Ultra-large-scale genomics approaches to improve cancer therapeutic response

Jose Mario Bello Pineda

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Robert K. Bradley, Chair

Marshall S. Horwitz

Stephen J. Tapscott

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Ultra-large-scale genomics approaches to improve cancer therapeutic response

Jose Mario Bello Pineda

Chair of the Supervisory Committee:

Robert K. Bradley

Director, Translational Data Sciences Integrated Research Center

Full Professor, Public Health Sciences and Basic Sciences Divisions

McIlwain Family Endowed Chair in Data Science

Fred Hutchinson Cancer Center

The advent of low-cost, high-throughput sequencing technologies has elucidated targetable cancer-specific genomic alterations and allowed the development of precision therapies and their deployment into clinical use. However, definitive determinants of response to targeted drugs remain elusive. Therefore, knowledge of genomic features relevant to mechanisms of oncogenesis and cancer susceptibility to therapeutic interception is imperative. In this dissertation, I detail

studies that identify such features, through the analysis of publicly available RNA sequencing (RNA-seq) datasets at a massive scale. First, we performed some of the largest RNA-seq analyses ever conducted to identify branchpoint nucleotide positions genome-wide, characterize the unexpected structural and regulatory complexity of human introns, and describe the unappreciated genome-wide prevalence of circular intron-derived RNAs. Second, we coupled multiple large genetic screens with experimental perturbations and RNA-seq analyses of patient-derived and cell line models of acute myeloid leukemia (AML) to identify determinants of drug response. We determined that splicing modulation is a unique AML susceptibility and identified specific splicing changes in the transcripts of spliceosomal components and apoptotic factors mediating sensitivity and resistance to the *BCL2* inhibitor venetoclax. Last, we analyzed large and diverse cancer cohorts using genomic, statistical, and machine learning methods to identify *DUX4* reactivation as a common mechanism of immune evasion in advanced cancers and central feature of metastatic cancer patients resistant to checkpoint immunotherapy. These works amalgamate genomic and clinical data to identify novel strategies to target cancer and improve the precision and efficacy of current treatment modalities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The person I am today was shaped by countless lessons from many kind and generous souls. I am indebted to all of them.

Words are insufficient to express my gratitude to my mentor Robert K. Bradley. Under his tutelage I found my identity as a scientist. I am thankful for the freedom and support he afforded me, and for his invaluable guidance through the numerous but important cycles failure and success. I had an immeasurable amount of personal and professional growth in my time as a scientist in his lab. I aspire to his standard of scientific rigor and creativity. Most of all, I am grateful for his patience and compassion. In the moments where the tides of life's events were overpowering, or even torpefying, he gave me the time and space to step away, to heal; and provided advice and words of comfort to nevertheless make me feel ascendant. Those moments meant the world to me. I will carry all of his lessons, especially those in kindness, for the rest of my career.

I am thankful to Wenying Shou, my first mentor. As an apprentice in her lab, I fell in love with biology. I feel fortunate for the focused experimental and quantitative training I received from her first-hand, for the trust she had in me to lead projects even as a novice, and for sharing her approach to original and meticulous science. I treasure her mentorship that has transcended geographic regions and persisted years beyond me exiting her lab.

I had the immense pleasure of working with members of the Bradley lab, past and present. They helped me gain the tools I needed to explore and discover my path as a scientist, I appreciate their generosity and patience. As a person with no computational experience coming into graduate

school, their help was instrumental. They made the workplace environment constantly enjoyable, something that always felt special to me. The friendships I formed in this lab I will cherish for a lifetime.

I am thankful to the members of my supervisory committee— Marshall S. Horwitz, Stephen J. Tapscott, Edith Wang, Judit Villen, and Cole Trapnell— and the UW Department of Genome Sciences and the Medical Scientist Training Program for their support. I am thankful to the Fred Hutchinson Cancer Center for the unique experience of working there for the last 12 years.

I am grateful for Lisa Peterson and the University of Washington Genomics Outreach for Minorities (UW GenOM) Project for a research internship which shaped the trajectory of my life and career.

I am thankful to all my friends, especially Morgan Todd and Racheal Jappert for their support and for pushing me to become the best version of myself. To my family who I adore, thank you: my dad, who fostered my love for science; my siblings Bon, Trina, Isa, and Jao for their support, love, and endless understanding. I thank my aunt Margaret Kessler-Allen for her kindness and assistance throughout the years, and for being my family away from home.

To Andrew Kessler, I am grateful to him for being in my life, his unwavering support, for restoring joy in times of my greatest sorrow, and for my life's pride and happiness: our love, our home, and our Henry.

Finally, I am most thankful for my mother Pia Valera Bello, my original teacher, who taught me how to work hard, put my heart in everything that is important, and to follow my bliss. I can never repay the sacrifices she made to give me a world where the limits of my potential are delineated merely by the dreams I cast in my mind. But I will spend the rest of my time making the most of this life in the hopes that I will one day find a way.

To my mother, Pia Valera Bello.

The light of my life, my love and inspiration.

For granting me the world and all its wonders,

I dedicate this body of work to you.

Chapter 1. INTRODUCTION

Precision oncology, treating cancer based on individual tumor-specific or tumor-associated genetic alterations, has been revolutionary: significantly improving patient outcomes compared to conventional cytotoxic or cytostatic chemotherapy agents. Perhaps the best example of this success was exhibited by imatinib mesylate, a small molecule inhibitor of the tyrosine kinase encoded by the hallmark BCR-ABL gene fusion of chronic myeloid leukemia (CML) (Carlisle et al., 2020; Nowell, 1962; Rowley, 1973). Early clinical studies demonstrated the efficacy of imatinib in treating CML and gastrointestinal tumors, even achieving complete remission in a number of patients (Druker, Sawyers, et al., 2001; Druker, Talpaz, et al., 2001; Goldman & Melo, 2001; Joensuu et al., 2001). It was approved in 2001 for the treatment of CML and 2002 for gastrointestinal stromal tumors (Savage & Antman, 2002). Recently, a study detailing the results of an 11-year observation of CML patients receiving imatinib showed long-lasting patient benefit with an overall survival rate at 10 years of 83.3% absent long-term toxicity (Hochhaus et al., 2017), cementing imatinib's remarkable success. At present, the incorporation of sequencing into routine cancer care directly informs clinical decisions. In addition, numerous directed therapies which target tumor characteristics resulting from specific mutations are widely available (Chae et al., 2017), demonstrating the shift towards treating cancer as a genomic disease.

Immunotherapy has similarly changed the way metastatic cancers are treated. For example, advanced melanoma patients now exhibit median overall survival greater than 60 months with combination checkpoint immunotherapy (Larkin et al., 2019), which contrasts the standard 6-month median overall survival that has historically plagued this patient population (Hodi et al., 2010). Response to immune checkpoint inhibition (ICI) can be predicted, to an extent, based on

genomic characteristics of a patient's cancer. Tumor mutational burden (TMB) is a clinically-important feature associated with ICI efficacy. TMB screening has become a mainstay in oncology clinics for guiding treatment choice— through sequencing or immunohistochemical assessment for the absence of DNA-repair enzymes. This biomarker is rooted in the biology of the disease process itself: consistent with the notion of high mutational load resulting in an increase of cancer neoantigens that could be recognized, a cancer vulnerability (Samstein et al., 2019). The discovery of TMB-associated response has been transformational. Pembrolizumab (*PD-1* inhibitor, also known as Keytruda), one of several ICI modalities available, became the first drug to receive FDA approval for use in any solid malignancies exhibiting high TMB. ICI treatment exhibits the possibility of cancer treatment as a personalized, mutation-directed, and tumor-type agnostic endeavor instead of relying on strategies based on population-averaged data from clinical trials.

While our understanding of cancer-specific molecular features which translate to precision therapies in the current age is unprecedented, there is much improvement to be desired (Tannock & Hickman, 2016). As of August 2022, approximately 220 oncological pharmacogenomic biomarkers are associated with FDA-approved drugs (USFDA, 2022). However, this represents a strikingly small minority of cancer-associated mutations that have been evaluated and approved for clinical use (Dienstmann et al., 2015). In addition, presence of a biomarker does not guarantee response to a particular targeted therapy in the current setting. For instance, majority of patients fail to respond, or exhibit acquired resistance, to ICI despite achieving durable and long-term effects in others (Schoenfeld & Hellmann, 2020). Cancer is biologically complex and requires further inquiry into other strategies to improve outcomes, and possibly additional biomarkers that could be used synergistically with existing knowledge to better guide treatment decisions (e.g., combination therapies).

The ‘big data’ approach is one avenue towards the aforementioned discovery effort. There is a massive amount of multimodal sequencing data available from the research and clinical spaces due in part to the dramatic reduction in the costs of sequencing. Since 2007, price of sequencing has reduced from $1,000,000 per human genome to an average of $1269 to $2058 per next-generation sequencing clinical test in 2021, outpacing Moore’s law for computing costs (Desai et al., 2021; NHGRI, 2021). Costs are expected to drop even further, with the rapid improvements in sequencing technologies. In this dissertation, I detail the use of publicly available sequencing data to reveal novel biology about cancer development, progression, sensitivity to available drugs, and alternative strategies for interception.

REFERENCES

Carlisle, B. G., Zheng, T., & Kimmelman, J. (2020). Imatinib and the long tail of targeted drug development. *Nature Reviews. Clinical Oncology*, *17*(1), 1–3. https://doi.org/10.1038/s41571-019-0287-0

Chae, Y. K., Pan, A. P., Davis, A. A., Patel, S. P., Carneiro, B. A., Kurzrock, R., & Giles, F. J. (2017). Path toward Precision Oncology: Review of Targeted Therapy Studies and Tools to Aid in Defining “Actionability” of a Molecular Lesion and Patient Management Support. *Molecular Cancer Therapeutics*, *16*(12), 2645–2655. https://doi.org/10.1158/1535-7163.MCT-17-0597

Desai, K., Hooker, G., Gilbert, K., Cropper, C., & Metcalf, R. (2021). Real-world trends in costs of next generation sequencing (NGS) testing in U.S. setting. *Journal of Clinical Oncology*, *39*(15).

Dienstmann, R., Jang, I. S., Bot, B., Friend, S., & Guinney, J. (2015). Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discovery*, *5*(2), 118–123. https://doi.org/10.1158/2159-8290.CD-14-1118

Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., Capdeville, R., & Talpaz, M. (2001). Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *The New England Journal of Medicine*, *344*(14), 1038–1042. https://doi.org/10.1056/NEJM200104053441402

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., & Sawyers, C. L. (2001). Efficacy and

safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, *344*(14), 1031–1037. https://doi.org/10.1056/NEJM200104053441401

Goldman, J. M., & Melo, J. v. (2001). Targeting the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, *344*(14), 1084–1086. https://doi.org/10.1056/NEJM200104053441409

Hochhaus, A., Larson, R. A., Guilhot, F., Radich, J. P., Branford, S., Hughes, T. P., Baccarani, M., Deininger, M. W., Cervantes, F., Fujihara, S., Ortmann, C.-E., Menssen, H. D., Kantarjian, H., O'Brien, S. G., Druker, B. J., & IRIS Investigators. (2017). Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. *The New England Journal of Medicine*, *376*(10), 917–927. https://doi.org/10.1056/NEJMoa1609324

Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., van den Eertwegh, A. J. M., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbé, C., … Urba, W. J. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *The New England Journal of Medicine*, *363*(8), 711–723. https://doi.org/10.1056/NEJMoa1003466

Joensuu, H., Roberts, P. J., Sarlomo-Rikala, M., Andersson, L. C., Tervahartiala, P., Tuveson, D., Silberman, S., Capdeville, R., Dimitrijevic, S., Druker, B., & Demetri, G. D. (2001). Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *The New England Journal of Medicine*, *344*(14), 1052–1056. https://doi.org/10.1056/NEJM200104053441404

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Lao, C. D., Cowey, C. L., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P. F., Smylie, M., Hogg, D., Hill, A., Márquez-Rodas, I., Haanen, J., Guidoboni, M., Maio, M., Schöffski, P., … Wolchok, J. D. (2019). Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *The New England Journal of Medicine*, *381*(16), 1535–1546. https://doi.org/10.1056/NEJMoa1910836

NHGRI. (2021). *DNA Sequencing Costs: Data*. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Nowell, P. C. (1962). The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut*, *8*, 65–66. https://doi.org/10.1007/BF01630378

Rowley, J. D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, *243*(5405), 290–293. https://doi.org/10.1038/243290a0

Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., Barron, D. A., Zehir, A., Jordan, E. J., Omuro, A., Kaley, T. J., Kendall, S. M., Motzer, R. J., Hakimi, A. A., Voss, M. H., Russo, P., Rosenberg, J., Iyer, G., Bochner, B. H., … Morris, L. G. T. (2019). Tumor mutational load predicts survival after immunotherapy

across multiple cancer types. *Nature Genetics*, *51*(2), 202–206. https://doi.org/10.1038/s41588-018-0312-8

Savage, D. G., & Antman, K. H. (2002). Imatinib mesylate--a new oral targeted therapy. *The New England Journal of Medicine*, *346*(9), 683–693. https://doi.org/10.1056/NEJMra013339

Schoenfeld, A. J., & Hellmann, M. D. (2020). Acquired Resistance to Immune Checkpoint Inhibitors. *Cancer Cell*, *37*(4), 443–455. https://doi.org/10.1016/j.ccell.2020.03.017

Tannock, I. F., & Hickman, J. A. (2016). Limits to Personalized Cancer Medicine. *The New England Journal of Medicine*, *375*(13), 1289–1294. https://doi.org/10.1056/NEJMsb1607705

USFDA. (2022). *Table of Pharmacogenomic Biomarkers in Drug Labeling*. https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling

Chapter 2. MOST HUMAN INTRONS ARE RECOGNIZED VIA MULTIPLE AND TISSUE-SPECIFIC BRANCHPOINTS

A version of this chapter has been previously published as:

Pineda, J. M. B., & Bradley, R. K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes & Development*, 32(7–8), 577–591. https://doi.org/10.1101/gad.312058.118

2.1 ABSTRACT

Although branchpoint recognition is an essential component of intron excision during the RNA splicing process, the branchpoint itself is frequently assumed to be a basal, rather than regulatory, sequence feature. However, this assumption has not been systematically tested due to the technical difficulty of identifying branchpoints and quantifying their usage. Here, we analyzed ∼1.31 trillion reads from 17,164 RNA sequencing data sets to demonstrate that almost all human introns contain multiple branchpoints. This complexity holds even for constitutive introns, 95% of which contain multiple branchpoints, with an estimated five to six branchpoints per intron. Introns upstream of the highly regulated ultraconserved poison exons of SR genes contain twice as many branchpoints as the genomic average. Approximately three-quarters of constitutive introns exhibit tissue-specific branchpoint usage. In an extreme example, we observed a complete switch in branchpoint usage in the well-studied first intron of *HBB* (β-globin) in normal bone marrow versus metastatic prostate cancer samples. Our results indicate that the recognition of most introns is unexpectedly complex and tissue-specific and suggest that alternative splicing catalysis typifies the majority of introns even in the absence of differences in the mature mRNA.

2.2 INTRODUCTION

RNA splicing proceeds via a two-step process defined by sequential transesterification reactions between three nucleotides: the first nucleotide of the 5′ splice site, the branch nucleotide (branchpoint) upstream of the 3′ splice site, and the last nucleotide of the 3′ splice site. In the first step of splicing, the 2′ OH group of the branchpoint engages in a nucleophilic attack on the phosphate between the upstream exon and the 5′ splice site, forming a 2′–5′ phosphodiester linkage (the "branch") characteristic of the lariat RNA intermediate and releasing the upstream exon. The 3′ OH group of the now-free upstream exon then engages in a nucleophilic attack on the phosphate between the 3′ splice site and the downstream exon, resulting in release of the intronic lariat and exon ligation (for review, see Wahl et al., 2009). The intronic lariat is then linearized via debranching and subsequently degraded.

The branchpoint therefore plays a critical role in RNA splicing catalysis, similar in importance to the splice sites themselves. The branchpoint's biochemical role in the splicing of specific substrate RNAs has been thoroughly studied accordingly. Nonetheless, the identification, selection, and potential regulation of branchpoints remains poorly understood, even relative to other intronic elements such as the polypyrimidine tract or intronic splicing silencers and enhancers that, like the branchpoint, do not appear in the final mRNA product (for review, see Fu and Ares, 2014; Scotti and Swanson, 2016).

The study of branchpoints has lagged behind the study of other sequence features that define introns and exons for several reasons. Experimentally identifying branchpoints is technically difficult, since lariats exist only as transient low-abundance RNAs. Computationally predicting branchpoints is similarly difficult due to the low information content of the human branchpoint consensus (Zhuang et al., 1989; Kol et al., 2005; Gao et al., 2008; Corvelo et al.,

2010). Finally, perhaps because branchpoints are an essential sequence feature required for splicing catalysis (in contrast to other intronic sequence elements that influence splice site recognition but are not universally required for splicing), branchpoints have frequently been assumed to play basal, rather than regulatory, roles.

Nonetheless, several lines of evidence suggest that branchpoint selection may frequently contribute to regulated splice site recognition in human cells. Detailed studies of specific introns revealed that both alternative and constitutive introns may have multiple branchpoints associated with a single 3′ splice site. This branchpoint degeneracy was initially observed in SV40 early pre-mRNA, which is alternatively spliced into the large T and small t mRNAs via an alternative 5′ splice site. This intron contains multiple branchpoints associated with a single 3′ splice site, six of which are used to generate large T mRNAs and one of which is used to generate small t mRNAs (Noble et al., 1987). Multiple branchpoints were subsequently found to be associated with single 3′ splice sites of the adenovirus E1a (Gattoni et al., 1988), rat Tpm1 (Helfman and Ricci, 1989), and human GH1 and HTR4 (Hartmuth & Barta, 1988; Hallegger et al., 2010) genes as well as within a majority of 52 introns of 20 human housekeeping genes (Gao et al., 2008).

Introns with multiple branchpoints can be subject to branchpoint competition during both constitutive and alternative splicing. Studies of a variant of the first intron of β-globin, which was engineered to contain a duplicated branchpoint sequence, revealed that branchpoint competition can occur even within constitutive introns (Zhuang et al., 1989). Branchpoint competition similarly contributes to alternative splicing. The contexts and positions of competing branchpoints can influence the recognition of competing 3′ splice sites (Reed & Maniatis, 1988; Smith et al., 1993; Bradley et al., 2012), competing 5′ splice sites (Noble et al., 1988), cassette exons (Kol et al., 2005; Corvelo et al., 2010) and mutually exclusive exons (Southby et al., 1999; Mullen et al., 1991).

