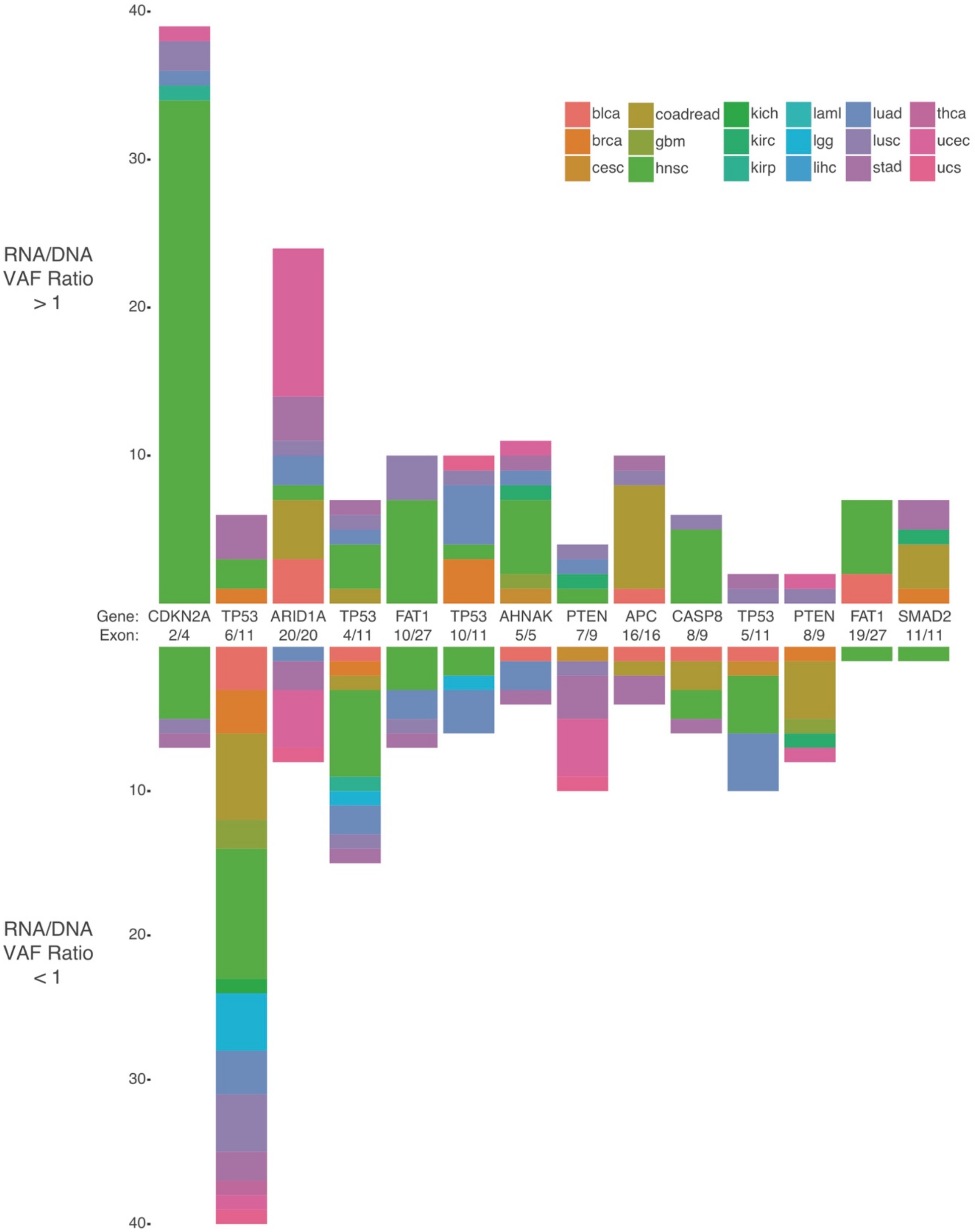


Figure A.3. Bar plot of total sites with RNA/DNA VAF Ratio greater than (top) and less than one (bottom) separated by gene.

Figure A.3 highlights the top 14 exons found to have the largest number of nonsense mutations and their RNA/DNA VAF Ratios. Many exons predicted to contain nonsense mutations that escape NMD are present in the last exon of the gene including ARID1A, AHNAK, APC, SMAD2, as expected. Of the 2,538 exons from 2,194 genes with at least one nonsense mutation with an RNA/DNA VAF greater than one, 1,069 of the exons are classified as the last exon of the gene of interest. This confirmed our first main finding that many nonsense mutations in cancer act as expected and are degraded effectively by the NMD pathway. Interestingly, there are still a large number of nonsense mutations that are present at a high variant frequency in RNA and DNA, suggesting evasion of the degradation pathway. We focused on a set of tumor suppressor genes that contained a large number of nonsense mutations with high RNA VAF including CDKN2A, TP53, SMAD2, CREBBP, ARID1A, PTEN, MAP3K1, KMT2C and KMT2D.

A subset of our mutations had available normalized RSEM gene expression data to compare the gene expression of the case to associated controls. For each of the genes mentioned above, we compared the case expression of the gene of interest to control samples in the same cancer type, and also compared genes within the same pathway to identify downstream genes with altered expression due to the activating truncation mutation. After identifying potential activating mutations, each mutation was annotated using TransVar(Zhou et al.), a multilevel variant annotator.

One sample had a mutation in exon 1 of CDKN2A in HNSC and was found to have much higher expression than control samples. CDKN2A expression leads to activation of INK4A and ARF which in turn inhibits MDM2, CDK4 and CDK6. Gene expression of the downstream genes showed significantly decreased expression of MDM2 and CDK6 but lower expression of TP53 and RB1 which should have increased expression with decreased expression of MDM2 and

CDK6. A closer look at the mutations present in this sample revealed a frame shift insertion in TP53 and upon inspection of the RNA-Seq showed very few reads spanning TP53 exons and present in a copy number neutral region and similar findings for RB1 (but without a mutation).