These previous studies of specific introns suggest that redundant branchpoints may be common, potentially permitting regulated or cell type-specific recognition of many splice sites and introns. However, systematically identifying roles for branchpoint selection in splicing regulation has been hindered by the lack of a genome-wide branchpoint annotation as well as the difficulty in quantifying branchpoint usage.

Recent studies have made significant progress toward generating partial genome-wide branchpoint annotations. Even though lariats are transient RNAs with unusual chemical linkages, they can nonetheless be reverse-transcribed and incorporated into cDNA libraries. The branchpoint associated with a given lariat can then be identified by sequencing the junction between the 5′ splice site and the lariat. Because reverse transcriptase frequently incorporates a mismatch, insertion, or deletion when traversing the 2′–5′ phosphodiester linkage at the branch, the precise branchpoint location can be mapped by identifying putative 5′ splice site–branchpoint junctions where the sequenced cDNA has a mismatch specifically at the inferred branchpoint location (Vogel et al., 1997; Gao et al., 2008). Taggart et al. (2012) exploited the occasional incorporation of lariats into cDNA libraries to perform the first de novo branchpoint identification using RNA sequencing (RNA-seq), identifying 862 branchpoints in the human genome. Mercer et al. (2015) later created lariat-enriched cDNA libraries with RNase R digestion and targeted RNA recovery to identify 59,359 branchpoints, the largest annotation to date. These genome-wide studies identified a minority of introns with multiple branchpoints (9% in Taggart et al. [2012] and 32% in Mercer et al. [2015]). However, most of those branchpoints were annotated based on just one or a few sequenced lariats. Therefore, it is possible that undersampling of lariats resulted in underestimates of branchpoint multiplicity in those previous studies. Furthermore, because only

one or a few lariats were observed for most branchpoints (an inevitable consequence of the very low abundance of lariat RNA species), quantitative estimates of branchpoint usage remain elusive.

Here, we sought to determine whether branchpoints are typically just basal sequence features of introns or whether branchpoint recognition is frequently complex or regulated. We performed a very large-scale analysis to systematically identify branchpoints and quantify their usage across diverse human tissues in both normal and diseased states. Our results indicate that almost all human introns have multiple branchpoints, which are frequently used in a tissue-specific manner. Branchpoint abundance correlates with alternative splicing. Our data demonstrate that branchpoint recognition is unexpectedly complex, giving rise to cell type-specific splice site recognition during both constitutive and alternative splicing.

2.3 RESULTS

2.3.1 *A large-scale analysis of RNA-seq data enables global branchpoint annotation*

We sought to create a genome-wide branchpoint annotation by taking advantage of the occasional reverse transcription of lariats and their subsequent incorporation into cDNA libraries. The transient nature of lariat RNA as well as the specific selection of polyadenylated RNA in many RNA-seq library construction protocols render lariat incorporation rare. Informative reads from lariats—those that span the junction between the 5′ splice site and a branchpoint rather than simply lying within the intron—are even rarer. To address this statistical challenge, we performed an extremely large-scale analysis of ~1.31 trillion reads from 17,164 RNA-seq data sets (**Supplemental Table S1**). These data sets were generated from healthy as well as diseased tissues, including ~550 billion reads from the Genotype-Tissue Expression (GTEx) project's survey of healthy tissues (Melé et al., 2015) and ~490 billion reads from The Cancer Genome Atlas's survey

of primary and peritumoral tissues from diverse cancers. Together, these 17,164 data sets represent a comprehensive survey of cell types and physiological states.

For each RNA-seq data set, we identified lariat-derived reads that spanned 5′ splice site–branchpoint junctions by sequentially aligning reads to the transcriptome, genome, 5′ splice sites, and 3′ splice sites (**Fig. 1A**; **Supplemental Fig. S1**). This alignment strategy was modeled after the "split-read" approach used by Mercer et al. (2015). In brief, we first prefiltered each RNA-seq data set by removing all reads that aligned to the transcriptome or genome. We then aligned the remaining reads to a database of all annotated 5′ splice sites, requiring a minimum of 20 nucleotides (nt) of aligned sequence. We trimmed each read alignment to remove the 5′ splice site sequence and then aligned each trimmed read to a database of all annotated 3′ splice sites, requiring complete alignment of the trimmed read with a minimum of 20 nt of aligned sequence within 250 nt of the 3′ splice site itself. We then restricted to reads that aligned to 5′ and 3′ splice sites within a single gene in the "inverted" pattern (e.g., where the end of the read maps upstream of the start of the read) expected of reads arising from lariats rather than linear introns. The inferred branchpoint location is then the last nucleotide of the alignment of the trimmed read to the 3′ splice site. Finally, in order to obtain nucleotide-level resolution of branchpoint locations, we restricted to reads with a mismatch at the inferred branchpoint location. Such mismatches are strongly associated with correctly inferred branchpoints, as reverse transcriptase frequently incorporates an incorrect nucleotide when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al., 1997; Gao et al., 2008).

Manual inspection of the resulting data set revealed that we comprehensively annotated 5′ splice site–branchpoint pairs for many genes. For example, we identified branchpoints within all but one intron of the gene encoding the splicing factor SRSF5. Our 5′ splice site–branchpoint pairs

revealed complex splicing patterns for SRSF5, including alternative 5′ splice site usage, skipping of multiple cassette as well as constitutive coding exons, and usage of branchpoints that were proximal as well as distal to 3′ splice sites (**Fig. 1B**).

2.3.2 *Comparison with published branchpoint annotations*

As the split-read alignment procedure can be confounded by gene duplications or the presence of other repetitive genomic DNA, we assigned a confidence level to each inferred 5′ splice site–branchpoint pair. For a 5′ splice site–branchpoint pair to meet the highest confidence level, we required (1) that ≥ 5% of supporting reads have mismatches at the branchpoint but no other mismatches or indels (insertions/deletions) in the 3′ splice site alignment and (2) that the 25 nt downstream from the 5′ splice site, 25 nt upstream of the inferred branchpoint, and 25 nt of the lariat centered on the inferred branchpoint all be unique (not present in the transcriptome or genome). We successively relaxed these criteria for the moderate and low confidence levels. We removed the sequence uniqueness criteria for moderate-confidence branchpoints and allowed additional mismatches and indels in the 3′ splice site alignment for low-confidence branchpoints. We identified a total of 136,998, 9182, and 48,935 5′ splice site–branchpoint pairs at high, moderate, and low confidence levels. Our branchpoint annotations were robust with respect to the specific details of the thresholds used. For example, requiring that ≥25% of supporting reads have mismatches at the branchpoint but not other mismatches or indels in the 3′ splice site alignment (a fivefold increase in stringency) resulted in only a 2.9% decrease in the number of high-confidence 5′ splice site–branchpoint pairs that we identified.

We next assessed the likely accuracy of our branchpoint inference procedure for each confidence level. Biochemical studies and lariat sequencing have revealed that adenosine is the most effective and frequent branchpoint ribonucleotide (Gao et al., 2008), suggesting that global

branchpoint adenine frequency correlates with inference accuracy. Branchpoints that we identified at high, moderate, and low confidence levels had adenine frequencies of ∼77%, 50%, and 32%, indicating that our confidence levels correlate with likely inference accuracy. Therefore, we restricted all subsequent global analyses to 5′ splice site–branchpoint pairs detected at the highest confidence level.

We next compared our branchpoint annotations with previously published branchpoint data sets (**Table 1**). We identified 70,935 and 94,216 more branchpoints than were reported in Mercer et al. (2015) and Taggart et al. (2017), the largest sets of branchpoint annotations published to date. (For comparison with Taggart et al. [2017], we used their "high-confidence" set of branchpoints.) Our annotation exhibited a branchpoint adenine frequency of ∼77% versus 78% and 55% for Mercer et al. (2015) and Taggart et al. (2017). The lower adenine frequency for the annotation of Taggart et al. (2017) may be due to differences in the methods that each study used to call branchpoints. Like Mercer et al. (2015), we restricted to reads with a mismatch at the inferred branchpoint, which is diagnostic of reverse transcriptase incorporating an incorrect nucleotide when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al., 1997; Gao et al., 2008). In contrast, Taggart et al. (2017) did not require a mismatch at the inferred branchpoint. Instead, they aligned putative branchpoint sequence contexts to the U2 small nuclear RNA (snRNA) sequence and called branchpoints at the inferred bulged nucleotide (Taggart et al., 2017).

2.3.3 *Parent gene expression and intron length determine branchpoint detection rate*

Despite the extremely large-scale nature of our analysis, we did not approach saturation. We estimated the branchpoint detection rate as a function of the number of sequenced lariats by randomly sampling from all branchpoint-spanning reads. Even though we detected many more

branchpoints than did previous studies, branchpoint detection continued to increase rapidly as a function of the number of sequenced lariats throughout the dynamic range of our study (**Fig. 1C**).

We detected one or more branchpoints within ~37% and 42% of U2- and U12-type constitutive introns present in the RefSeq annotation (O'Leary et al., 2016), where we defined constitutive introns as those that were present in all child transcripts of a given RefSeq gene (**Fig. 1D**). Fifty percent of detected branchpoints fell within RefSeq constitutive introns, while 35% fell within nonconstitutive introns present in the RefSeq, University of California at Santa Cruz (UCSC), Ensembl, or Mixture of Isoforms (MISO) isoform databases (**Fig. 1E**; Katz et al., 2010; Flicek et al., 2013; Meyer et al., 2013). An unexpectedly large percentage (14%) of 5′ splice site–branchpoint pairs corresponded to introns that were not annotated in any of those isoform databases, resulting from skipping of one or more ostensibly constitutive exons. While some such cases may correspond to stable isoforms with potential cellular functions, many may simply represent by-products of splicing mistakes.

The numbers of branchpoints that we detected per intron or gene were highly variable. We obtained seemingly near-complete annotations for some genes (e.g., SRSF5 in **Fig. 1B**) and few or no branchpoints for any introns of other genes. This high level of variability in branchpoint detection could arise from many factors, including differences in parent gene expression, intron length, and lariat stability. We tested whether each of these factors contributed to differences in branchpoint detection rate. We restricted these power analyses to constitutive introns in order to avoid additional complexities arising from alternative splicing. The branchpoint detection rate was strongly positively and negatively correlated with parent gene expression and intron length, as expected from random sampling of lariats (**Supplemental Fig. S2A–C**). Branchpoints from very

short (<200-base-pair [bp]) introns were underrepresented, presumably because most RNA-seq protocols intentionally deplete such short RNAs during library preparation.

We tested whether some lariats had unusually short or long half-lives by comparing the observed abundance of each lariat with its expected abundance, defined as the ratio of its parent gene expression to intron length. The distribution of observed to expected abundances followed a normal distribution across all sequenced lariats, consistent with a model in which lariats are degraded randomly. Lariats with a guanine branchpoint exhibited a 1.6-fold greater abundance than expected, suggesting that they are frequently more stable than lariats with adenine, cytosine, or thymine/uracil branchpoints (**Supplemental Fig. S2D**). These findings are consistent with a previous report that lariats formed from nonadenine mutants of the rabbit HBB gene were resistant to debranching relative to lariats formed via the wild-type adenine branchpoint (Hornig et al., 1986).

2.3.4 *Distal branchpoints contribute to alternative exon and intron recognition*

We used our genome-wide branchpoint annotation to identify sequence features contributing to branchpoint recognition and usage. Branchpoints within constitutive introns were most frequently adenine (82.5%), followed by guanine (8.7%), thymine/uracil (4.7%), and cytosine (4.1%). We observed a modest preference for thymine/uracil at the −2 position relative to the branchpoint, as reported previously (Gao et al., 2008). However, this preference was restricted to adenine branchpoints, with no site-specific sequence preferences at any other nucleotides for nonadenine branchpoints (**Fig. 2A**).

Branchpoints exhibited a tightly constrained spatial distribution, as reported by previous studies (Taggart et al. 2012, 2017; Mercer et al. 2015). Branchpoints within U2-type constitutive introns were positioned at a median of 28 nt upstream of the 3′ splice site, with 80% of such

branchpoints found within the positions −49 and −20 nt. Branchpoints within U12-type introns exhibited a bimodal distribution (**Fig. 2B**). Approximately half of such U12-type branchpoints were found in close proximity (within 20 nt) of the 3′ splice site, as observed previously (R. C. Dietrich et al., 2001; Taggart et al., 2017). In contrast, approximately half of U12-type introns were located only modestly closer to the 3′ splice site than we observed for U2-type branchpoints. We classified introns as U2- or U12-type by finding the best match between each 5′ splice site sequence to the U2- and U12-type consensus sequences. The U2- and U12-type 5′ splice site consensus sequences are distinct (Sheth et al., 2006), making frequent misclassification unlikely. However, we cannot rule out the possibility that classification error contributes to the unexpected bimodal spatial distribution for U12-type branchpoint positions.

While most branchpoints were positioned proximal to the 3′ splice site, a subset was located further upstream. Distal branchpoints, located ≥50 nt upstream, constituted only 9.5% of branchpoints in U2-type constitutive introns (**Fig. 2B**). In contrast, distal branchpoints frequently occurred in introns associated with alternative splicing events, consistent with previous reports (Corvelo et al., 2010; Taggart et al., 2012, 2017). We quantified alternative splicing across 16 human tissues and restricted to "switch-like" events that exhibited changes in isoform ratio ("switch scores") of ≥ 25% between tissues. This restriction focused our analysis on regulated tissue-specific splicing rather than low-abundance isoforms that might result from stochastic splicing. Distal branchpoints occurred at frequencies of 39.5% and 28.7% within U2-type introns that were frequently retained or positioned upstream of cassette exons (**Fig. 2C,D**). Far-distal branchpoints, located ≥100 nt upstream, occurred at frequencies of 4.6%, 22.0%, and 13.9% in U2-type constitutive introns, retained introns, and introns upstream of cassette exons. This

unexpectedly strong enrichment for distal and far-distal branchpoints in switch-like retained introns strongly suggests that branchpoint position contributes to regulated intron recognition.

2.3.5 *Almost all constitutive introns have multiple branchpoints*

We anecdotally noticed that many introns contained multiple annotated branchpoints. This branchpoint multiplicity was common even in constitutive introns, which are not subject to alternative splice site usage yet frequently contain an unexpectedly large number of branchpoints. Given this surprising degree of branchpoint multiplicity, we sought to confirm the results of our high-throughput branchpoint inference procedure with direct lariat sequencing. We selected four constitutive introns within *MBNL1*, *POLR3A*, *SNX9*, and *VASP*, each of which exhibited high branchpoint multiplicity, with six or seven branchpoints discovered within RNA-seq libraries from the K562 erythroleukemic cell line alone. We generated cDNA libraries from K562 cell lysate, used nested PCR to specifically amplify lariats from each of those four introns, performed Sanger sequencing on single amplicons with colony sequencing, and inferred branchpoints from each sequenced amplicon (**Supplemental Fig. S3A,B**). For each intron, we validated the majority of computationally inferred branchpoints and furthermore discovered new branchpoints (**Fig. 3A,B**). In addition to experimentally confirming the striking branchpoint multiplicity that we inferred for many introns, these results demonstrated that many or most introns are still undersampled despite our very large-scale RNA-seq analysis.

We next tested whether branchpoint multiplicity was an unusual feature of specific introns or was instead a common characteristic of many introns. Accurately estimating branchpoint abundance is challenging for two reasons. First, as revealed by our power analysis and targeted lariat sequencing experiments, our study has not approached saturation of lariat sequencing even

for introns with many annotated branchpoints (**Figs. 1C, 3A,B**). Second, our lariat sequencing depth varied by orders of magnitude for different introns.

We simultaneously controlled for both of those effects by stratifying all analyses by per-intron lariat sequencing depth. Simply binning each intron according to the number of sequenced lariats revealed that the vast majority of constitutive introns contained multiple lariats. Ninety-five percent of constitutive introns with the greatest sequencing depth (≥ 250 sequenced lariats) contained two or more distinct branchpoints, with a mean of 6.75 branchpoints detected per intron (**Fig. 3C,D**). Sequencing just five to 10 lariats per intron was sufficient to detect multiple branchpoints in the majority of introns. Ninety-five percent is probably an accurate estimate of the fraction of constitutive introns with multiple branchpoints, as an asymptote is clearly evident in our power analysis (**Fig. 3C**). In contrast, additional lariat sequencing will probably reveal novel branchpoints for the 95% of introns exhibiting branchpoint multiplicity (no asymptote is visible in the relevant power analysis) (**Fig. 3D**), consistent with our discovery of novel branchpoints via direct lariat sequencing of *MBNL1*, *POLR3A*, *SNX9*, and *VASP* introns.

Our estimates of branchpoint multiplicity could potentially be confounded by small nontemplated insertions or deletions generated by reverse transcriptase when traversing the 2′–5′ phosphodiester linkage at the branch (Vogel et al., 1997; Gao et al., 2008). Deletions do not confound our analysis, as they do not result in a mismatch at the branchpoint. However, insertion of a single nucleotide could result in incorrect inference of a branchpoint at the +1 position with respect to the actual branchpoint. (Insertion of two or more nucleotides, which is relatively infrequent, would result in multiple mismatches. Such reads would not satisfy our criteria for high-confidence branchpoints, except for the unlikely case where the randomly inserted nucleotides matched the genomic sequence.) To test whether our branchpoint multiplicity estimates were

biased by this potential source of error, we took the conservative approach of collapsing all adjacent branchpoints into a single branchpoint. Even after applying this merge procedure, we estimated that ~94% of constitutive introns contained multiple branchpoints, with a mean of five branchpoints per intron for introns with the most lariat sequencing coverage (**Supplemental Fig. S4A,B**). We conclude that high branchpoint multiplicity typifies the vast majority of human introns.

2.3.6 *Branchpoint position strongly influences branchpoint usage*

We next attempted to identify sequence features that contribute to basal branchpoint recognition and selection in the face of high branchpoint multiplicity. We took advantage of the large-scale nature of our study to quantitatively estimate branchpoint usage across 54 healthy human tissues. We removed transcriptome- or genome-aligning reads from the ~550 billion reads sequenced by the GTEx project (Melé et al., 2015) and aligned the remaining reads to all lariat sequences (5′ splice site–branchpoint pairs). We restricted to reads with a lesion (mismatch or small indel) specifically at the branchpoint in order to help ensure that the reads originated from reverse transcription of branched RNA. For each tissue, we collated reads sampled from different individuals in order to increase our lariat sequencing depth. We then estimated branchpoint usage by computing the frequency with which a particular 5′ splice site–branchpoint pair was used relative to all branchpoints associated with that 5′ splice site. As we sought to identify sequence features that influenced basal branchpoint recognition independent of potential cis- or transacting regulation, we estimated basal branchpoint usage by averaging branchpoint usage across all 54 tissues.

We focused on the two key features that define a branchpoint: its location relative to the 3′ splice site and its complementarity to the U2 snRNA sequence. We restricted to U2-type introns and focused our analysis on the two most frequently used branchpoints within each intron. Plotting quantitative branchpoint usage as a function of branchpoint position revealed that the majority of most frequently used branchpoints resided within a narrow window, consistent with the restricted genome-wide distribution of all branchpoint positions (**Fig. 3E**). While a few far-distal branchpoints were predominantly used, such examples were relatively uncommon.

In contrast to branchpoint position, complementarity to the U2 snRNA was not strongly correlated with branchpoint usage. We computed the binding energy of each branchpoint sequence context to the U2 snRNA sequence AUGAUGUG, with the exception of the branchpoint itself, which appears as a bulge in the structure. While previous studies have shown that U2 snRNA complementarity is associated with branchpoint recognition, this association is very weak (Mercer et al. 2015). Consistent with previous results, we observed little association between U2 snRNA complementarity and quantitative branchpoint usage (**Fig. 3F**). Many branchpoints were very poor matches to the U2 snRNA yet were predominantly used. In contrast, most branchpoints within U12-type constitutive introns were comparatively better matches to the U12 snRNA sequence AGGAAUG (**Supplemental Fig. S4C**).

2.3.7 *Highly regulated ultraconserved introns have an unusually high number of branchpoints*

Since constitutive introns exhibited such a surprising degree of branchpoint multiplicity, we hypothesized that introns that were associated with alternative splicing might exhibit even more. Manual inspection of specific introns flanking highly regulated cassette exons, such as the "poison" exons of *SRSF5* and *SRSF3*, supported this hypothesis (**Figs. 1B, 4A**). *SRSF5* and *SRSF3*

are members of the SR gene family, each of which contains a highly regulated "poison" splicing event that introduces an in-frame premature termination codon into the mature transcript. Poison exons contribute to SR splicing factor homeostasis and overlap with ultraconserved or highly conserved genomic sequence (Lareau et al., 2007; Ni et al., 2007).

We first confirmed that the large branchpoint cluster upstream of the *SRSF3* poison exon was correctly annotated with direct lariat sequencing. As with our studies of constitutive introns, direct lariat sequencing of the *SRSF3* intron both confirmed the computationally inferred branchpoint cluster and revealed novel branchpoints (**Fig. 4B**). These branchpoints were spread throughout the highly conserved intronic region upstream of the poison exon, suggesting that they likely contribute to the purifying selection acting on this genomic sequence.

We next tested whether branchpoint abundance was associated with alternatively spliced sequences at a genome-wide level. We classified introns as constitutive, retained, upstream of or downstream from cassette exons, containing a cassette exon, or containing competing 5′ or 3′ splice sites. We considered introns that were associated with poison exons of SR genes as a distinct class. All introns associated with the inclusion of alternatively spliced sequence were enriched for branchpoints relative to constitutive introns, with retained introns displaying the greatest enrichment (~43%) (**Fig. 4C**). While introns upstream of as well as downstream from cassette exons were enriched for branchpoints, introns corresponding to cassette exon exclusion exhibited a modest depletion relative to constitutive introns, suggesting that some branchpoints downstream from cassette exons are used only in the context of exon inclusion. (This comparison was made possible by our enumeration of 5′ splice site–branchpoint pairs rather than branchpoints alone.) We observed the same trend, although with much greater branchpoint multiplicity, for introns associated with poison exons of SR genes.

In addition to exaggerated multiplicity, branchpoints within introns that were associated with alternative splicing exhibited other unusual characteristics. Adenine is found at ~83% of branchpoints within constitutive introns but only ~58% of branchpoints within retained introns (**Fig. 4D**). Adenine frequencies are highest for 5′ splice site–branchpoint pairs corresponding to cassette exon exclusion (~88%), mirroring our observation that cassette exon exclusion is associated with reduced branchpoint multiplicity even relative to constitutive intron splicing. Introns flanking the poison exons of SR genes contained branchpoints within sequence contexts that were unusually poor matches to the U2 snRNA consensus, with an average of −0.4 kcal/mol for introns upstream of SR poison exons versus −1.1 kcal/mol for constitutive introns (**Fig. 4E**). Together, our data indicate that an abundance of branchpoints, many of which are suboptimal, likely contributes to regulated alternative splicing.

2.3.8 *Branchpoint usage is frequently tissue-specific*

The branchpoint multiplicity that characterizes most introns theoretically permits tissue-specific branchpoint selection and intron recognition even for constitutive introns. We anecdotally noticed a striking example of this within the first intron of *HBB* (encoding β-globin), a well-studied splicing substrate. Early biochemical studies demonstrated that *HBB*'s first intron forms a lariat RNA via an adenine branchpoint at position −37 nt relative to the 3′ splice site (Ruskin et al., 1984). Mutating this branchpoint to a guanine did not abolish in vitro splicing of its parent intron. Instead, branchpoint usage shifted to a 3′ splice site-proximal adenine located at position −24 nt (Ruskin et al., 1985). While the dominant branchpoint at position −37 nt is an excellent match to the U2 snRNA, with a binding energy of −5.2 kcal/mcol, the cryptic branchpoint at position −24 nt has a binding energy of just −0.5 kcal/mol. This difference may explain why the downstream branchpoint was used in vitro only when the dominant branchpoint was mutated.

Given these biochemical studies, we expected to observe exclusive usage of the −37-nt branchpoint in our own data for *HBB*. Unexpectedly, we found that branchpoint usage was instead highly tissue-specific, with nonoverlapping sets of branchpoints used in blood versus metastatic prostate cancer (**Fig. 5A**). The −37-nt branchpoint was present in 79% of lariats sequenced from normal or leukemic peripheral blood or bone marrow, with infrequent usage of other branchpoints at positions −78, −41, and −24 nt. In contrast, in metastatic prostate cancer samples, branchpoints at positions −30 nt and −26 nt constituted 31% and 58% of branchpoint usage. The −37-nt branchpoint was virtually unused.

Since we observed such striking variation in branchpoint usage even within the well-studied first intron of *HBB*, we hypothesized that tissue-specific branchpoint usage might be more common than is currently recognized. We therefore sought to use direct lariat sequencing to identify differentially used branchpoints within the *VASP* and *SRSF3* introns studied above as exemplars of constitutive and alternative splicing. We first confirmed that our direct lariat sequencing protocol was sufficiently reproducible to quantify differential branchpoint usage. We amplified, cloned, and sequenced lariats from the *VASP* and *SRSF3* introns in technical duplicates from five tissues obtained from healthy donors (peripheral blood, cerebellum, spleen, fetal spleen, and testis). We estimated false discovery rates (FDRs) for each intron by measuring the frequencies of differential branchpoint usage between the technical replicates for each of the [(number of branchpoints) × (number of tissues)] trials, where we defined differential branchpoint usage as differences in usage of ≥ 10% with a P-value of ≤ 0.01. We estimated FDRs of 2.5% and 2.4% for *VASP* and *SRSF3*, indicating that our assay is robust with respect to experimental variability.

We therefore used our lariat sequencing assay to quantify branchpoint usage for all branchpoints within the *VASP* and *SRSF3* introns. We observed frequent differential branchpoint

usage for both introns, including differences between tissues as well as between fetal and adult samples from the same tissue (**Fig. 5B,C**). Both the *VASP* and *SRSF3* introns contained "switch-like" branchpoints, which were never used in some tissues but were used frequently in others. The four and five differentially used branchpoints that we detected in *VASP* and *SRSF3* far exceeded the number expected from experimental variability alone, with associated P-values of 5.6×10^{-3} and $1.9 \times 10{-}5$ (**Fig. 5D**).

We next extended our targeted experimental analysis of *VASP* and *SRSF3* to a genome-wide RNA-seq-based measurement of differential branchpoint usage. We estimated branchpoint usage across healthy human tissues in the GTEx data set and performed a power analysis similar to our approach for estimating genome-wide branchpoint multiplicity (**Fig. 3C,D**). We focused our analysis on constitutive introns, since branchpoint multiplicity and differential branchpoint usage in the context of constitutive splicing was so unexpected. We restricted to constitutive introns with two or more branchpoints, binned each intron according to the total number of sequenced lariats across all tissues, and tested whether each intron exhibited tissue-specific differences in branchpoint usage.

Our power analyses suggested that most branchpoints within constitutive introns are used relatively frequently within one or more tissues and that a majority of constitutive introns undergoes tissue-specific branchpoint usage (**Fig. 5E**). After binning introns by the total number of sequenced lariats, we found that ~87% of branchpoints within the highest-coverage bin (total of ≥ 250 sequenced lariats over all tissues) were used at rates of $\geq 10\%$ in one or more tissues. Fifty-eight percent of constitutive introns within this highest-coverage bin exhibited tissue-specific branchpoint usage. We obtained even more striking results after binning introns by the mean number of sequenced lariats per tissue. This method more accurately controlled for how variable

sequencing depth affected our power to quantify branchpoint usage. Approximately 96% of branchpoints within the highest-coverage bin (mean of ≥ 10 sequenced lariats per tissue) were used at rates of ≥ 10% in one or more tissues, and 81% of constitutive introns in this coverage bin exhibited tissue-specific branchpoint usage. We conclude that even constitutive introns commonly undergo tissue-specific branchpoint usage.

2.4 DISCUSSION

In addition to providing a comprehensive genome-wide branchpoint annotation, our study has several important implications for future studies of splicing mechanisms and regulation. First, our finding that most introns have multiple branchpoints suggests that any perturbation of branchpoint recognition may have unexpectedly profound consequences for global splicing. Branchpoint multiplicity may be particularly important in the context of cancer-associated mutations that alter normal splicing mechanisms and regulation (Dvinge et al., 2016). Second, our study demonstrates that branchpoint selection is unexpectedly complex in healthy tissues, even for constitutive introns. The striking tissue-specific variability in branchpoint usage that we observed suggests that introns are recognized in mechanistically distinct ways in different cell types.

The discovery of recurrent cancer-associated mutations affecting the splicing factor *SF3B1* created intense interest in understanding how these mutations might alter normal splicing (Papaemmanuil et al., 2011; Quesada et al., 2011; L. Wang et al., 2011; Yoshida et al., 2011). *SF3B1* is a core component of U2 snRNP that binds pre-mRNA near the branchpoint. While the mechanistic consequences of *SF3B1* mutations have not been fully elucidated, several studies have demonstrated that these lesions are associated with abnormal 3′ splice site recognition, including usage of cryptic 3′ splice sites and alternate branchpoints (Darman et al., 2015; DeBoever et al., 2015; Alsafadi et al., 2016). However, relatively few splicing changes have been identified to date

in *SF3B1* mutant cells (Obeng et al., 2016). Given the unexpected branchpoint multiplicity and tissue-specific regulation revealed by our study, we speculate that *SF3B1* mutations might have more profound and pervasive consequences for global splicing than is currently recognized.

While our study highlights the complexity of recognizing even constitutive introns, the mechanistic origins of tissue-specific branchpoint usage remain mysterious. For example, it is unclear why different sets of branchpoints underlie *HBB* splicing in blood versus metastatic prostate cancer samples. This tissue specificity is not readily explained by somatic mutations, as the analyzed metastatic prostate cancer samples were not enriched for recurrent splicing factor mutations (Robinson et al., 2015a). Interestingly, within each of the two distinct sets of branchpoints, the branchpoint with the best match to the U2 snRNA was dominant (**Fig. 5A**). Binding of trans-acting factors to the *HBB* intron might prevent branchpoint recognition by physical occlusion, and competition between nonoccluded branchpoints might govern subsequent patterns of branchpoint selection in a given tissue. Even a single splicing factor can occlude multiple branchpoints; for example, *CELF2* can bind sites flanking a branchpoint cluster to simultaneously prevent usage of any branchpoint in the cluster (Dembowski & Grabowski, 2009). Further studies are required to test whether this or other mechanisms enforce the tissue specificity of branchpoint usage within *HBB*, *VASP*, *SRSF3*, and other genes.

What are the functional consequences of branchpoint multiplicity? We speculate that having multiple branchpoints might confer both fitness advantages and expanded regulatory potential to introns. First, branchpoint multiplicity may confer mechanistic robustness. Having multiple branchpoints may render introns resilient to otherwise deleterious transcriptional errors, somatic mutations, or genetic variation. Second, branchpoint multiplicity may facilitate splicing regulation by rendering splice site recognition more plastic. For example, an intron with multiple

branchpoints could be regulated by the intronic binding of tissue-specific splicing factors that promote or repress individual branchpoints. Such an intron might have more inherent regulatory potential than would an intron with a single branchpoint. This hypothesis is supported by our finding that introns associated with the highly regulated SR poison exons are rich with branchpoints. Third, branchpoint multiplicity may enable regulated retention of introns, including ostensibly constitutive introns. A majority of human introns, most of which are not associated with alternative splice site or exon usage, exhibits detectable intron retention in specific healthy and/or cancerous cell types (Braunschweig et al., 2014; Dvinge & Bradley, 2015). With rare exceptions, the mechanistic origins of intron retention are not understood. However, our observation that the most frequently retained introns have more branchpoints than do other introns, including nonadenine and 3′ splice site-distal branchpoints, strongly suggests that branchpoint selection is an important contributor to regulated intron retention. While our understanding of branchpoint selection remains incomplete, it is clear that the branchpoint plays a more important regulatory role in both constitutive and alternative splicing than is generally recognized.

2.5 FIGURES

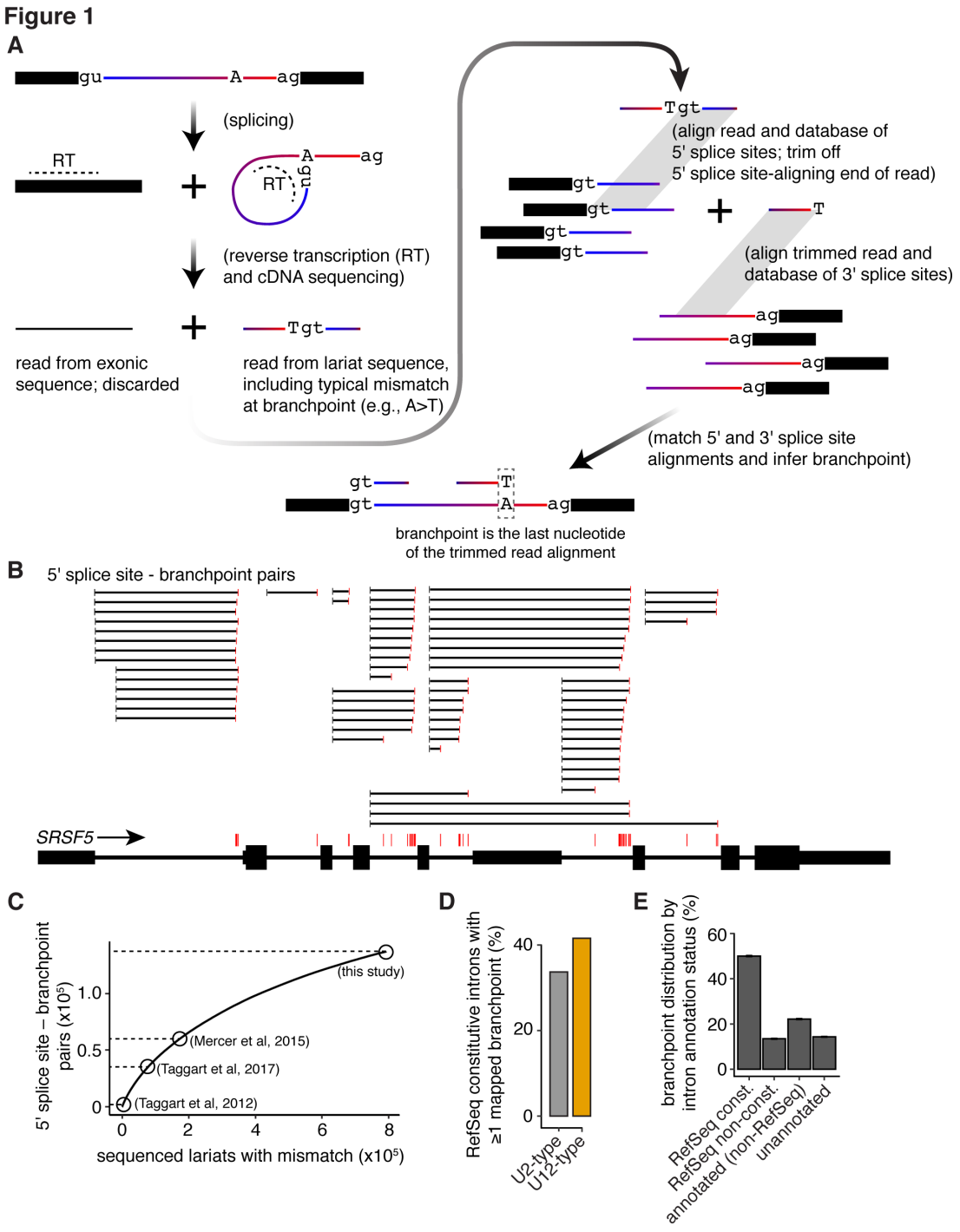


Figure 1. Genome-wide branchpoint annotation from RNA-seq data.
(**A**) Overview of our branchpoint detection algorithm. See also **Supplemental Figure S1** (**Figure S1**).
(**B**) Branchpoint annotation of SRSF5. For simplicity, only the intron-distal splice site of a competing 5' splice site event within the first intron is illustrated in the exon-intron structure.

Vertical red bars, branchpoints. Horizontal black lines, 5' splice site - branchpoint pairs. Plot based on image from the UCSC Genome Browser (Meyer et al. 2013).