For each gene of interest we identified missense, silent, nonsense and splice site variants deemed to be activating mutations by our gene expression permutation test. We used a lolliplot to visualize the predicted activating mutations within each functional domain. Our findings identified a number of “hotspots” where mutations cluster that create similar truncating activating mutations.

We defined genes as expressed if they had a RSEM value greater than the lower quartile of the distribution of RSEM control values. Furthermore, we focused on sites that have RNA and DNA VAF ratio greater than 1 and RSEM value greater than the lower quartile for the distributed RSEM control values.

Measuring allele specific expression by integrating RNA and DNA variant allele fractions

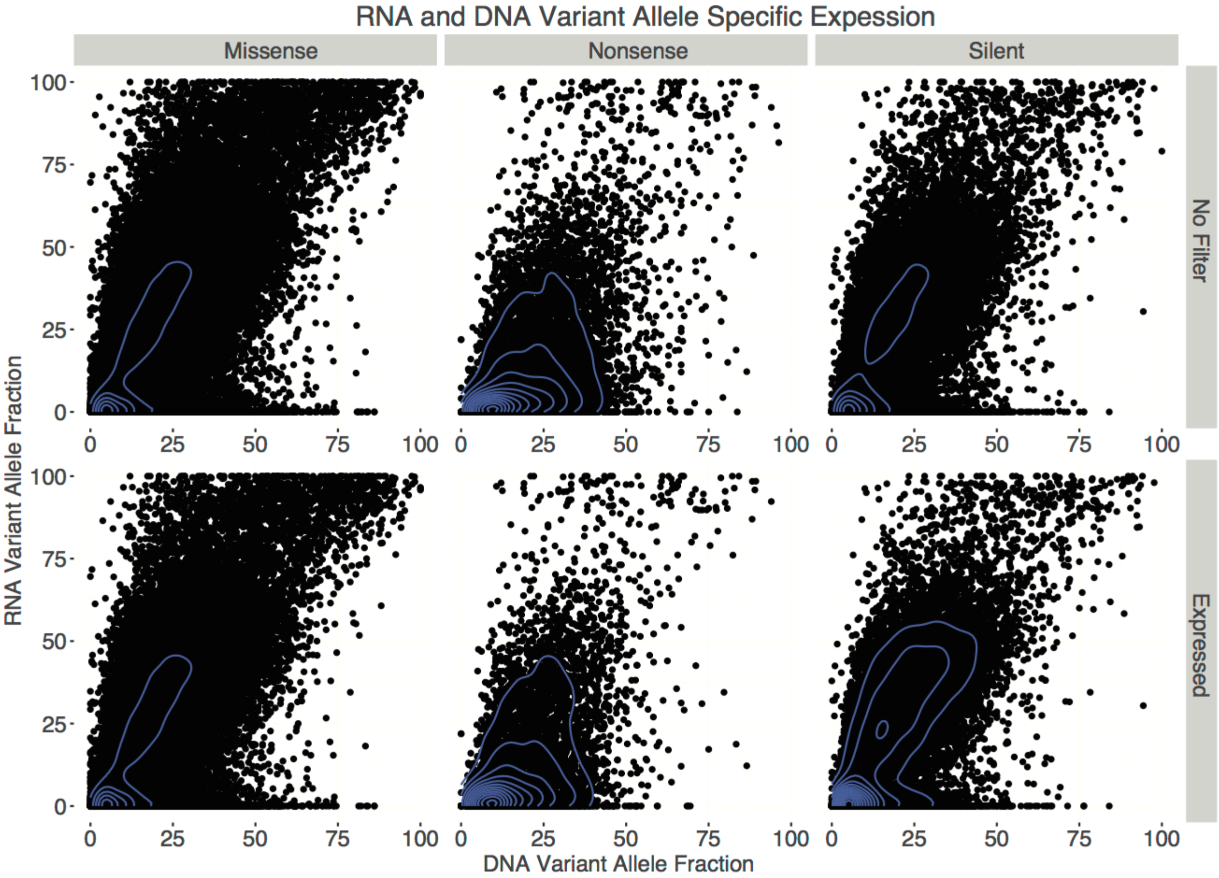


Figure A.4. Comparison of RNA Variant Allele Fraction and DNA allele fraction by conventional annotation type. Top panels indicate all sites and bottom panels are only expressed.

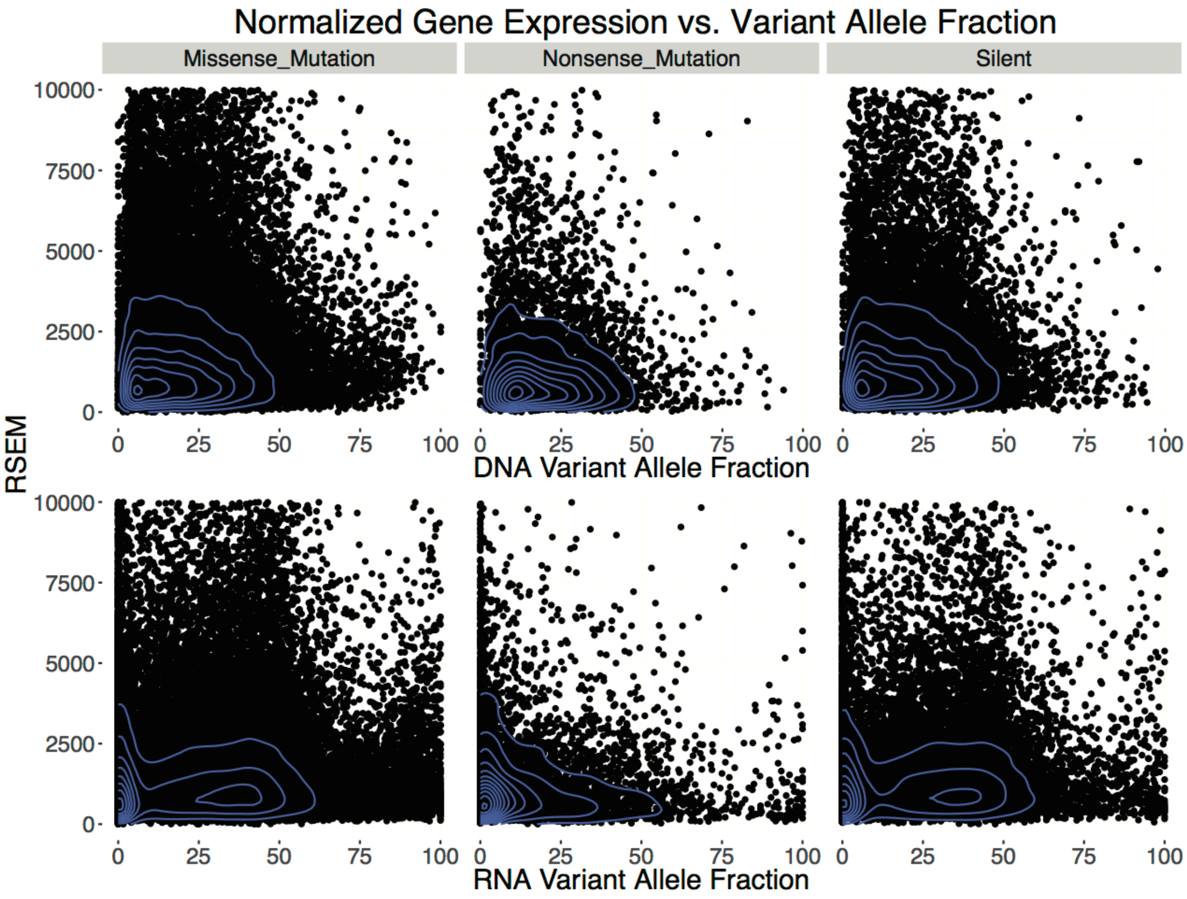


Figure A.5. Normalized gene expression (RSEM) and variant allele fraction for all sites.

Table A.3. Total number of variants in each RNA/DNA VAF ratio category

	No Filter			**Expressed**		
Mutation Type	Ratio >= 1	Ratio < 1	Fisher (Silent)	Ratio >= 1	Ratio < 1	Fisher (Silent)
Missense Mutation	33304	24727	8.863e-15	24632	17556	2.143e-12

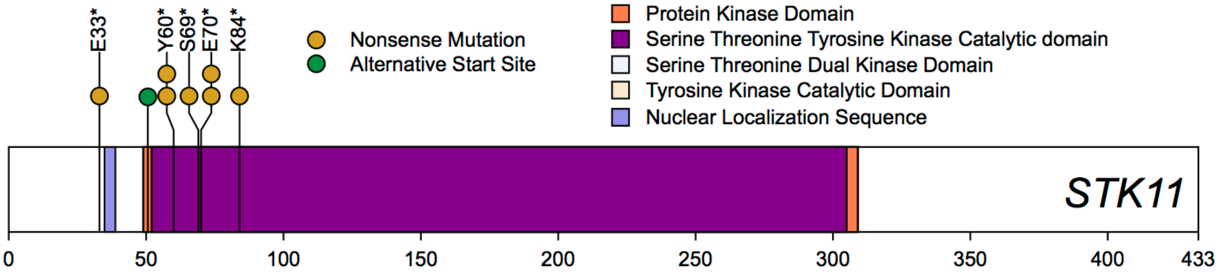
Nonsense Mutation	2219	6329	<2.2e-16	1556	3623	<2.2e-16
Silent Mutation	11960	10049	-	8793	7150	-

76,501 mutations fall into copy number neutral sites and 63,489 mutations have case RSEM greater than the lower quartile of matched cancer control samples lacking mutations in the gene of interest. We classify these mutations as expressed relative to their control samples.

We chose to leverage variant allele fractions (VAFs) measured by DNA and RNA sequencing of the tumor samples to define the relationship between the genomic position and transcriptome effect. One would expect a comparative relationship between RNA and DNA VAF if allelic content and transcription were directly proportional. But in practice there are a number of variables that can cause a deviation from this proportional relationship including post-transcriptional modifications such as RNA degradation. Figure A.4 highlights the relationship between DNA and RNA VAF across missense, nonsense and silent mutations while table 1 gives a numerical representation of the number of mutations falling above a slope of 1. Nonsense mutations show an increased number of mutations with higher DNA VAF than RNA VAF, suggesting a higher level of degradation efficiency compared to missense and silent mutations. Overall we can see a higher level of RNA VAF enrichment of missense mutations (positive selection) compared to silent mutations while nonsense mutations show on overall negative selection.

Using the silent mutation RNA and DNA VAF distribution, we can make the assumption silent mutations follow a null distribution. We used fishers exact test to compare the proportion of mutations greater than and less than a RNA DNA VAF Ratio of 1 to determine if the difference in proportions is significant between both missense and nonsense with silent mutations as the null distribution.