(**C**) Branchpoint detection rate as a function of the number of sequenced lariats. We randomly sampled from all sequenced lariats analyzed in our study and computed the number of distinct 5' splice site - branchpoint pairs detected. As 5' splice site - branchpoint pairs were not reported by other studies, we illustrated the number of reported branchpoints instead. For Taggart et al, 2017, we illustrated their "high-confidence" set of branchpoints.

(**D**) Fraction of all RefSeq constitutive introns with one or more mapped branchpoints.

(**E**) Distribution of mapped branchpoints among different annotation classes. RefSeq const., RefSeq constitutive introns. RefSeq non-const., RefSeq non-constitutive introns. annotated (non-RefSeq), introns present in the UCSC, Ensembl, or MISO annotation databases, but not RefSeq. unannotated, introns formed by unannotated ligation of annotated 5' and 3' splice sites.

Figure 2

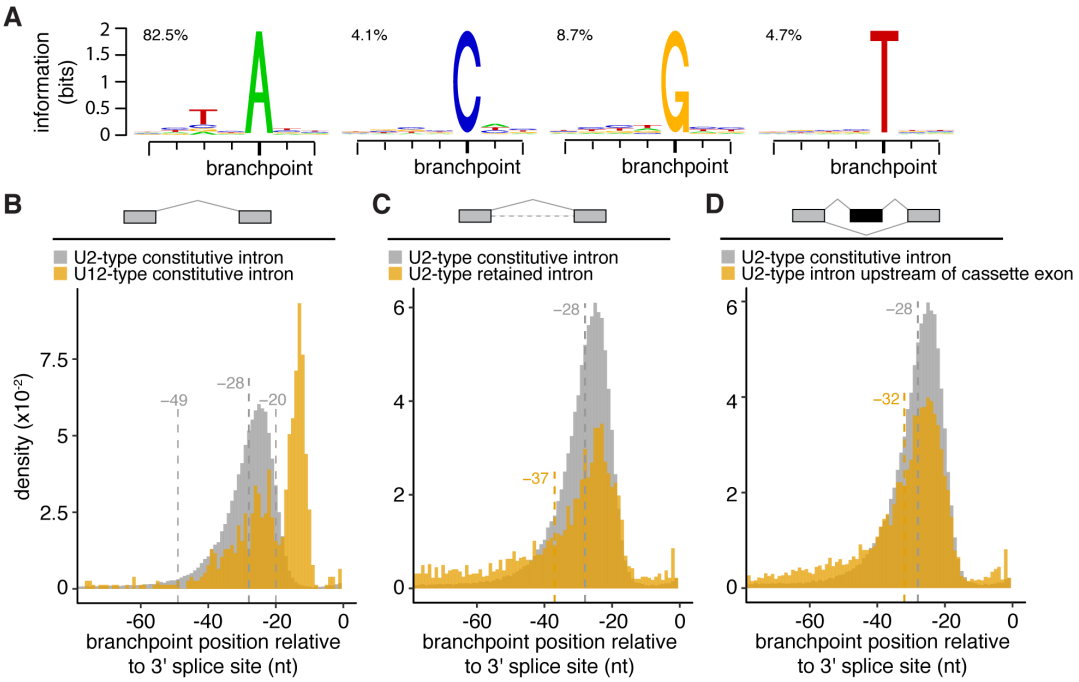


Figure 2. Branchpoint position, but not sequence context, is constrained.

(**A**) Sequence logos of branchpoint contexts. Plot restricted to branchpoints within RefSeq constitutive introns.

(**B**) Histogram of branchpoint positions relative to the 3' splice site, where position -1 nt corresponds to the last intronic nucleotide. Vertical dashed lines at -20, -28, and -49 nt illustrate the 10th, 50th, and 90th percentiles of positions for U2-type introns. Plot restricted to branchpoints within RefSeq constitutive introns.

(**C**) As (B), but for U2-type introns classified as constitutive or retained. To ensure that the analyzed sets of introns were disjoint, we restricted to constitutive introns that did not overlap introns annotated as potentially retained in the MISO v2.0 annotation, even if those introns did not exhibit retention in our data. Vertical dashed line at -28 nt illustrates the median position for constitutive introns.

(**D**) As (B), but for U2-type introns classified as constitutive or upstream of a cassette exon. retained. To ensure that the analyzed sets of introns were disjoint, we restricted to constitutive introns that did not overlap introns associated with cassette exons, even if those cassette exons did not exhibit alternative splicing in our data. Vertical dashed line at -28 nt illustrates the median position for constitutive introns.

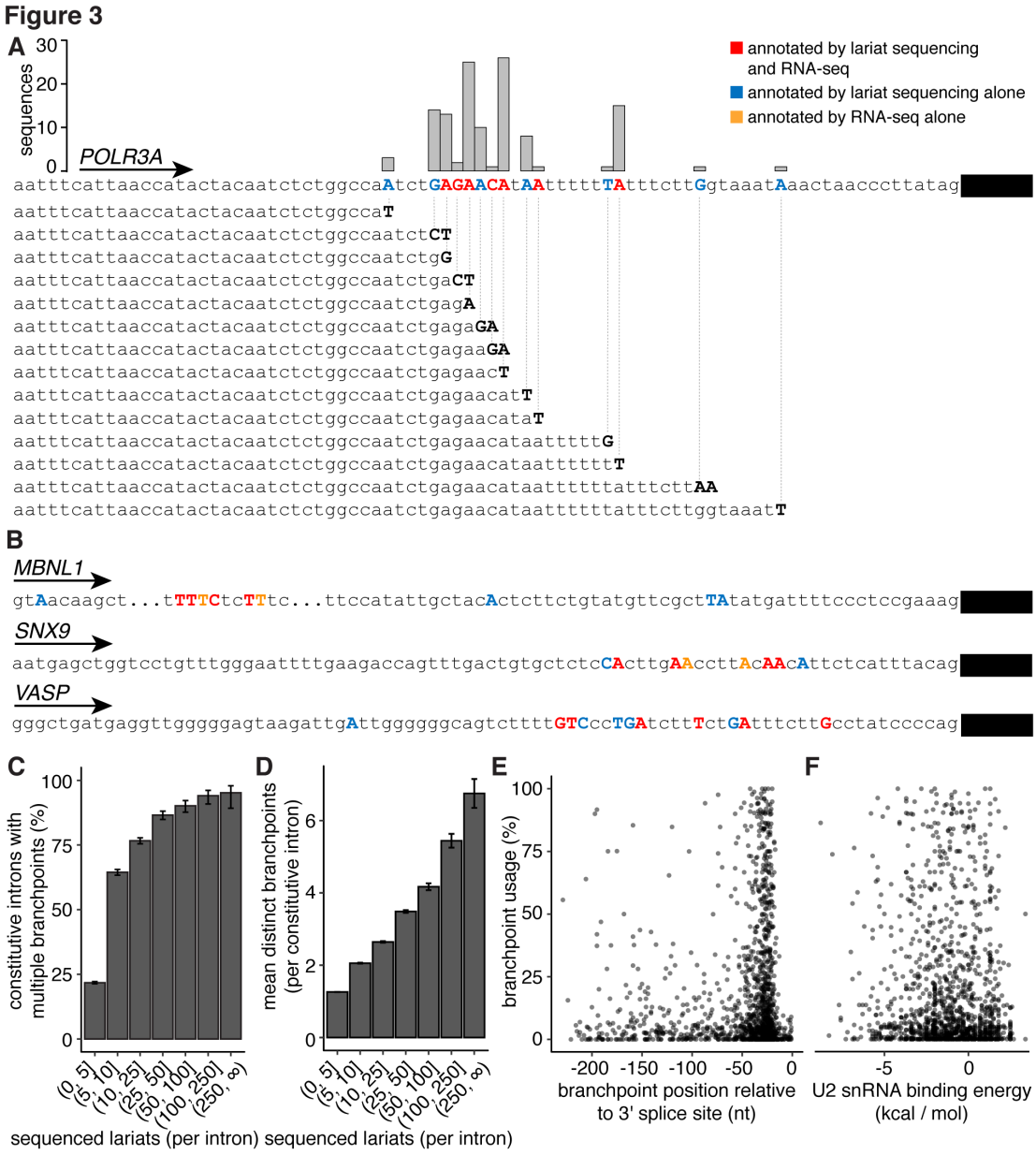


Figure 3. Most constitutively spliced introns contain multiple branchpoints.
(**A**) Branchpoint annotations of introns within *POLR3A* and (**B**) *MBNL1*, *SNX9*, and *VASP*, based on RNA-seq analysis as well as direct lariat sequencing. Colors indicate the evidence supporting each branchpoint. Examples of sequenced lariats are shown for *POLR3A*.
(**C**) Fraction of constitutive introns with multiple branchpoints as a function of the number of sequenced lariats with a mismatch at the branchpoint. Error bars, 95% confidence interval estimated with a proportion test.
(**D**) As (C), but illustrates the mean number of branchpoints per intron. Error bars, standard deviation of the mean estimated by bootstrapping.
(**E**) Branchpoint usage as a function of the relative branchpoint position. Branchpoint usage is defined as the number of sequenced lariats supporting a given 5' splice site - branchpoint pair divided by the total number of sequenced lariats mapped to that 5' splice site. Each point

corresponds to a single branchpoint. Plot restricted to constitutive introns with two or more branchpoints. The two most commonly used branchpoints per intron are illustrated.
(**F**) As (E), but illustrates estimated binding energy to the U2 snRNA sequence AUGAUGUG for each branchpoint context.

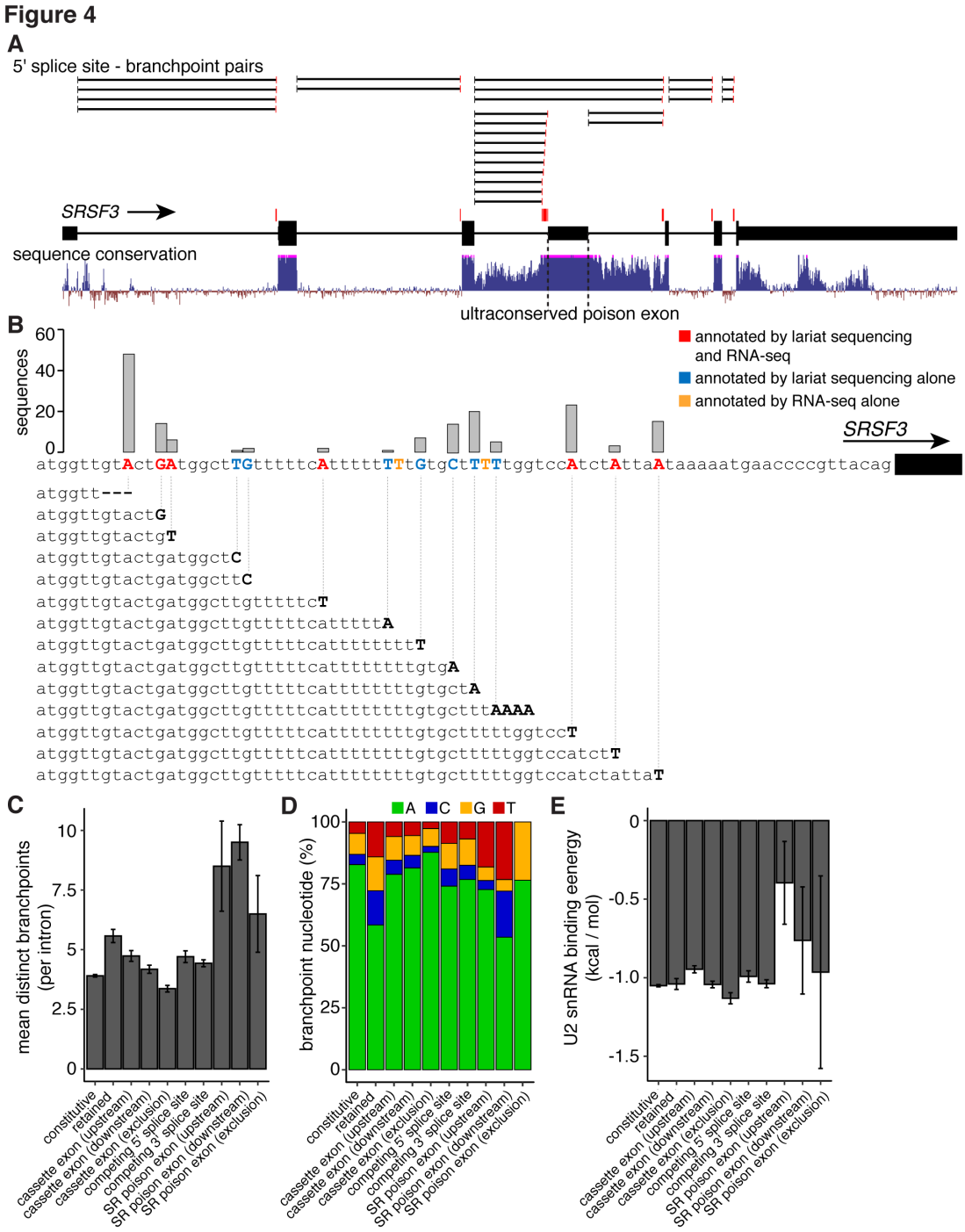


Figure 4. Regulated alternative splicing is associated with high branchpoint multiplicity.
(**A**) Branchpoint annotation for *SRSF3*. Sequence conservation, phastCons 100 vertebrates conservation track (Siepel et al., 2005). Plot based on image from the UCSC Genome Browser (Meyer et al. 2013).
(**B**) Branchpoint annotation for intron upstream of the *SRSF3* poison exon, based on RNA-seq analysis as well as direct lariat sequencing. Colors indicate the evidence supporting each branchpoint.

(**C**) Mean number of branchpoints detected in each of the illustrated classes of introns. Alternative splicing annotations were based on the MISO v2.0 isoform database (Katz et al. 2010). Plot restricted to introns with ≥ 25 sequenced lariats to help control for intron-specific variability in lariat sequencing depth. Error bars, standard deviation of the mean estimated by bootstrapping.
(**D**) As (C), but illustrates the frequencies with which each branchpoint nucleotide occurs.
(**E**) As (C), but illustrates the mean estimated U2 snRNA binding energy. Error bars, standard deviation of the mean estimated by bootstrapping.

Figure 5

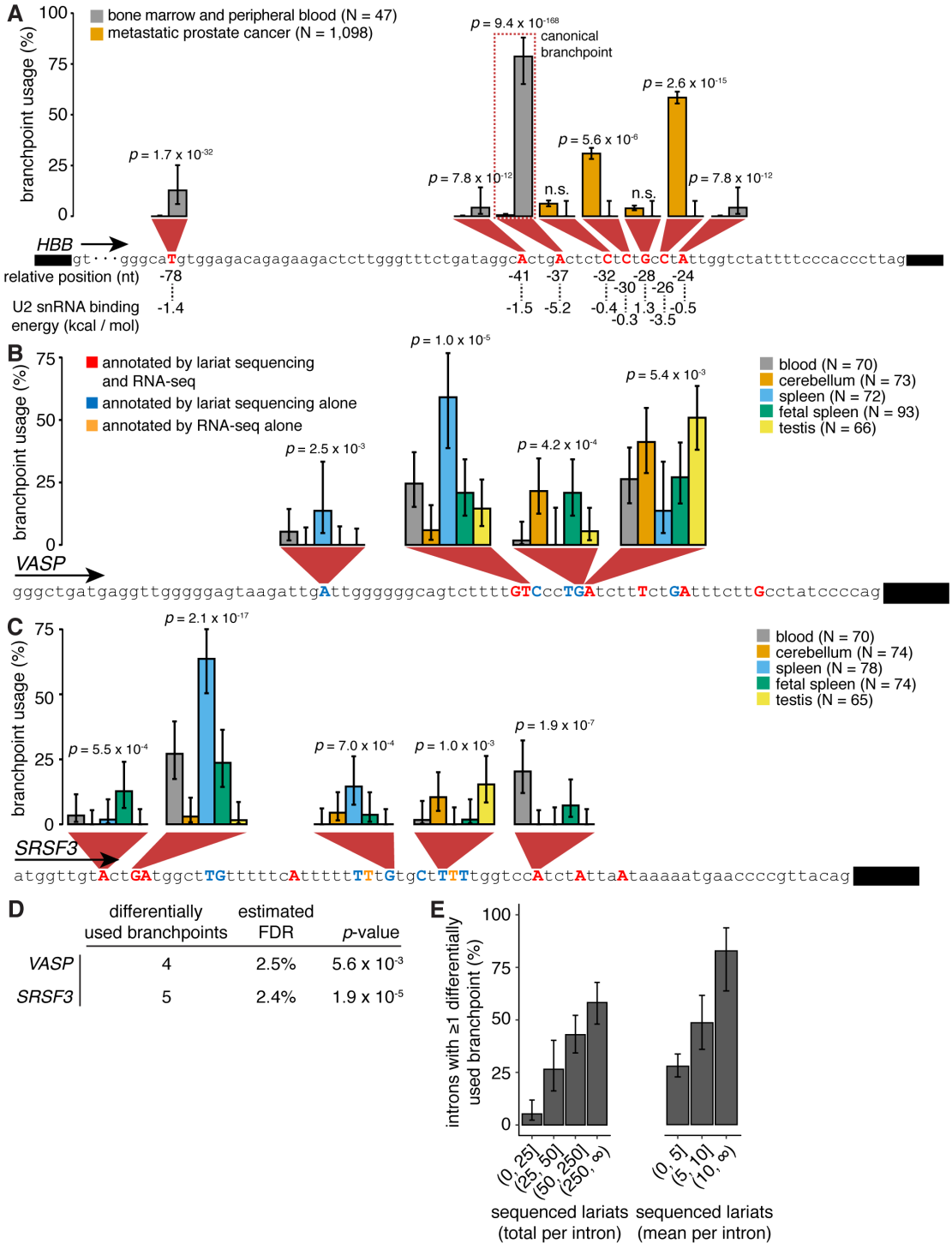


Figure 5. Tissue-specific branchpoint usage is common.