Novel inactivation of STK11



Amino Acid Change	Sample	Cancer	RNA VAF	DNA VAF	Case RSEM	Control (Lower Quartile)	RNA DNA VAF Ratio	Expression Classification
p.E33Ter	TCGA-86-8359-01A-11D-2323-08	luad	86.78	30.3	1492.1021	634.40235	2.86403	High Expression
p.Y60Ter	TCGA-53-7813-01A-11D-2167-08	luad	63.63	40.91	232.2713	634.40235	1.55	Not Expressed
p.Y60Ter	TCGA-69-7763-01A-11D-2167-08	luad	37.50	28.57	544.9978	634.40235	1.3125	Not Expressed
p.S69Ter	TCGA-NJ-A55O-01A-11D-A25L-08	luad	27.59	11.11	810.9685	634.40235	2.48335	Expressed
p.E70Ter	TCGA-49-4487-01A-21D-1855-08	luad	33.93	20	593.9343	634.40235	1.6965	Not Expressed
p.E70Ter	TCGA-62-A46P-01A-11D-A24D-08	luad	84.44	57.69	1067.6692	634.40235	1.46369	Expressed
p.K84Ter	TCGA-44-A479-01A-31D-A24D-08	luad	36.17	14.29	479.5687	634.40235	2.53114	Not Expressed

Figure A.6. Lolliplot of novel inactivating mutations in STK11. Statistics of STK11 variants below.

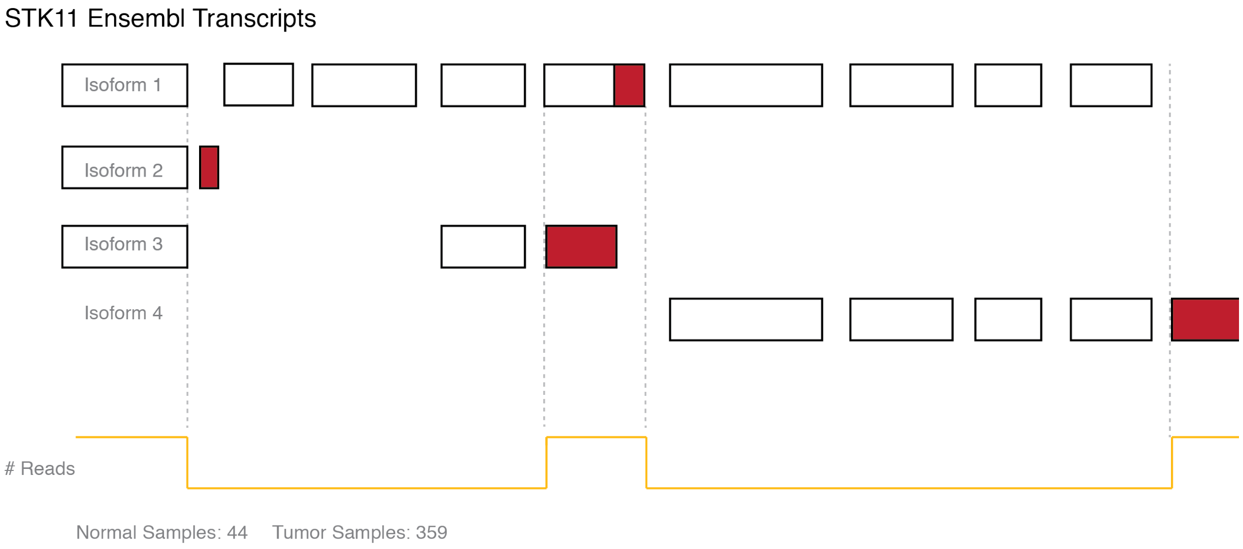


Figure A.7. All Alternative isoforms of indicated STK11 mutants.

We identified 7 mutations in the canonical transcript of exon in ensembl transcript ENST00000326873. We used TransVar[1] to perform equivalence annotation to map the mutation of interest to all other transcripts defined by Ensembl based on the genomic coordinates of the variant. To determine which transcript is readily expressed in the tissue type, we determined the total number of reads supporting the first and last base of the exon unique to each Ensembl transcripts. For the example of STK11 we determined transcript X is expressed in LUAD by assessing number of reads supporting the unique exon for that transcript.

Studies found that the subcellular localization of kinase deficient mutants were found in the nucleus while mutants with disrupted NLS were localized in the cytoplasm. Furthermore, the mutant with the disrupted NLS could still induce G1 cell cycle arrest comparable to the wild type protein. We hypothesize a subset of nonsense mutations in the first exon of STK11 utilize a downstream start codon to create N terminally truncated proteins lacking the N terminal localization sequence.

Identification of activating mutations using allele specific expression

Our analysis was able to pick up well known activating mutations in key cancer genes including EGFR. Of the 28 EGFR mutations tested in our gene expression test comparing case to control gene expression, 16 had RNA VAF greater than 10% suggesting some level of transcription supporting the variant allele. 9 of the 16 mutations are functionally characterized activating mutations in EGFR including: L858R[4] (4 - LUAD), L861Q (1 - LUSC), G719A[4] (1 - LUAD), G598A,V (2 –LGG), R108K[5] (1 - GBM). Furthermore, a tool using protein structure guided discovery identified R252C/P (2 – LGG), S768I (1 – LUAD) and L833V (1 – LUAD) as potential activating mutations due to their proximity to known activating mutations[6].

Kelch-like erythroid cell derived protein with CNC homology (ECH)-associated protein 1 (Keap1) is essential in the regulation of cytoprotective and detoxifying defense systems. Keap1 is responsible for sequestering nuclear factor erythroid 2-related factor 2 (Nrf2 or NFE2L2) in the cytoplasm and interacts with Cul3-E3 ubiquitin ligase complex to target Nrf2 for ubiquitination and degradation by the proteosome. Many mutations in lung cancer disrupt the binding of Keap1 to Nrf2, thereby increasing the presence of free active Nrf2, or the degradation efficiency of Nrf2 by disrupting binding to Cul3[7]. In 2014, several "superbinder" mutants were found to increase the levels of Nrf2 within the nucleus without disrupting the interaction between KEAP1 and Nrf2[8]. To evaluate how this could be the case, the authors looked at protein turnover of Nrf2 and found the superbinder mutants R320Q and R470C dramatically stabilized Nrf2 more so than the wild-type Keap1 but maintained successful ubiquitination. This surprising result could suggest that the increased affinity for Keap1 for Nrf2 suppresses substrate turnover by disrupting a degradation step post ubiquitination.

Assessing degradation efficiency of nonsense mutations using allele specific expression

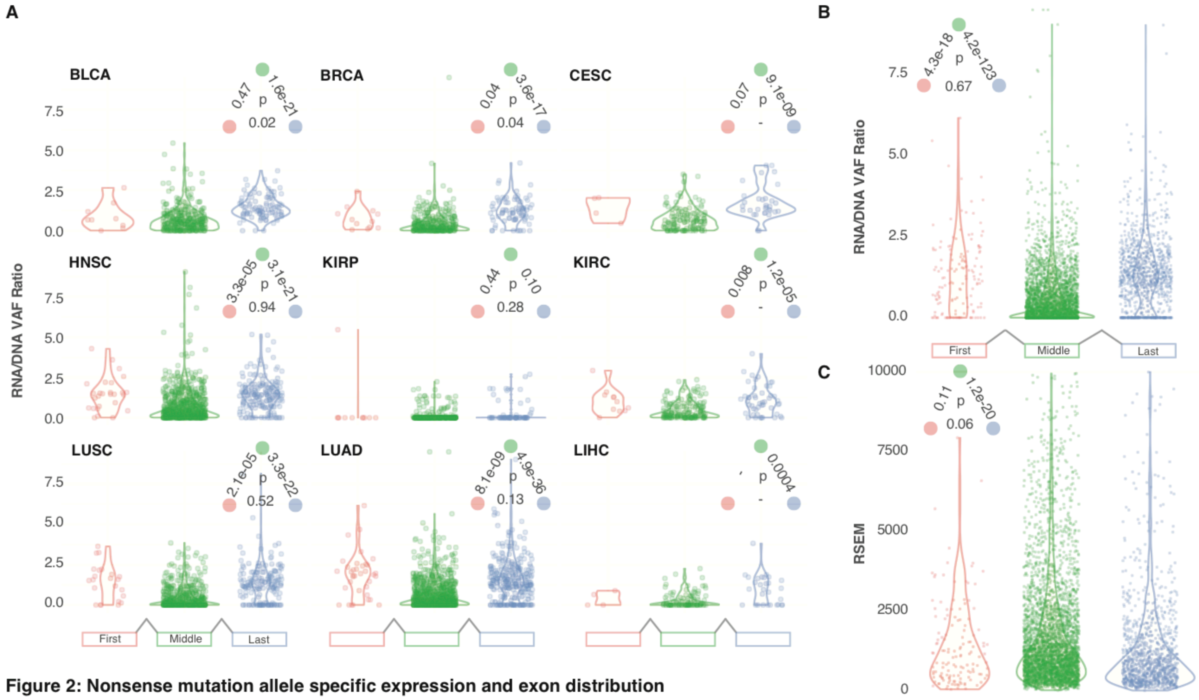


Figure 2: Nonsense mutation allele specific expression and exon distribution

Figure A.8. (A) Allele specific expression denoted by RNA/DNA VAF Ratio of nonsense mutations across exons. Red corresponds to mutations found in the first exon, green in middle exon and blue in last exon. Wilcox test was used to compare groups of nonsense mutations located in the first, middle and last exons of a gene. (B) RNA/DNA VAF Ratio distribution across all cancer types for mutations found in the first, middle and last exons. (C) RSEM distribution across all cancer types for mutations found in first, middle and last exons.

Nonsense mutations in the last exon tend to have higher allele specific expression than mutations in another exon, as expected. Interestingly, nonsense mutations in the first exon show a similar trend as the last exon mutations (higher allele specific expression). N terminally truncated proteins can lose their localization sequences, thereby altering the cellular localization of the protein. Studies of β-globin show a very distinct transition between nonsense mutations in the first exon that are able to bypass NMD and those that are subject to degradation. The potential of the transcript to evade degradation is dependent upon the presence of a downstream start codon to reinitiate translation.

All nonsense mutations from the first and last exon were annotated to all possible transcripts using TransVar to determine mapping potential to alternative isoforms. 143 of the 207 nonsense mutations annotated to the first exon maintained the nonsense mutation prediction on alternative isoforms. Of the 143 variants annotated solely to the first exon, 58 variants had an RNA VAF less than 10% suggesting selection against the variant allele in the tumor or the presence of a subclonal mutation. The presence of the remaining 85 variants with greater than 10% RNA VAF supporting the N-terminal nonsense mutation could be explained by translation readthrough or reinitiation at a downstream start codon. Two studies reported leak expression when nonsense mutations were introduced up to residue 70 of the Shaker voltage-gated potassium (Kv) channel[10] and residue 26 of B-globin[9], suggesting a sharp divide between mutations susceptible and resistant to NMD. Both studies came to the conclusion that the likely mechanism is due to reinitiation of translation at a downstream translation start site, and in the case of the Kv channel, non-canonical start codons were used (AAG, AGG). Additionally, translation isn’t limited to the next start codon, but alternative start sites farther downstream can be used by a fraction of ribosomes.