(**A**) Branchpoint annotation and estimated branchpoint usage for the first intron of *HBB*. N, number of sequenced lariats with a mismatch at the inferred branchpoint. Error bars, 95% confidence intervals estimated with the binomial proportion test. *p*-values were estimated with the binomial proportion test. Branchpoints at positions -32, -37 (the canonical branchpoint annotated

biochemically (Ruskin et al. 1984)), and -41 nt were annotated with moderate, rather than high, confidence due to the non-uniqueness of the *HBB* intronic sequence.
(**B**) As (A), but for the indicated introns of *VASP* and (**C**) *SRSF3*. Data is from direct lariat sequencing. *p*-values were estimated with the multinomial proportion test. Plot restricted to branchpoints exhibiting differential branchpoint usage across the indicated samples, defined as a tissue-specific difference in branchpoint usage of ≥ 10% with an associated *p*-value ≤ 0.01 (two-sided test for difference in proportion). The illustrated percentages do not add up to 100% because the plot is restricted to differentially used branchpoints.
(**D**) Detection of tissue-specific branchpoint usage in *VASP* and *SRSF3* relative to the empirical false-discovery rate (FDR) for each intron. Empirical FDRs were estimated by identifying differential branchpoint usage between technical replicates. *p*-values were estimated by comparing the frequency of differential branchpoint usage detected between tissues and between technical replicates (two-sided test for difference in proportion).
(**E**) Fraction of constitutive introns exhibiting tissue-specific branchpoint usage within the GTEx dataset. Left panel, introns binned by the total number of sequenced lariats across all 54 tissues sampled by the GTEx project. Right panel, introns binned by the mean number of sequenced lariats per tissue. Error bars, 95% confidence intervals estimated with a proportion test.

Figure S1

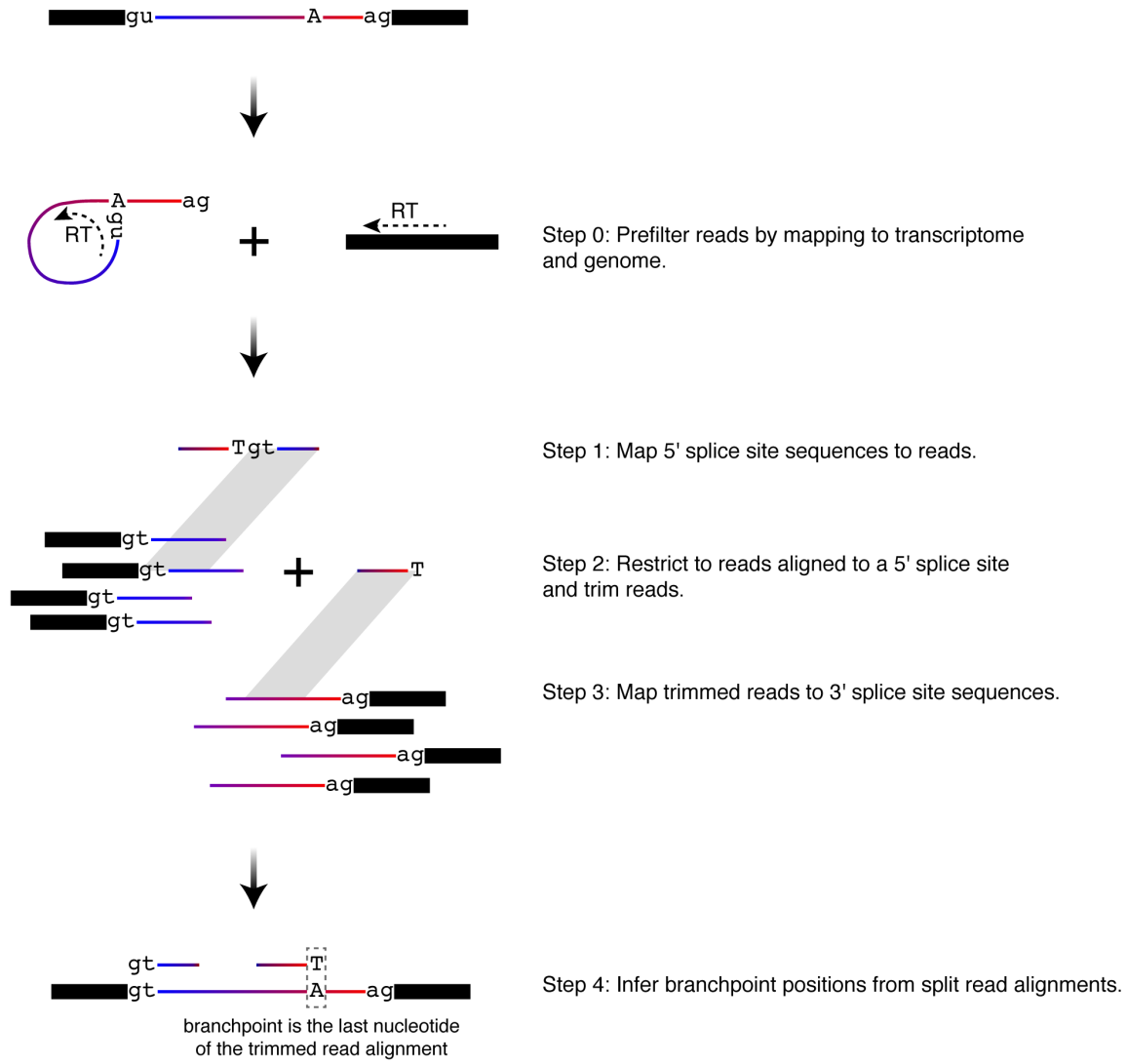


Figure S1. Strategy for branchpoint discovery and classification from RNA-seq data.
Flow chart of our branchpoint detection algorithm, which is based on the split-read alignment strategy of Mercer et al, 2015. The details of each step and subsequent assignment of confidence levels to each 5' splice site - branchpoint pair are described in Methods.

Figure S2

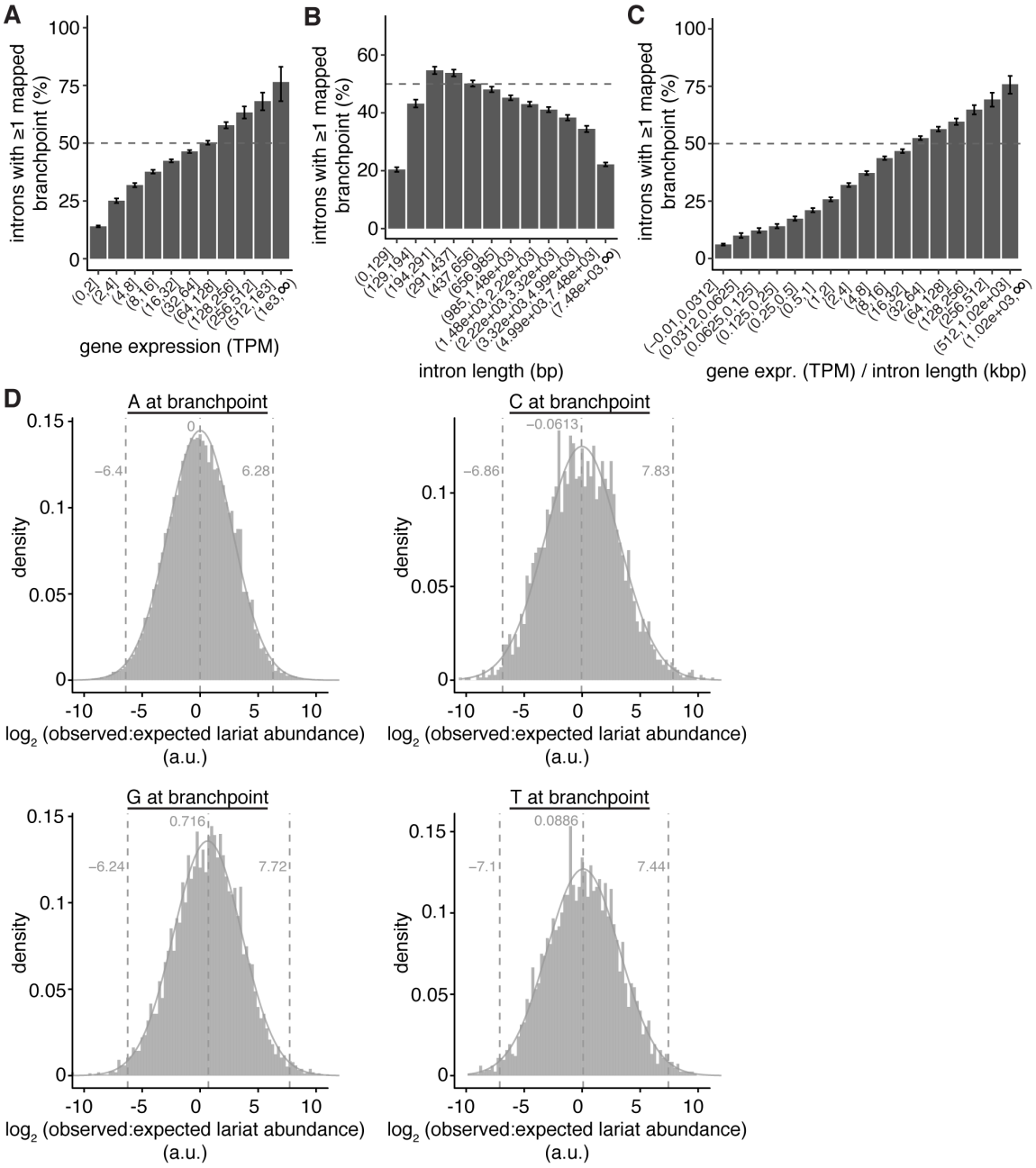


Figure S2. Branchpoint detection rate is influenced by intron length and parent gene expression.

(**A**) Fraction of constitutive introns with at least one mapped branchpoint as a function of parent gene expression. Gene expression was computed as the median expression across the 16 human tissues analyzed in the Illumina Body Map 2.0 dataset. Plot restricted to introns of length >= 200 bp to avoid the confounding effects of short introns. Dashed line indicates where 50% of introns have at least one annotated branchpoint.

(**B**) As (A), but for intron length. Plot restricted to introns within genes with median gene expression of >= 1 transcript per million (TPM).

(**C**) As (A), but for parent gene expression divided by intron length. Plot restricted to introns of length >= 200 bp to avoid the confounding effects of short introns.
(**D**) Histograms of observed to expected lariat abundance for constitutive introns. Lariats were binned according to the nucleotide at the branchpoint. Expected lariat abundance was defined for each intron as parent gene expression / intron length plus a regularization factor of 0.5 to ensure that the log transformation was defined. Histograms were centered by dividing by the median of the distribution for lariats with A at the branchpoint prior to performing a log transformation. The same scaling factor for centering was used in order to permit direct comparison of the different distributions. Dashed lines illustrate the 1st, 50th, and 99th percentiles. a.u., arbitrary units.

Figure S3

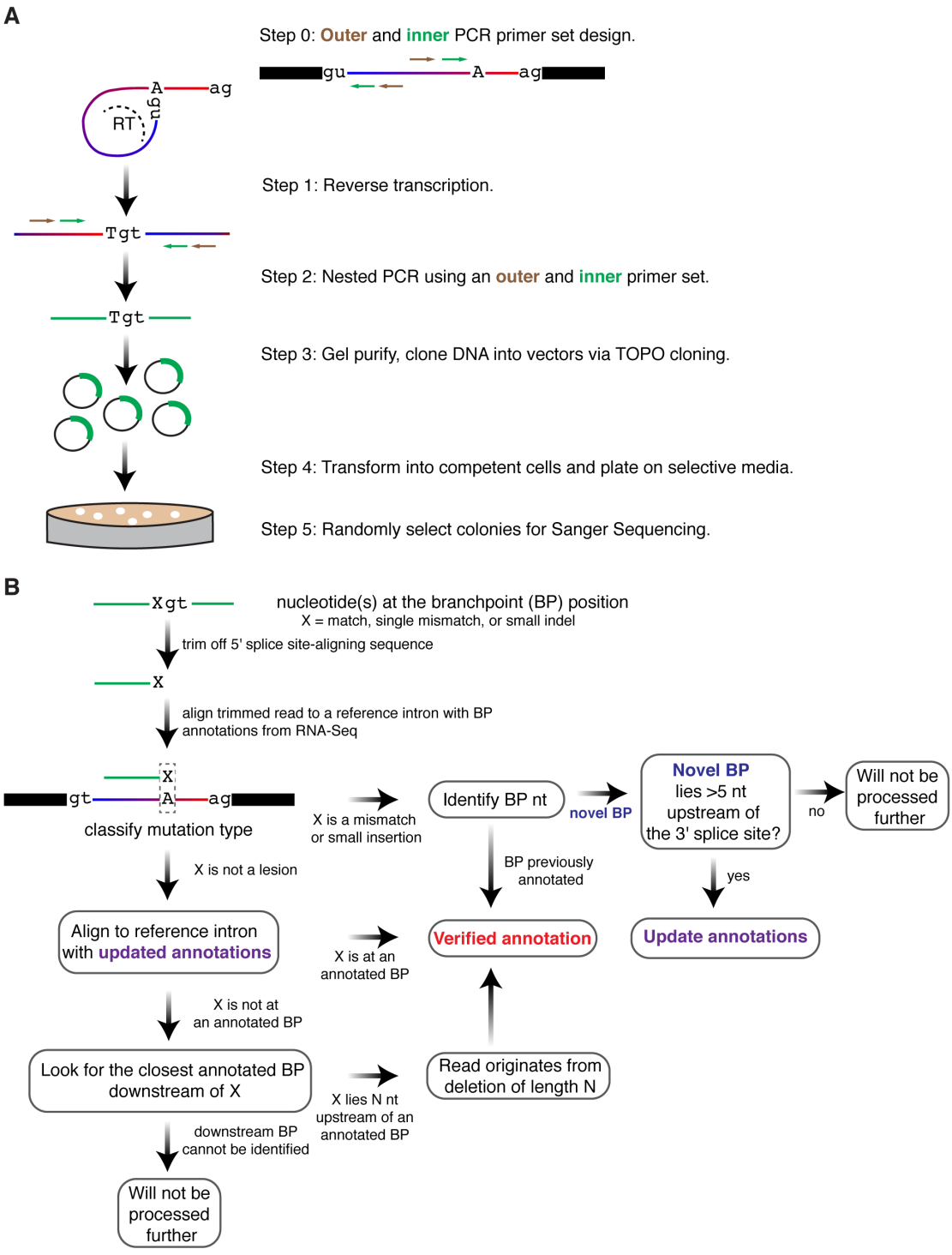


Figure S3. Strategy for branchpoint discovery from direct lariat sequencing.
(**A**) Experimental procedure for direct lariat sequencing.
(**B**) Flow chart of our procedure for annotating or validating branchpoints with data from direct lariat sequencing.

Figure S4

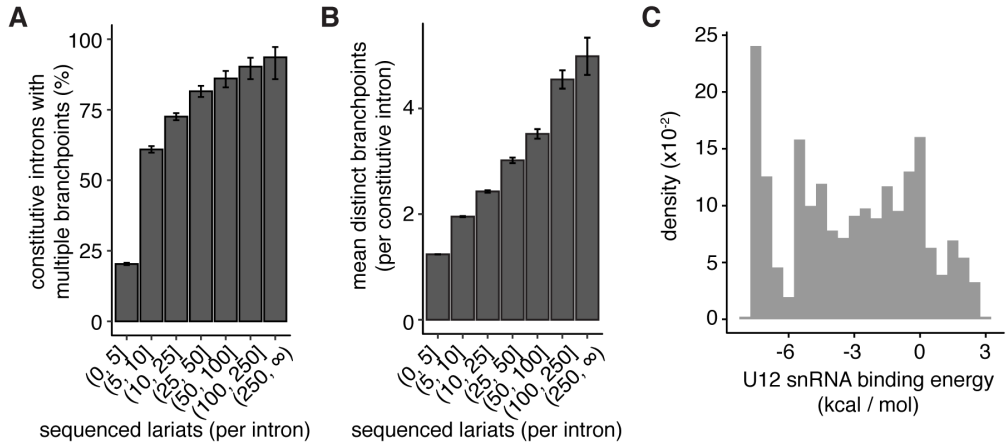


Figure S4. Branchpoint multiplicity and snRNA complementarity.
(**A**) As Figure 3C, but with all adjacent branchpoints merged into a single branchpoint.
(**B**) As Figure 3D, but with all adjacent branchpoints merged into a single branchpoint.
(**C**) Histogram of estimated U12 snRNA binding energies for all branchpoints within U12-type introns.

2.6 TABLES

study	RNA-seq reads analyzed	branchpoints	5' splice site - branchpoint pairs	A frequency at branchpoint
Gao et al, 2008	NA	60	60	85%
Taggart et al, 2012	~1.2 billion	862	not reported	39%
Mercer et al, 2015	~3 billion	59,359	not reported	78%
Taggart et al, 2017	~11.3 billion	36,078	not reported	55%
this study: high confidence	~1.31 trillion	130,294	136,998	77%
this study: moderate confidence	~1.31 trillion	8,220	9,182	50%
this study: low confidence	~1.31 trillion	47,894	48,935	32%

Table 1. Comparison of published branchpoint annotations. The high, moderate, and low confidence categories used in our study are mutually exclusive. The numbers for Taggart et al, 2017 correspond to their "high-confidence" set of branchpoints.