5.8 Viral integration

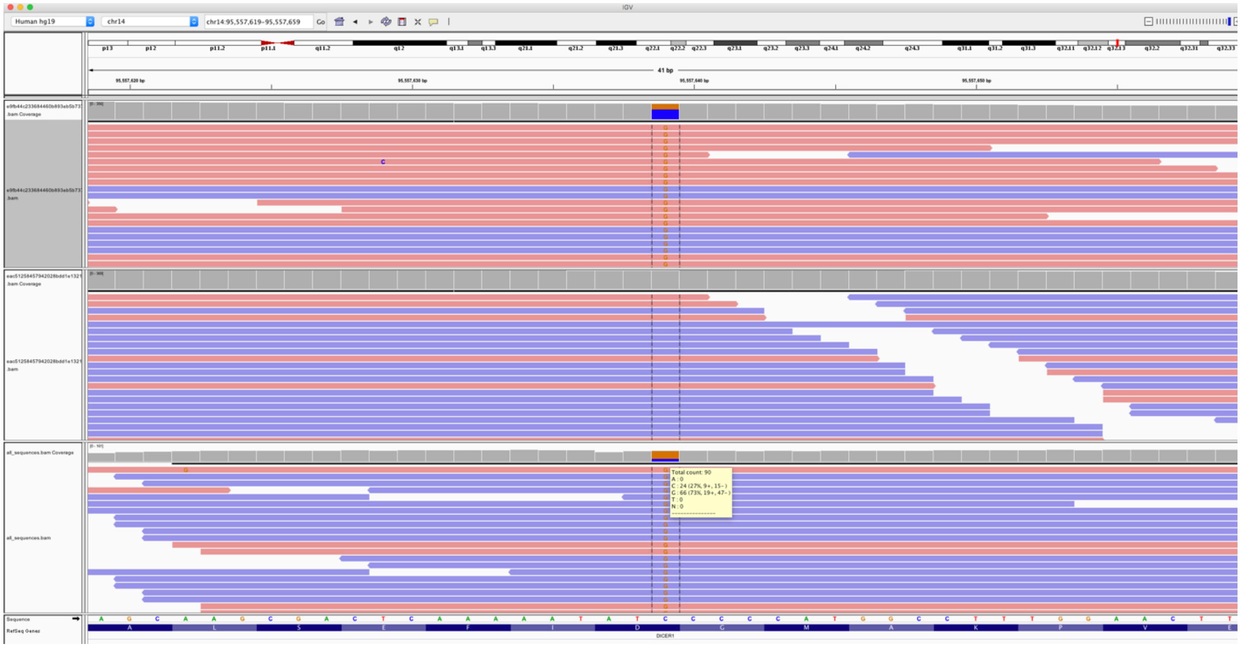
Contribution: I helped with gene expression figures related to discordant pair analysis. All Figures and supplemental documentation can be seen in the following publication.

Divergent viral presentation among human tumors and adjacent normal tissues, Scientific Reports, 2016 Jun. doi:10.1038/srep28294

Cao S., Wendl M., Wyczalkowski M.A., Wylie K., Ye K., **Jayasinghe R.G.** et al.

5.9 DROSHA Mutation Analysis

Through this collaboration we identified novel mutations in DROSHA and Dicer and evaluated RNA-Seq expression for associated downstream products that could be altered. Of note, for one of the mutated samples, MDM2 expression was increased, matching our collaborators findings in a patient of interest.