Dataset name	Reads (billion)	Accession number(s)	PMID(s)
2010/illumina.body_map_2	2.4	ERP000546 (ENA)	NA
2012/akimitsu.rna_stability	2.3	DRA000345, DRA000346, DRA000347, DRA000348, DRA000350, DRP000622 (DDBJ)	22369889, 23064110
2012/encode-gingeras.shortRNA_subcellular_fractions	13	GSE24565 (GEO)	22955620
2012/encode-gingeras.subcellular_fractions	5.1	GSE30567 (GEO)	22955620
2012/lopez-otin-quesada.chronic_lymphocytic_leukemia	2.4	EGAD00001000083 (EGA)	22158541, 23187290
2013/abdel-wahab.acute_myeloid_leukemia	1.4	NA	25965569
2013/aerts-cools.acute_lymphoblastic_leukemia	2.3	NA	24367274
2013/bradley-ramakrishnan.chronic_myelomonocytic_leukemia	1.1	NA	25965569
2013/bradley-ramakrishnan.spliceosomal_mutations	3.3	GSE58871, GSE65349 (GEO)	25267526, 25965569
2013/bradley-tapscott.fshd	1.4	GSE45883 (GEO)	24278031
2013/bradley.nonsense_mediated_decay	1.1	GSE58335, GSE61410 (GEO)	25385641
2013/carroll.B-lymphoblastic_leukemia	1.4	SRP009840 (SRA)	23377183
2013/ecker-ren.ucsd_epigenome_mapping	2.5	GSE16256 (GEO)	21289626
2013/geuvadis-dermitzakis.1000_genomes	11	E-GEUV-1 (ENA)	24037378
2013/gray.breast_cancer_celllines	4.2	GSE48213 (GEO)	24176112
2013/khaitovich-gelfand-chen.aging_brain	1.1	SRP005169 (SRA)	23340839
2013/paddison.glioblastoma	1.8	GSE75147 (GEO)	23651857
2013/sauvageau.acute_lymphoblastic_leukemia	2.3	GSE49601 (GEO)	24069164
2013/tapscott.fshd_patients	3.6	GSE56787 (GEO)	24861551
2013/TCGA.ACC	3.2	NCI Genomic Data Commons	
2013/TCGA.BLCA	1.4	NCI Genomic Data Commons	
2013/TCGA.BRCA	78	NCI Genomic Data Commons	
2013/TCGA.CESC	18	NCI Genomic Data Commons	
2013/TCGA.CHOL	2	NCI Genomic Data Commons	

2013/TCGA.COAD	4.1	NCI Genomic Data Commons	
2013/TCGA.DLBC	4.4	NCI Genomic Data Commons	
2013/TCGA.GBM	38	NCI Genomic Data Commons	
2013/TCGA.HNSC	4.1	NCI Genomic Data Commons	
2013/TCGA.KICH	2.4	NCI Genomic Data Commons	
2013/TCGA.KIRC	14	NCI Genomic Data Commons	
2013/TCGA.KIRP	2.8	NCI Genomic Data Commons	
2013/TCGA.LAML	43	NCI Genomic Data Commons	
2013/TCGA.LIHC	4.6	NCI Genomic Data Commons	
2013/TCGA.LUAD	27	NCI Genomic Data Commons	
2013/TCGA.LUSC	6.5	NCI Genomic Data Commons	
2013/TCGA.HNSC	29	NCI Genomic Data Commons	
2013/TCGA.OV	86	NCI Genomic Data Commons	
2013/TCGA.PAAD	11	NCI Genomic Data Commons	
2013/TCGA.PRAD	30	NCI Genomic Data Commons	
2013/TCGA.SARC	14	NCI Genomic Data Commons	
2013/TCGA.STAD	20	NCI Genomic Data Commons	
2013/TCGA.TGCT	6.9	NCI Genomic Data Commons	
2013/TCGA.THCA	7.4	NCI Genomic Data Commons	
2013/TCGA.THYM	9.7	NCI Genomic Data Commons	
2013/TCGA.UCEC	22	NCI Genomic Data Commons	
2013/west.breast_cancer_stroma	3.8	GSE42948 (GEO)	24342436
2013/white-mcnerney.acute_myeloid_leukemia	1.9	SRP017262 (SRA)	23212519
2014/conboy.erythropoiesis	1.1	GSE53635 (GEO)	24442673

2014/gallagher.erythropoiesis	1.1	GSE53983 (GEO)	24637361
2014/kim.colorectal_cancer	4.1	GSE50760 (GEO)	25049118
2014/myers.breast_cancer	27	GSE581350 (GEO)	24929677
2014/su2c.prostate_cancer	25	phs000915.v1.p1 (dbGaP)	26000489
2014/target-meshinchi.acute_myeloid_leukemia	43	phs000218.v18.p7 (dbGaP)	NA
2014/varmus.u2af1_mutations	4.8	GSE80136 (GEO)	27776121
2015/boultwood.sf3b1_mutations	2.3	GSE63569 (GEO)	25428262
2015/buonamici.sf3b1_mutations	2.2	GSE72790 (GEO)	26565915
2015/koeffler.zrsr2_mutations	1.2	GSE63816 (GEO)	25586593
2015/mattick.branchpoint_discovery	2.2	GSE53328 (GEO)	25561518
2016/berglund-wang.myotonic_dystrophy	3.7	GSE86356 (GEO)	27681373
2016/cairns.embryogenesis	9.6	GSE85632 (GEO)	28459457
2016/chen-chen.acute_lymphoblastic_leukemia	12	Chinese Genotype-phenotype Archive	27428428
2016/fioretos.acute_lymphoblastic_leukemia	5.1	EGAD00001002112 (EGA)	27265895
2016/gtex.tissues	550	phs000424.v6.p1 (dbGaP)	25954002
2016/mano.acute_lymphoblastic_leukemia	13	JGAS00000000047 (JGA)	27019113
2016/stern.uveal_melanoma	3.9	NA	26842708

Table S1. RNA-seq datasets analyzed in this study. Published or publicly available datasets with >1 billion available reads that were analyzed in this study. Accession numbers and/or PMIDs are listed when available.

Gene	Intron	PCR	Direction	Sequence (5'->3')
SRSF3	3	outer	F	TCAAAATCTTGCCCCTTTTG
SRSF3	3	outer	R	CAAAAGCCAACACTCAGCAC
SRSF3	3	inner	F	ACAGCACACTGTTGCCCATC
SRSF3	3	inner	R	CATACCCCAAATTACACCCAAC
POLR3A	4	outer	F	TATTGGGAAACGGACCTCTC
POLR3A	4	outer	R	GAGAAAAGCTGACTCCCGAAC
POLR3A	4	inner	F	TATTGGGAAACGGACCTCTC
POLR3A	4	inner	R	AACAGAAGACAGTGAGTGAAAAGG
MBNL1	7	outer	F	AAGCCTGTTTGTGTCAATTTTC
MBNL1	7	outer	R	GGAAATGGACTTGCCCAATAG
MBNL1	7	inner	F	TGTCAATTTTCTTGATTTGATGG
MBNL1	7	inner	R	ATTTTGAGGGGCTGTGAGG
SNX9	10	outer	F	AGGGAGATAGAGTGGGAGCTG
SNX9	10	outer	R	CTTCTGGCAGGCAGTTCTTC
SNX9	10	inner	F	AATGAGCTGGTCCTGTTTGG
SNX9	10	inner	R	AGGTTTCTGTCCCCTCACTG
VASP	9	outer	F	GCACCCTTATAGGAGAGTCAGG
VASP	9	outer	R	TAGTTCCTGTGGCTGGACTG
VASP	9	inner	F	GCACCCTTATAGGAGAGTCAGG
VASP	9	inner	R	GGCTGGACTGGGCACTCTAC

Table S2. Primers for lariat sequencing. PCR primers used for direct lariat sequencing.

2.7 MATERIALS AND METHODS

2.7.1 *Method Details*

2.7.1.1 Genome annotations

We generated a genome annotation by merging the UCSC knownGene (Meyer et al., 2013), Ensembl 71 (Flicek et al., 2013), and MISO version 2.0 (Katz et al., 2010) annotations for the UCSC hg19 (GRCh37) genome assembly. We created an expanded intron annotation for subsequent branchpoint mapping by enumerating all possible combinations of annotated 5′ and 3′ splice sites within each gene.

2.7.1.2 Gene expression and alternative splicing analysis

We estimated gene expression and alternative splicing across the 16 tissues in the Body Map 2.0 database as described previously (Dvinge et al., 2014). Briefly, we first mapped all reads to the transcriptome with RSEM (RNA-seq by expectation maximization) version 1.2.4 (B. Li & Dewey, 2011), which produces gene-level expression estimates. We modified RSEM to invoke Bowtie (Langmead et al., 2009) with the option "-v 2." We then mapped remaining unaligned reads to the genome and the splice junction database described above (equivalent to the expanded intron annotation) with TopHat version 2.0.8b (Trapnell et al., 2009). We merged the read alignments produced by RSEM and TopHat and used those as input to MISO with its version 2.0 annotation (Katz et al., 2010) to quantify isoform expression.

2.7.1.3 Branchpoint detection algorithm

Our branchpoint detection algorithm was based on the split-read alignment strategy used in Mercer et al. (2015).

Prefilter reads

First, filter out reads with > 5% Ns or other ambiguous characters. Next, sequentially invoke Bowtie2 as follows for the transcriptome and genome: bowtie2 -x <index file for transcriptome or genome> --end-to-end --sensitive --score-min L,0,-0.24 -k 1 --n-ceil L,0,0.05 -U <FASTQ file of reads>. Finally, discard the aligned reads and use the unaligned reads as input for the next step.

Map 5′ splice site sequences to reads

First, build a Bowtie index for a FASTA file of prefiltered reads. Next, build a FASTA file holding 5′ splice site sequences (the first 20 nt of each intron or, alternately, the 20 nt downstream from each 5′ splice site, including the 5′ splice site itself). Finally, map 5′ splice site sequences to reads as follows: bowtie2 -x <index file for reads> --end-to-end --sensitive --k 10000 --no-unal -f -U <FASTA file of 5′ splice site sequences>.

Restrict to reads aligned to a 5′ splice site and trim reads

First, restrict to alignments between 5′ splice site sequences and reads with no mismatches and no indels. Second, restrict to reads that align to a single 5′ splice site. Third, trim off the portion of each read starting at the 5′ splice site alignment and continuing to the end of the read. Finally, restrict to trimmed reads of ≥ 20-nt length.

Map trimmed reads to 3′ splice site sequences

First, build a FASTA file holding the 3′ splice site sequences (the last 250 nt of each intron or, alternately, the 250 nt upstream of each 3′ splice site, including the 3′ splice site itself). Next, build a Bowtie index for these sequences. Finally, map trimmed reads to 3′ splice sites as follows: bowtie2 -x <index file for 3′ splice sites> --end-to-end --sensitive -k 10 --no-unal -f -U <FASTA file of trimmed reads>.

Infer branchpoint positions from split-read alignments

First, restrict to trimmed read alignments with five or fewer mismatches, ≤10% mismatch rate, and at most a single indel of ≤3-nt length in the 3′ splice site-aligning portion of the read. Second, restrict to alignments that score as well as the best-scoring alignment for each read (e.g., remove lower-scoring alignments). Third, restrict to reads with inverted alignments (e.g., where the "left" half of the read aligns near the 3′ splice site, while the "right" half of the read aligns to the 5′ splice site). Fourth, restrict to reads for which the 5′ and 3′ splice site-aligning portions of the read map to splice sites within a single gene. Fifth, compute the branchpoint position as the last nucleotide of the trimmed read alignment. Sixth, restrict to reads with a mismatch at the inferred branchpoint position. Finally, assemble a final set of 5′ splice site–branchpoint pairs.

Assign confidence levels to each 5′ splice site–branchpoint pair

First, for each identified 5′ splice site–branchpoint pair, extract these sequences and identify nonunique sequences as follows: (1) 5′ splice site sequence (25 nt of sequence downstream from the 5′ splice site, including the 5′ splice site itself; test whether each sequence aligns to more than one location in the genome with no mismatches or gaps), (2) upstream branchpoint sequence (25 nt of sequence upstream of the branchpoint, including the branchpoint itself; test whether each sequence aligns to more than one location in the genome with no mismatches or gaps), and (3) lariat sequence (concatenation of the branchpoint and 5′ splice site sequence; test whether each sequence aligns to the transcriptome or genome with two or fewer mismatches, ≤ 5% mismatches, and no gaps). Next, assign confidence levels to each 5′ splice site–branchpoint pair based on the "hits" (branchpoint-spanning reads used to infer the branchpoint location) as follows: (1) low (one or more hits with mismatch at the branchpoint and ≥ 5% of hits with mismatches at the branchpoint), (2) moderate (one or more hits with mismatch at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read and ≥ 5% of hits with mismatches at

the branchpoint and no other mismatches or indels in the 3′ splice site region of the read), and (3) high (one or more hits with mismatch at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read, ≥ 5% of hits with mismatches at the branchpoint and no other mismatches or indels in the 3′ splice site region of the read, and unique 5′ splice site, upstream branchpoint, and lariat sequences).

2.7.1.4 Branchpoint sequence analysis

Branchpoint sequence analysis was performed within the R programming environment. All analyses relied on Bioconductor tools, including the AnnotationHub, BSgenome, GenomicAlignments, GenomicFeatures, and GenomicRanges packages (Lawrence et al., 2013; Huber et al., 2015). All plots and figures were generated with the dplyr (http://CRAN.R-project.org/package=dplyr) and ggplot2 (Wickham, 2009) packages.

Introns were classified as U2 or U12 type by computing the best match between the 5′ splice site of the intron and position weight matrices (PWMs) representing consensus U2- or U12-type 5′ splice sites (Sheth et al., 2006). The binding energy between a given branchpoint context (excluding the branchpoint itself, which appears as a bulge) and the U2 snRNA sequence AUGAUGUG was computed with ViennaRNA (Lorenz et al., 2011).

Several statistics were recomputed in order to create Table 1. The A frequency at branchpoints reported by Gao et al. (2008) was recomputed in order to be consistent with other studies. Gao et al. (2008) reported this statistic using all sequenced lariats rather than distinct called branchpoints. The A frequency was similarly recomputed for Taggart et al. (2012, 2017). For Taggart et al. (2017), the “high-confidence” set of branchpoints was used. This set of branchpoints was obtained by collapsing branchpoint calls (branchpoints in Supplemental Table S2 of Taggert

et al. 2017 without motif model values of “template_switching” or “circle”) (AJ Taggart and WG Fairbrother, pers. comm.).

2.7.1.5 Quantification of branchpoint usage across tissues

Branchpoint usage was computed across the 54 tissues represented in the GTEx data set (Melé et al. 2015) as follows. All reads were prefiltered by aligning them to the transcriptome and genome and then discarding aligned reads (as for the branchpoint discovery algorithm). Unaligned reads were collated across individuals for each tissue and aligned to a database of lariat sequences. The lariat sequence database consisted of sequences spanning the branchpoint itself, with the length of flanking sequence upstream of and downstream from the branchpoint chosen such that an aligned read must have at least 10 nt aligned to either side of the branchpoint. The database of lariat sequences is therefore dependent on the query read length. Unaligned reads were mapped to the lariat database with Bowtie2 (Langmead & Salzberg, 2012). Alignments were restricted to those with three or fewer mismatches and one or fewer indel of ≤ 3-nt length. Reads were permitted to align to up to 25 different lariat sequences; however, multimapping reads were downweighted proportional to the number of lariats to which they aligned. Usage of a given branchpoint was then estimated as the number of reads supporting usage of that branchpoint divided by the total number of reads supporting usage of the 5′ splice site that was associated with that branchpoint. Usage of any branchpoint therefore must fall within the interval 0% – 100%. This is analogous to the Ψ value commonly used for estimating usage of alternatively spliced sequence (E. T. Wang et al., 2008) For each branchpoint, a minimum of 20 reads per tissue was required in order to estimate branchpoint usage; if < 20 reads were available, then that data point was not subjected to further analysis.

An intron was said to exhibit tissue-specific branchpoint usage if we observed tissue-specific differences in branchpoint usage of ≥ 10% (absolute, rather than relative, value) with a P-value of ≤ 0.05 by a two-sided proportion test.

2.7.1.6 RNA extraction and cDNA synthesis

K562 cells were lysed using TRIzol (Thermo Fisher Scientific). K562 total RNA was isolated from the cell lysate according to the manufacturer's protocol and cleaned using the Qiagen RNeasy minikit with DNase treatment (Qiagen, RNase-free DNase set). Total RNA from human peripheral blood mononuclear cells, cerebella, testes, spleens, and fetal spleens were purchased from Takara Bio. cDNA was synthesized from the total RNA using random hexamer priming and the SuperScript III first strand synthesis system (Thermo Fisher Scientific).

2.7.1.7 Direct lariat sequencing and analysis

Primers (Integrated DNA Technologies) were designed to amplify the branchpoint–5′ splice site junction of a specific lariat via nested PCR as illustrated in **Supplemental Figure S3A**. A first round of gradient PCR (30 cycles) was performed with Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific) using the “outer” primer set and the K562 cDNA as a template. The annealing temperatures used were in the range of Tm ± 3°C. The reactions were pooled together, cleaned, and concentrated using the QIAquick PCR purification kit (Qiagen). The concentrated DNA served as the template for the second round of gradient PCR (30 cycles) using the “inner” primer set and an annealing temperature range analogous to that used for the first round of PCR. The reactions were combined and subjected to 2% agarose gel electrophoresis. Bands of sizes consistent with lariat amplification were excised, and DNA was extracted using the MinElute gel extraction kit (Qiagen). Purified DNA fragments were cloned into the pCR-Blunt II-TOPO vector (Thermo Fisher Scientific) and transformed into TOP10 chemically competent Escherichia coli

(Thermo Fisher Scientific) using the ZeroBlunt TOPO PCR cloning kit (Thermo Fisher Scientific). The transformants were plated on 50 μg/mL LB + kanamycin plates, and random colonies were selected for Sanger sequencing (Genewiz) after growth. Inner and outer primer sets are listed in **Supplemental Table S2**.

Branchpoints were annotated using Sanger-sequenced amplicons with the algorithm outlined in **Supplemental Figure S3B**. Only reads that contained a mismatch or small insertion at the inferred branchpoint position were used to identify novel branchpoints. Reads with no lesion or those containing a deletion at the inferred branchpoint were used only to experimentally confirm branchpoints annotated via RNA-seq analysis, not to annotate new branchpoints. Tissue-specific branchpoint usage was quantified and analyzed analogously to the method described above for the GTEx data set.

2.8 ACKNOWLEDGMENTS

J.M.B.P. was supported by the ARCS Foundation. R.K.B. is a Scholar of The Leukemia and Lymphoma Society (1344-18) and was supported by the Edward P. Evans Foundation, National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK103854), and NIH/National Heart, Lung, and Blood Institute (NHLBI) (R01 HL128239). The results published here are based in part on data generated by The Cancer Genome Atlas Research Network (http://cancergenome.nih.gov). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the NIH and by the National Cancer Institute, National Human Genome Research Institute, NHLBI, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke. The GTEx data used for the analyses described in this manuscript were obtained from the

Database of Genotypes and Phenotypes (dbGaP; accession no. phs000424.v6.p1) on February 8, 2017.

2.9 Author Contributions

J.M.B.P. and R.K.B. performed the experiments, analyzed the data, and wrote the paper.

2.1 Competing Interests

The authors declare that they have no competing interests.

2.2 References

Alsafadi, S., Houy, A., Battistella, A., Popova, T., Wassef, M., Henry, E., Tirode, F., Constantinou, A., Piperno-Neumann, S., Roman-Roman, S., Dutertre, M., & Stern, M. H. (2016). Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms10615

Bradley, R. K., Merkin, J., Lambert, N. J., & Burge, C. B. (2012). Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biology*, *10*(1), e1001229. https://doi.org/10.1371/journal.pbio.1001229

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., & Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, *24*(11), 1774–1786. https://doi.org/10.1101/gr.177790.114

Corvelo, A., Hallegger, M., Smith, C. W. J., & Eyras, E. (2010). Genome-wide association between branch point properties and alternative splicing. *PLoS Computational Biology*, *6*(11), 12–15. https://doi.org/10.1371/journal.pcbi.1001016

Darman, R. B., Seiler, M., Agrawal, A. A., Lim, K. H., Peng, S., Aird, D., Bailey, S. L., Bhavsar, E. B., Chan, B., Colla, S., Corson, L., Feala, J., Fekkes, P., Ichikawa, K., Keaney, G. F., Lee, L., Kumar, P., Kunii, K., MacKenzie, C., … Buonamici, S. (2015). Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Reports*, *13*(5), 1033–1045. https://doi.org/10.1016/j.celrep.2015.09.053

DeBoever, C., Ghia, E. M., Shepard, P. J., Rassenti, L., Barrett, C. L., Jepsen, K., Jamieson, C. H. M., Carson, D., Kipps, T. J., & Frazer, K. A. (2015). Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3' Splice Site Selection in SF3B1-mutated Cancers. *PLoS Computational Biology*, *11*(3), 1–19. https://doi.org/10.1371/journal.pcbi.1004105

Dembowski, J. A., & Grabowski, P. J. (2009). The CUGBP2 splicing factor regulates an ensemble of branchpoints from perimeter binding sites with implications for autoregulation. *PLoS Genetics*, *5*(8), e1000595. https://doi.org/10.1371/journal.pgen.1000595

Dietrich, R. C., Peris, M. J., Seyboldt, A. S., & Padgett, R. A. (2001). Role of the 3' splice site in U12-dependent intron splicing. *Molecular and Cellular Biology*, *21*(6), 1942–1952. https://doi.org/10.1128/MCB.21.6.1942-1952.2001

Dvinge, H., & Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, *7*(1), 45. https://doi.org/10.1186/s13073-015-0168-9

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews. Cancer*, *16*(7), 413–430. https://doi.org/10.1038/nrc.2016.51

Dvinge, H., Ries, R. E., Ilagan, J. O., Stirewalt, D. L., Meshinchi, S., & Bradley, R. K. (2014). Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(47), 16802–16807. https://doi.org/10.1073/pnas.1413374111

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., … Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Research*, *41*(D1), 48–55. https://doi.org/10.1093/nar/gks1236

Fu, X.-D., & Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews. Genetics*, *15*(10), 689–701. https://doi.org/10.1038/nrg3778

Gao, K., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Research*, *36*(7), 2257–2267. https://doi.org/10.1093/nar/gkn073

Gattoni, R., Schmitt, P., & Stevenin, J. (1988). In vitro splicing of adenovirus E1A transcripts: characterization of novel reactions and of multiple branch points abnormally far from the 3' splice site. *Nucleic Acids Research*, *16*(6), 2389–2409. https://doi.org/10.1093/nar/16.6.2389

Hallegger, M., Sobala, A., & Smith, C. W. J. (2010). Four exons of the serotonin receptor 4 gene are associated with multiple distant branch points. *RNA (New York, N.Y.)*, *16*(4), 839–851. https://doi.org/10.1261/rna.2013110

Hartmuth, K., & Barta, A. (1988). Unusual branch point selection in processing of human growth hormone pre-mRNA. *Molecular and Cellular Biology*, *8*(5), 2011–2020. https://doi.org/10.1128/mcb.8.5.2011-2020.1988

Helfman, D. M., & Ricci, W. M. (1989). Branch point selection in alternative splicing of tropomyosin pre-mRNAs. *Nucleic Acids Research*, *17*(14), 5633–5650. https://doi.org/10.1093/nar/17.14.5633

Hornig, H., Aebi, M., & Weissmann, C. (1986). Effect of mutations at the lariat branch acceptor site on beta-globin pre-mRNA splicing in vitro. *Nature*, *324*(6097), 589–591. https://doi.org/10.1038/324589a0

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

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

Kol, G., Lev-Maor, G., & Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Human Molecular Genetics*, *14*(11), 1559–1568. https://doi.org/10.1093/hmg/ddi164

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

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., & Brenner, S. E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, *446*(7138), 926–929. https://doi.org/10.1038/nature05676

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

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

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 122–128. https://doi.org/10.1186/1748-7188-6-26

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. v, Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., … Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, *348*(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. *Genome Research*, *25*(2), 290–303. https://doi.org/10.1101/gr.182899.114

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., … Kent, W. J. (2013). The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Research*, *41*(D1), 64–69. https://doi.org/10.1093/nar/gks1048

Mullen, M. P., Smith, C. W., Patton, J. G., & Nadal-Ginard, B. (1991). Alpha-tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice. *Genes & Development*, *5*(4), 642–655. https://doi.org/10.1101/gad.5.4.642

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E., & Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & Development*, *21*(6), 708–718. https://doi.org/10.1101/gad.1525507

Noble, J. C., Pan, Z. Q., Prives, C., & Manley, J. L. (1987). Splicing of SV40 early pre-mRNA to large T and small t mRNAs utilizes different patterns of lariat branch sites. *Cell*, *50*(2), 227–236. https://doi.org/10.1016/0092-8674(87)90218-2

Noble, J. C., Prives, C., & Manley, J. L. (1988). Alternative splicing of SV40 early pre-mRNA is determined by branch site selection. *Genes & Development*, *2*(11), 1460–1475. https://doi.org/10.1101/gad.2.11.1460

Obeng, E. A., Chappell, R. J., Seiler, M., Chen, M. C., Campagna, D. R., Schmidt, P. J., Schneider, R. K., Lord, A. M., Wang, L., Gambe, R. G., McConkey, M. E., Ali, A. M., Raza, A., Yu, L., Buonamici, S., Smith, P. G., Mullally, A., Wu, C. J., Fleming, M. D., & Ebert, B. L. (2016). Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. *Cancer Cell*, *30*(3), 404–417. https://doi.org/10.1016/j.ccell.2016.08.006

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-45. https://doi.org/10.1093/nar/gkv1189

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., Godfrey, A. L., Rapado, I., Cvejic, A., Rance, R., McGee, C., Ellis, P., Mudie, L. J., Stephens, P. J., McLaren, S., … Campbell, P. J. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *New England Journal of Medicine*, *365*(15), 1384–1395. https://doi.org/10.1056/NEJMoa1103283

Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Beà, S., Pinyol, M., Martínez-Trillos, A., López-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M., Rozman, M., Delgado, J., Giné, E., Hernández, J. M., … López-Otín, C. (2011). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics*, *44*(1), 47–52. https://doi.org/10.1038/ng.1032

Reed, R., & Maniatis, T. (1988). The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes & Development*, *2*(10), 1268–1276. https://doi.org/10.1101/gad.2.10.1268

Robinson, D., van Allen, E. M., Wu, Y.-M., Schultz, N., Lonigro, R. J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C. C., Attard, G., Beltran, H., Abida, W., Bradley, R. K., Vinson, J., Cao, X., Vats, P., Kunju, L. P., Hussain, M., Feng, F. Y., …

Chinnaiyan, A. M. (2015). Integrative clinical genomics of advanced prostate cancer. *Cell*, *161*(5), 1215–1228. https://doi.org/10.1016/j.cell.2015.05.001

Ruskin, B., Greene, J. M., & Green, M. R. (1985). Cryptic branch point activation allows accurate in vitro splicing of human beta-globin intron mutants. *Cell*, *41*(3), 833–844. https://doi.org/10.1016/s0092-8674(85)80064-7

Ruskin, B., Krainer, A. R., Maniatis, T., & Green, M. R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell*, *38*(1), 317–331. https://doi.org/10.1016/0092-8674(84)90553-1

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews. Genetics*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research*, *34*(14), 3955–3967. https://doi.org/10.1093/nar/gkl556

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Smith, C. W., Chu, T. T., & Nadal-Ginard, B. (1993). Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular and Cellular Biology*, *13*(8), 4939–4952. https://doi.org/10.1128/mcb.13.8.4939-4952.1993

Southby, J., Gooding, C., & Smith, C. W. (1999). Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons. *Molecular and Cellular Biology*, *19*(4), 2699–2711. https://doi.org/10.1128/MCB.19.4.2699

Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E., & Fairbrother, W. G. (2012). Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nature Structural & Molecular Biology*, *19*(7), 719–721. https://doi.org/10.1038/nsmb.2327

Taggart, A. J., Lin, C. L., Shrestha, B., Heintzelman, C., Kim, S., & Fairbrother, W. G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. *Genome Research*, *27*(4), 639–649. https://doi.org/10.1101/gr.202820.115

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Vogel, J., Hess, W. R., & Börner, T. (1997). Precise branch point mapping and quantification of splicing intermediates. *Nucleic Acids Research*, *25*(10), 2030–2031. https://doi.org/10.1093/nar/25.10.2030

Wahl, M. C., Will, C. L., & Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell*, *136*(4), 701–718. https://doi.org/10.1016/j.cell.2009.02.009

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., Zhang, L., Zhang, W., Vartanov, A. R., Fernandes, S. M., Goldstein, N. R., Folco, E. G., Cibulskis, K., Tesar, B., Sievers, Q. L., Shefler, E., … Wu, C. J. (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England Journal of Medicine*, *365*(26), 2497–2506. https://doi.org/10.1056/NEJMoa1109016

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., Chalkidis, G., Suzuki, Y., Shiosaka, M., Kawahata, R., Yamaguchi, T., Otsu, M., Obara, N., Sakata-Yanagimoto, M., Ishiyama, K., … Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, *478*(7367), 64–69. https://doi.org/10.1038/nature10496

Zhuang, Y. A., Goldstein, A. M., & Weiner, A. M. (1989). UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(8), 2752–2756. https://doi.org/10.1073/pnas.86.8.2752

Chapter 3. MOST HUMAN GENES EXPRESS INTRON-DERIVED CIRCULAR RNAS

Jose Mario Bello Pineda*, Taylor R. Nicholas*, and Robert K. Bradley

*These authors contributed equally

A version of this chapter has been submitted for publication on November 6, 2022.

3.1 ABSTRACT

Circular RNAs are closed-loop molecules that can arise from diverse processes, with non-co-linear splicing or "backsplicing" of exons producing the majority of known circular RNAs in human cells. Here, we report that many human genes also express intron-derived circular RNAs. These "cintrons" correspond to full-length introns that have been excised from pre-mRNA and circularized, with the frequent addition of non-templated adenosines at the point of circularization. Sequence analyses and biophysical modeling suggest that cintrons arise from debranched intron lariats, with cintron formation strongly associated with the expected proximity of the 5' and 3' intron ends in the lariat conformation due to RNA secondary structure, RNA flexibility, and lariat tail length. Cintron expression is variable across tissues, with highest levels in the brain and testis. Our results describe a widespread class of circular RNAs and motivate future studies of their potential functional roles.

3.2 INTRODUCTION

Circular RNAs are conserved across all domains of life and can be composed of coding or non-coding sequences or both. Exonic circular RNAs (circRNAs) are produced by backsplicing, a non-canonical splicing event in which a 5′ splice site (5′ss) is joined to an upstream 3′ splice site (3′ss),

resulting in looping, covalent closure, and ultimate production of a transcript consisting of one or more exons. Exonic circular RNAs are common transcriptional products in human cells (reviewed in Kristensen et al., 2019). Self-splicing introns can also generate circular RNAs. This occurs for the intervening sequence of Tetrahymena ribosomal RNA (Been and Cech, 1985; Cech et al., 1981; Grabowski et al., 1981; Kruger et al., 1982; Zaug et al., 1983) and in the context of the cyclization pathways of group I (Dalgaard and Garrett, 1992; Nielsen et al., 2003; Nielsen and Johansen, 2009; Vader et al., 2002; Vader et al., 1999; Vicens and Cech, 2009) and group II introns (Li-Pook-Than and Bonen, 2006; Molina-Sanchez et al., 2006; Monat and Cousineau, 2016; Murray et al., 2001). Although self-splicing introns are absent from the human genome, intron-derived, closed-loop RNAs with 2′-5′ linkages at their points of closure have been described. Such RNAs arise from intronic lariats for which the lariat tail, the sequence from the branchpoint to the 3′ss, has been trimmed by exonuclease activity (Saini et al., 2019; Talhouarne and Gall, 2018; Zhang et al., 2013). Recently, a large-scale study of human branchpoints reported that ~3% of identified branchpoints coincide with the 3'ss, consistent with expression of circular RNAs derived from introns that are distinct from lariats (Taggart et al., 2017). These findings suggested that intron-derived circular RNAs analogous to those arising from self-splicing introns are more common in human cells than previously appreciated, although the extent, pervasiveness, and mechanistic origins of such circular RNAs are unknown.

3.3 RESULTS

3.3.1 *Detecting cintrons from sequencing data*

We undertook a focused search to identify and characterize intron-derived circular RNAs in human cells. To search for such circular RNAs, we took advantage of the “split-read” method for high-throughput branchpoint identification that leverages the unique sequence signatures of lariat-

identifying reads: the branchpoint (BP) and 5′ss sequences joined in an inverted orientation, with a mutation at the BP nucleotide position due to low-fidelity reverse transcription (RT) at the lariat 2′-5′ linkage (Mercer et al., 2015; Pineda and Bradley, 2018). A mismatch uniquely identifies the branchpoint position, while lack thereof is consistent with either correct nucleotide incorporation while traversing a 2′-5′ linkage or normal traversal of a 5′-3′ linkage (**Fig. 1A**).

In order to find intron-derived circular RNAs with 5′-3′ linkages at their points of closure, we modified this methodology by removing the mismatch requirement at the nucleotide immediately adjacent to the 5′ss sequence of the inverted read (the "junction nucleotide"), which indicates the point of circularization (**Fig. S1A**). Applying this detection algorithm to ~2 trillion RNA-seq reads from 24,813 datasets, we found a dramatic increase in read density at the -1 nt position when we removed versus kept the requirement for a mismatch at the junction nucleotide, consistent with enhanced detection of full-length circular introns without RT polymerization errors at the point of closure, consistent with 5′-3′ linkages. Some inverted reads mapped to positions in the BP region of the intron even when no mismatch was required, likely representing BPs that were not mutated during RT, which occurs ~25% of the time (Mercer et al., 2015) (**Fig. 1B**). We termed these intron-derived circular RNAs "cintrons" (**Fig. 1A**). We restricted our analyses to RefSeq constitutive (not alternatively spliced) introns to eliminate the possibility that the cintronic signals were artifacts of alternative splicing. This detection algorithm allows simultaneous identification of BP positions and cintron-generating introns, as in the introns at the 3′ end of *EWSR1* (**Fig. 1C**).

To experimentally validate cintrons, we performed RT-PCR with nested outward facing primer sets (**Fig. S1B**). The primer design allows amplification of both the 5′ss-BP lariat junction and the concomitant 5′ss-3′ss cintron junction. We used TOPO cloning on the amplicons and

subsequent Sanger sequencing to quantify cintron abundance. As an exemplar, we performed direct cintron and lariat sequencing on the intron upstream of the *SRSF3* ultraconserved poison exon. In addition to the branchpoint cluster previously annotated computationally and experimentally (Pineda and Bradley, 2018), we observed cintronic sequences notable in their lack of mutations at the junction nucleotide. Intriguingly, cintron sequences featured frequent addition of non-templated adenosines at the circularization junction (**Fig. 1D**).

To determine if the cintrons in *SRSF3* were in fact circular, we treated samples with RNase R to digest linear RNA and performed RT-PCR with outward facing primers (**Fig. 1E**). The control *FBXW4* linear RNA was degraded by RNase R while the control *FBXW4* circRNAs and both *SRSF3* cintrons and lariats resisted exonuclease digestion. Sanger sequencing of the circularization junctions from RNase R-digested RNA confirms lariat and cintron persistence and recovery from these conditions, validating cintrons as full-length circular introns (**Fig. 1F**).

3.3.2 *Cintron formation is prevalent genome-wide*

We sought to understand if cintron-formation was an attribute generalizable to other introns by performing direct lariat and cintron sequencing on 10 distinct intronic sequences. Cintron abundance varied between exceeding lariats, as seen in the introns of *DUSP1* and *BRD9*; and zero, as in the *GLUL*, *HBB*, and *VASP* introns (**Fig. 2A**, p-value $< 2.2\times10^{-16}$ via a 10-sample proportion test). These experiments suggest the existence of context-specific features that permit cintron formation and support the plurality of cintron-producing introns. Of note, the rate of mutation incorporation at the junction nucleotide diverges between cintrons and lariats in our experiments. Lariats harbor mutation rates matching RT error over the 2′-5′ phosphodiester measured from RNA-seq (Mercer et al., 2015), while cintrons display dramatically lower mutation rates congruent

with accurate transcription over their junctions (**Fig. 2B**). These results suggest that the cintron junction is covalently linked by a 5′-3′ bond, making it structurally distinct from lariats.

We next assessed the preponderance of cintron-producing genes among all coding genes. By stratifying each gene based on the number of sequenced lariats and cintrons, we obtained an estimate of the proportion of cintron-yielding genes as a function of sequencing depth. This power analysis suggested that the majority of human coding genes (~56%) generate cintrons (**Fig. 2C**). In similar power analyses, we estimated that ~23% of U2-type and ~52% of U12-type introns produce cintrons (**Fig. 2D**). We identified ~15,000 unique cintrons in our analysis, but a power analysis indicates that we were far from saturation (**Fig. 2E**).

3.3.3 *Lariat tail length is correlated with cintron formation*

We wanted to understand if cintron-forming (cintron+) introns have specific sequence features that distinguish them from their intronic counterparts that do not form cintrons (cintron–). To control for factors that might influence cintron detection rate, such as differences in gene expression, we restricted our analyses to introns with at least one detected BP. We observed that cintron+ introns have shorter polypyrimidine tracts than cintron– introns (**Fig. 3A**). We found that the BP sequence motifs and nucleotide preferences in U2- and U12-type cintron+ and cintron– introns are largely similar. However, we observed a statistically significant difference in the overall BP nucleotide preference between cintron+ and cintron– U2-type introns, arising from differences in usage of non-canonical BP nucleotides. Further, U12-type cintron+ introns have branchpoint sequences which better conform to the canonical CCUUNAN minor intron branchpoint sequence (Hall and Padgett, 1994) (**Fig. 3B-D**). U12-type cintron+ introns also have significantly lower predicted snRNA binding energies compared to cintron– introns (**Fig. 3E**). Taken together with the observed

conservation of the canonical branchpoint motif, U12-type cintron+ introns have branchpoint sequences which may permit better association with the minor spliceosome.