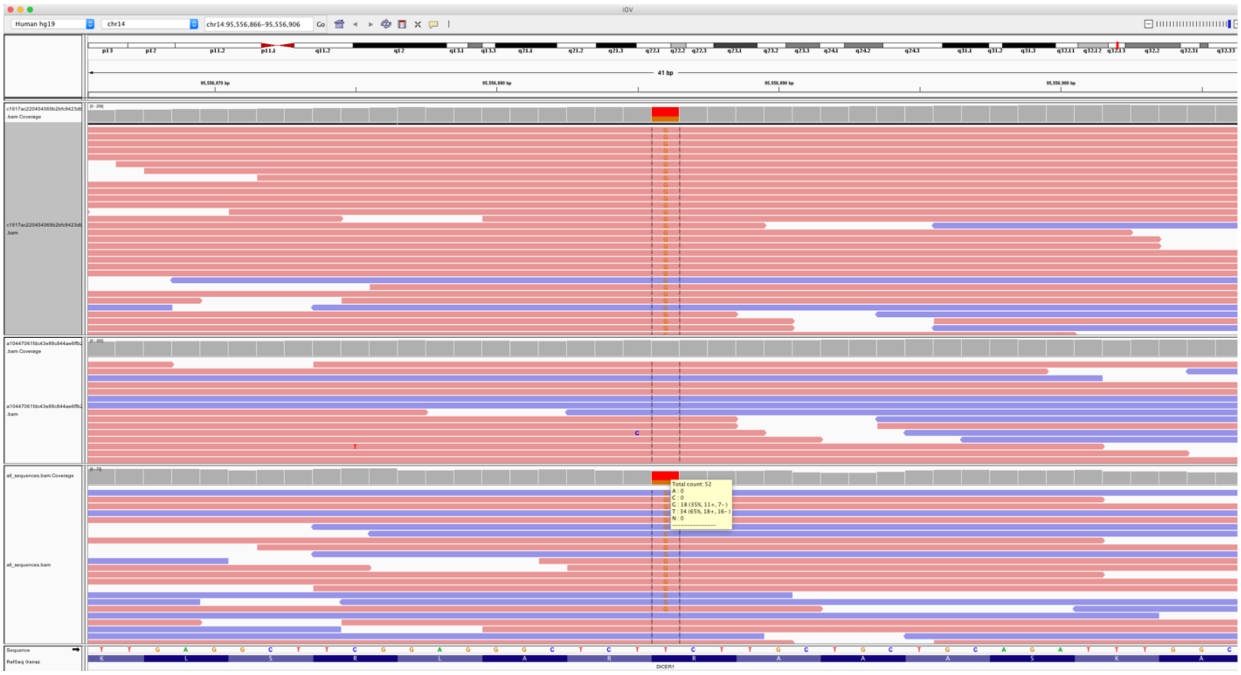
Human hg19

chr14

chr14:95,557,619-95,557,659

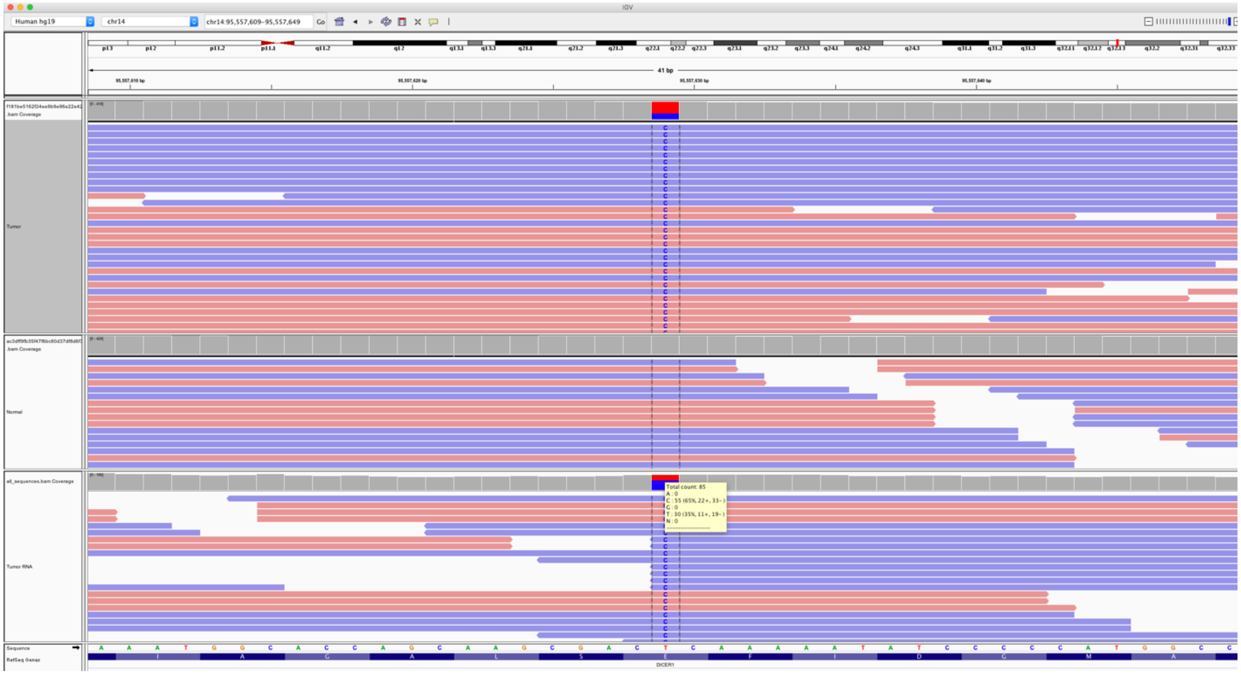
Total count: 90
A : 0
C : 24 (27%, 9+, 15-)
G : 66 (73%, 19+, 47-)
T : 0
N : 0

Figure A.9. DICER1_14_95556886_T_G_TCGA-EL-A3GO



Total count: 52
A : 0
C : 0
G : 18 (35%, 11+, 7-)
T : 34 (65%, 18+, 16-)
N : 0

Figure A.10. DICER1_14_95556886_T_G_TCGA-EM-A2CT



Total count: 85
A : 0
C : 55 (65%, 22+, 33-)
G : 0
T : 30 (35%, 11+, 19-)
N : 0

Figure A.11. DICER1_14_95557629_T_C_TCGA-EL-AD35

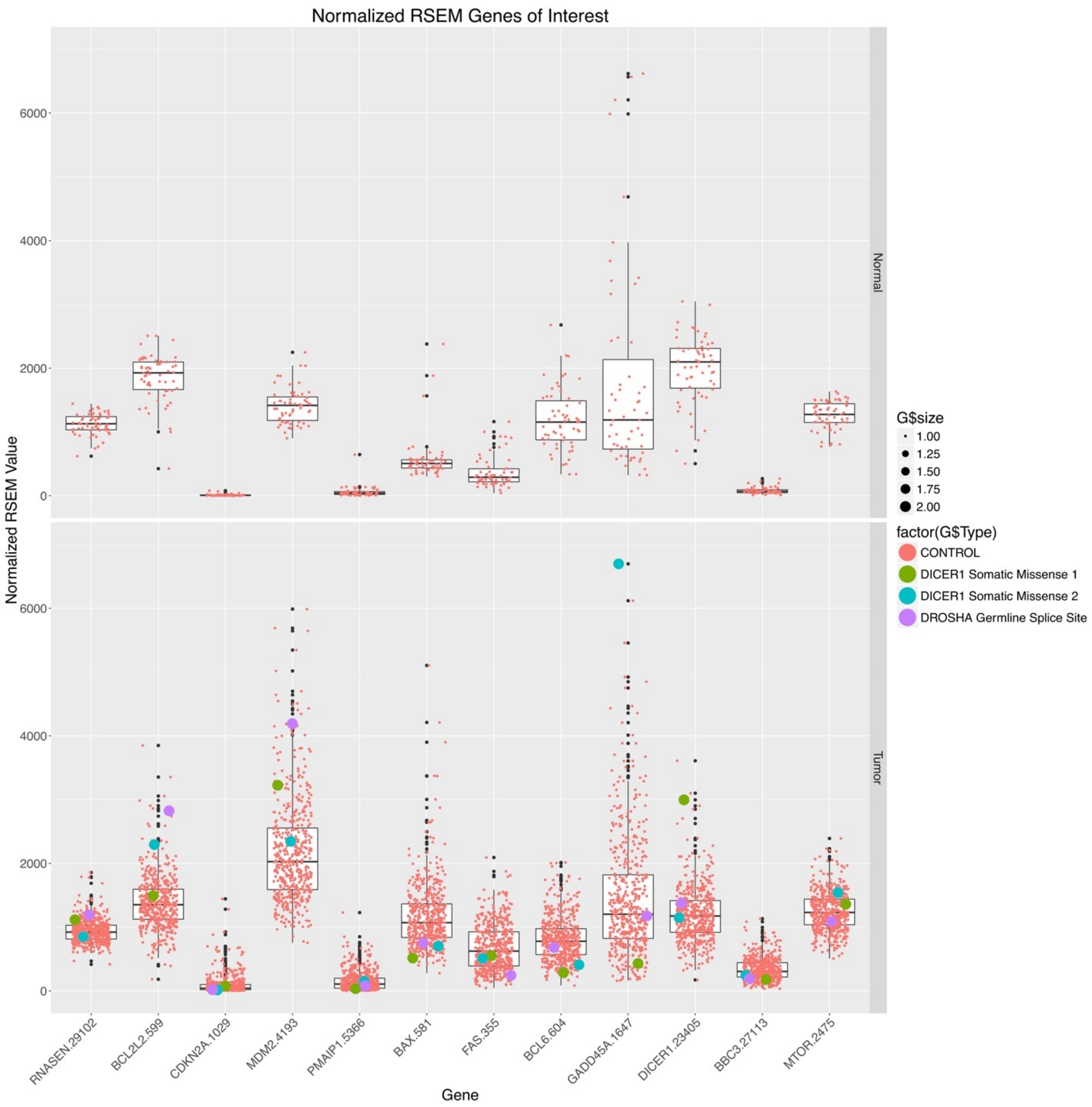


Figure A.12. Distribution of normalized expression for downstream genes and DICER1 for mutants vs. control samples

Table A.4: Samples with DROSHA Mutations

Cancer	Diagnosis (Age)	**Chr**	**start**	**Ref**	**Var**	**amino_acid_change**	**sample**	**Normal WXSV AF**	**Tumor WXSV AF**	**Tumor RNA VAF**	**Selection**
LAML	45	5	31451686	A	-	p.L879fs	TCGA-AB-2983	0.57	0.48	0.24	Not under selection
STAD	75	5	31526994	G	A	p.R16*	TCGA-HU-8610	0.53	0.37	0.15	Not under selection
THCA	77	5	31527020	C	T	e2-1	TCGA-EM-A2CO	0.42	0.54	NA	Under selection
UCEC	65	5	31431787	T	C	e23-2	TCGA-D1-A2G7	0.39	0.38	NA	Not under selection
UCEC	59	5	31521318	G	A	p.R287*	TCGA-AX-A1CJ	0.54	0.43	NA	Not under

											selection

Uncategorized References

Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nature structural & molecular biology *16*, 107-113.

Brooks, A.N., Choi, P.S., de Waal, L., Sharifnia, T., Imielinski, M., Saksena, G., Pedamallu, C.S., Sivachenko, A., Rosenberg, M., Chmielecki, J*., et al.* (2014). A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. PLoS One *9*, e87361.

Gao, Q., Liang, W.W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L*., et al.* (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep *23*, 227-238 e223.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A *100*, 189-192.

Lu, C., Xie, M., Wendl, M.C., Wang, J., McLellan, M.D., Leiserson, M.D., Huang, K.L., Wyczalkowski, M.A., Jayasinghe, R., Banerjee, T*., et al.* (2015). Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun *6*, 10086.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M., Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev *21*, 708-718.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev *20*, 153-158.

Rajan, P., Elliott, D.J., Robson, C.N., and Leung, H.Y. (2009). Alternative splicing and biological heterogeneity in prostate cancer. Nature reviews Urology *6*, 454-460.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Weischenfeldt, J., Damgaard, I., Bryder, D., Theilgaard-Monch, K., Thoren, L.A., Nielsen, F.C., Jacobsen, S.E., Nerlov, C., and Porse, B.T. (2008). NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev *22*, 1381-1396.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J.S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B.T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol *13*, R35.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, C., McMichael, J.F., Schmidt, H.K., Miller, C.a., Bradley, a.*, et al.* (2014). Age-related cancer mutations associated with clonal hematopoietic expansion. Nature Medicine.

Ye, K., Wang, J., Jayasinghe, R., Lameijer, E.W., McMichael, J.F., Ning, J., McLellan, M.D., Xie, M., Cao, S., Yellapantula, V.*, et al.* (2016). Systematic discovery of complex insertions and deletions in human cancers. Nat Med *22*, 97-104.

Zhang, C., Li, H.R., Fan, J.B., Wang-Rodriguez, J., Downs, T., Fu, X.D., and Zhang, M.Q. (2006). Profiling alternatively spliced mRNA isoforms for prostate cancer classification. BMC Bioinformatics *7*, 202.

Zhou, W., Chen, T., Chong, Z., Rohrdanz, M.A., Melott, J.M., Wakefield, C., Zeng, J., Weinstein, J.N., Meric-Bernstam, F., Mills, G.B.*, et al.* (2015). TransVar: a multilevel variant annotator for precision genomics. Nat Methods *12*, 1002-1003.