We next wanted to understand the relationship between BP position and cintron formation. We hypothesized that BP proximity to the 3′ss might influence cintron formation, since that distance influences the proximity of the RNA ends in the lariat structure. Since branchpoint multiplicity is a feature of most human introns (Pineda and Bradley, 2018), we restricted our analyses to the most 3′ss-proximal branchpoint within each intron. Compared to U2-type cintron– introns, U2-type cintron+ introns have BP positions closer to the 3′ss. U12-type constitutive introns display a bimodal distribution of branchpoint positions relative to the 3′ss. Interestingly, U12 cintron+ introns occupy positions in the peak more proximal to the 3′ss (**Fig. 3F**). By binning U2- and U12-type introns by branchpoint position, we found a strong correlation between branchpoint position and the proportion of cintron-generating introns (**Fig. 3G**).

Branchpoint position and the polypyrimidine tract length are in essence determinants of lariat tail length. Our sequence analyses suggest that introns producing lariats with shorter tail lengths exhibit higher rates of cintron formation (**Fig. 3H**). U12-type introns naturally have shorter lariat tails than U2-type introns due to their branchpoint distribution and lack of a polypyrimidine tract, which may explain the higher rate of cintron formation observed in these introns.

3.3.4 *The spatial configuration of the BP and 3'ss within the lariat impacts cintron formation*

Given that lariat tail length is strongly associated with the propensity to form cintrons, we investigated other tail-related features that might influence cintron formation. We hypothesized that 3′ss-BP proximity facilitated by RNA secondary structure favors cintron formation, motivated by our observation that the predicted secondary structures of the 10 introns that we profiled experimentally exhibited starkly different profiles when stratified by their propensity to form

cintrons (**Fig. 4A**). We investigated this observation genome-wide by measuring the average 3′ss-BP distance after lariat tail folding in BP-annotated RefSeq constitutive introns. We enumerated a list of ViennaRNA-predicted secondary structures for each lariat tail sequence (Lorenz et al., 2011). We then calculated the ensemble-average 3′ss-BP distance, weighted by folding state probability (Ding and Lawrence, 2003). Since most human introns produce multiple unique lariats due to BP multiplicity, we calculated an intron-level, BP usage frequency-weighted average of the respective ensemble distances per intron. We found that introns which have, on average, shorter 3′ss guanosine and BP nucleotide separations within the lariat conformation are correlated with increased propensity to form cintrons (**Fig. 4B**).

Our secondary structure analysis also revealed that a notable proportion of introns bear lariats whose tails have no predicted folding structures (9,492 U2-type introns and 202 U12-type introns). In these cases, we hypothesized that the 3′ss and the BP could come into close proximity simply via random fluctuations of the lariat tail in three-dimensional space. To test this hypothesis, we modeled the ring-closure probabilities of single-stranded RNA using the Worm-like Chain model for semi-flexible chains (Guerin, 2017; Kratky and Porod, 1949). We specifically estimated the probability that the BP and the 3′ss will be separated by a length of 2 nucleotides at most, as a function of lariat tail length. This biophysical analysis revealed that lengths between 10 and 15 nucleotides are predicted to have the highest probability of end-to-end proximation (**Fig. 4C**). Interestingly, our genomic analysis of actual cintron abundance produced concordant results; introns with unstructured lariat tails whose lengths fall between 10 and 15 nucleotides produce the most cintrons, consistent with biophysical predictions (**Fig. 4D**). Our analyses suggest that the distance of the BP to the 3′ss within the lariat conformation may facilitate cintron formation, and

that secondary structure or random thermal fluctuations of the lariat tail are possible mechanisms to achieve juxtaposition (**Fig. 4E**).

3.3.5 *Cintron expression displays tissue-specific differences*

To determine if cintrons show tissue-specific expression patterns, we analyzed RNA-seq data from the Genotype-Tissue Expression (GTEx) project to assess cintron expression per tissue type. We found that cintrons are most abundant in neural tissue and testis, similar to previous reports of tissue-specific enrichment of exonic circular RNAs (**Fig. 5A**) (Maass et al., 2017; Memczak et al., 2013; Rybak-Wolf et al., 2015; Szabo et al., 2015). We experimentally recapitulated this finding by performing direct lariat and cintron sequencing of the *SRSF3* cintron-producing intron among eight distinct human tissues and quantifying total cintron abundance (**Fig. 5B**). These tissue-specific differences may reflect the differences in rates of cellular turnover, such as dilution with higher rates of cell division (e.g., bladder and gastrointestinal tissues) and accumulation with lower rates of cell division (e.g., nervous tissues). To assess if there are particular gene networks associated with cintron production, we performed Gene Ontology (GO) enrichment analyses to compare genes containing at least one cintron+ intron against the set of parental genes of cintron– introns. We found a diverse range of enriched terms, notably including many terms related to gene functions in RNA processing and metabolism (**Fig. 5C**).

3.4 DISCUSSION

Circular RNAs were first shown to be abundant in human cells a decade ago (Salzman et al., 2012). Since then, regulatory functions and disease relevance are now widely recognized for exon-derived circular RNAs, the best-studied class of circular RNAs (Kristensen et al., 2022). Our study

describes the genome-wide prevalence and possible provenance of cintrons, a recently described and understudied class of intronic circular RNAs.

Although the molecular mechanisms underlying cintron formation remain to be elucidated, our data suggest possible origins. In particular, our observation of frequent, non-templated, poly-adenosine (poly-A) stretches within the cintron junction suggests a lariat origin for this circular RNA species. 3′-end polyadenylation by the Trf4/5-Air1/2-Mtr4 (TRAMP) complex directs non-coding transcripts towards exosome-mediated RNA decay (reviewed in: (Houseley et al., 2006; Houseley and Tollervey, 2009; Schmid and Jensen, 2008; Schmidt and Butler, 2013). Intron lariats can associate with the exosome complex, as seen in the Drosophila mirtron maturation pathway, where formation of miRNAs from lariat precursors (Okamura et al., 2007; Ruby et al., 2007) involves lariat tail trimming by the exosome (Flynt et al., 2010). The poly-A signature that we observe is consistent with lariat tail polyadenylation and escape from degradation, followed by circularization.

We hypothesize two potential modes of intron circularization from a lariat precursor. In the first mode, the lariat 2′-5′ linkage is debranched via the debranching enzyme *DBR1* or the debranching activity of the spliceosome itself (Chen et al., 2018), with subsequent ligation of the 5′ and 3′ ends. *RTCB* (also known as *HSPC117*) is the sole known 5′-3′ RNA ligase in humans whose cognate substrates are tRNAs following intron removal via a splicing endonuclease (Chakravarty and Shuman, 2012; Chakravarty et al., 2012; Popow et al., 2011). However, a recent preprint reported that the uncharacterized gene *C12orf29* encodes an RNA ligase as well, suggesting an additional path for ligation (Yuan et al., 2022). In the second mode, circularization could be achieved through a transesterification reaction involving nucleophilic attack of the 2′-5′ phosphodiester bond by the 3′OH of the 3′ss, a debranching-independent process. More work is

needed to test if these or other mechanisms are important contributors to cintron formation. Both of these modes of circularization require adjacency of the 3′ss to the 5′ss-BP junction, which could be subject to the RNA structural and biophysical constraints that we studied, mediated by RNA-binding proteins, or influenced by as-yet-unknown factors (**Fig. 4E**).

Cintrons could play potential functional roles through diverse means. First, cintrons could possibly modulate parental gene expression by sequestering miRNAs, directly engaging with the transcriptional machinery, or recruiting regulators of transcription to promoter regions. Second, citrons could play roles in splicing regulation by "sponging" intron-binding splicing factors or by promoting their degradation if bound simultaneously with their corresponding ubiquitin ligases. Finally, if translated like some exon-derived circular RNAs, cintrons could be an unrecognized source of proteomic diversity. Similar functions have been reported for exon-derived circular RNAs (Kristensen et al., 2019; Li et al., 2018; Patop et al., 2019).

3.5 FIGURES

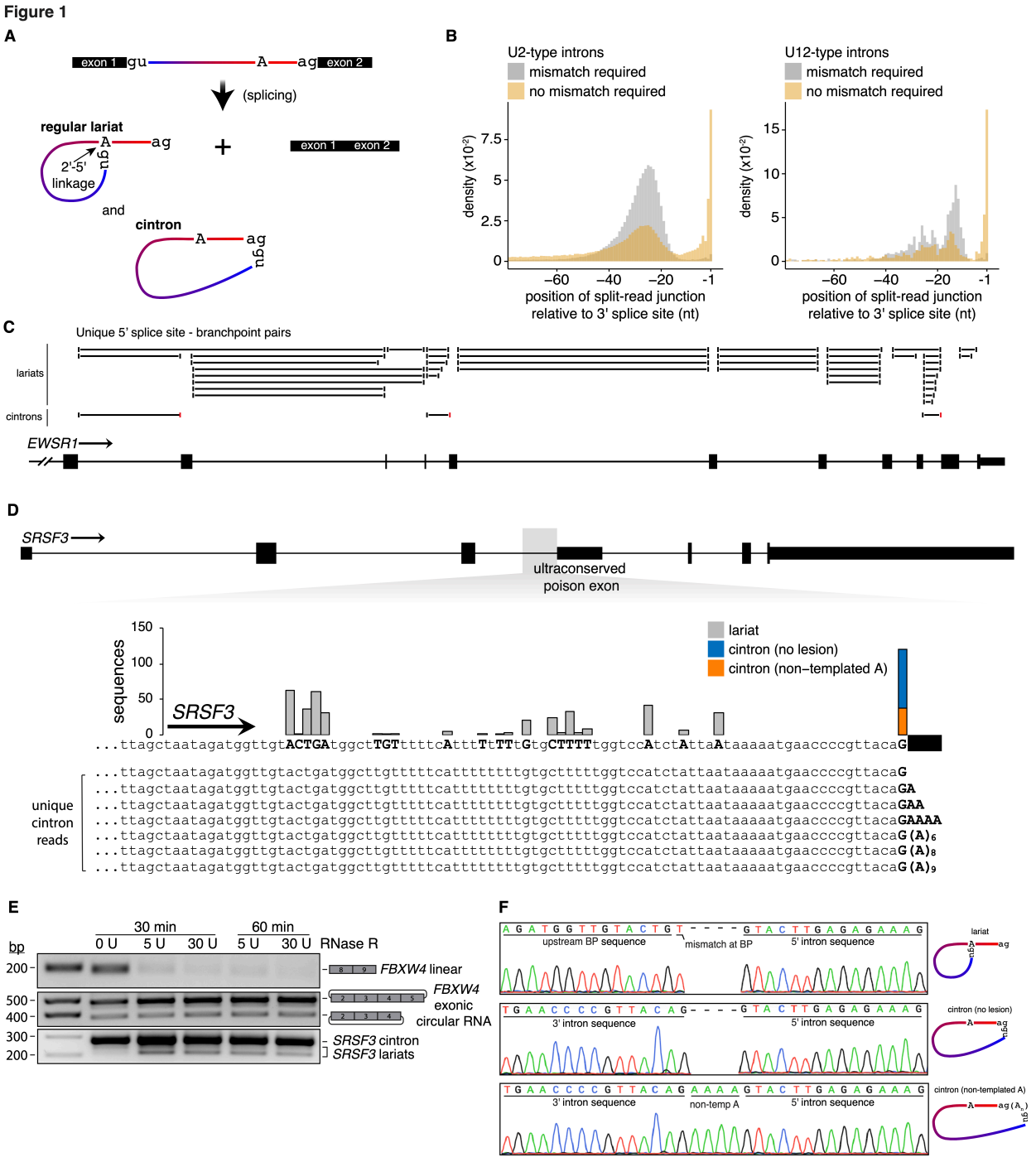


Figure 1. Human introns fully circularize with frequent addition of non-templated adenosines

(**A**) Cintrons are splicing products distinct from intron lariats.
(**B**) Histogram of branchpoint positions, relative to the 3′ss, for RefSeq constitutive introns. Color coding indicates hits identified by requiring or not requiring the mismatch at the junction nucleotide. The −1 nt position corresponds to the last intronic nucleotide.
(**C**) Branchpoint positions and cintron-producing introns annotated for the 3′ end of *EWSR1* from RNA-seq analysis. Unique pairings (horizontal black lines) of the 5′ss (upstream, vertical black bars) with either the branchpoint (downstream, vertical black bars) or the 3′ss guanosine (downstream, vertical red bars) are shown. The diagram was generated through the University of California at Santa Cruz (UCSC) Genome Browser (Meyer et al., 2013).
(**D**) The number of lariat- or cintron-identifying reads for the intron upstream of the *SRSF3* ultraconserved poison exon from direct lariat and cintron sequencing. The junction nucleotide positions (bold uppercase letters) are shown in the intronic reference sequence. For simplicity, the 5′ss sequences of the unique cintron reads are not displayed.
(**E**) RT-PCR gel from RNase R digestion experiments showing exonic (*FBXW4* control) and intronic (*SRSF3* lariat and cintron) circular RNAs resistant to exonucleolytic degradation.
(**F**) Representative chromatograms from Sanger sequencing of the 5′ss-BP or 5′ss-3′ss junctions for *SRSF3* lariats or cintrons, respectively. The lariat and cintron sequences displayed are from the 60 min 5 U RNase R and 60 min 30 U RNase R digestion conditions, respectively.

Figure 2

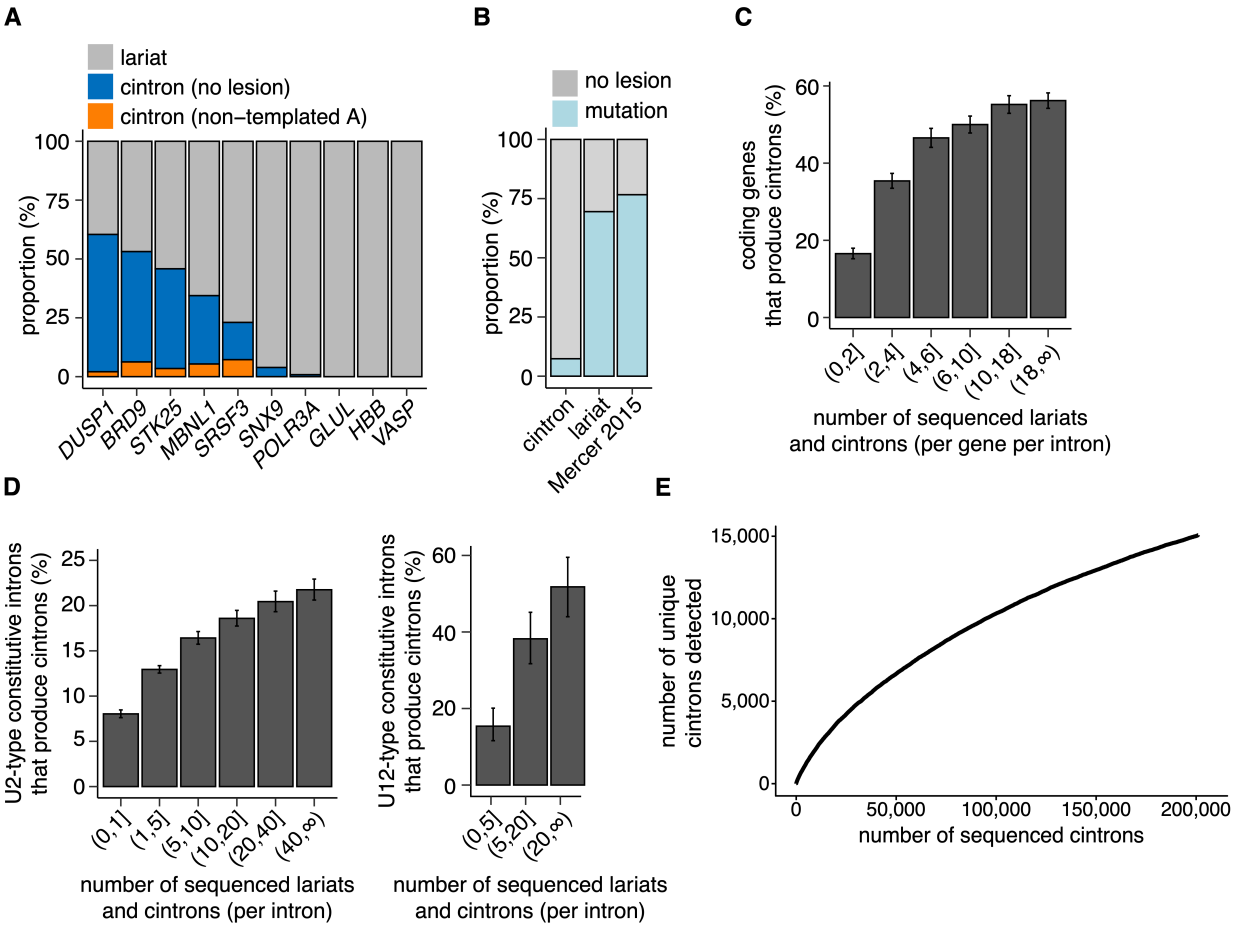


Figure 2. Cintrons have circularization junctions linked via a 5′-3′ phosphodiester bond and are widespread across the genome

(**A**) The proportion of lariats and cintrons from nested PCR using outward facing primers for selected RefSeq constitutive introns.

(**B**) The proportion of inverted reads containing a mutation at the junction nucleotide from direct lariat and cintron sequencing. Experimental cintronic junction sequences containing non-templated adenosines were excluded from the analysis.

(**C**) The proportion of coding genes that produce cintrons as a function of lariat and cintron sequencing depth standardized by the number of introns per gene. The 95% confidence intervals were estimated via a proportion test.

(**D**) The proportions of cintron-producing U2- or U12-type introns as a function of lariat and cintron sequencing depth.

(**E**) Cintron detection rate as a function of sequenced cintrons. The number of unique cintrons are displayed from random sampling of all sequenced 5′ss-3′ss junctions.

Figure 3

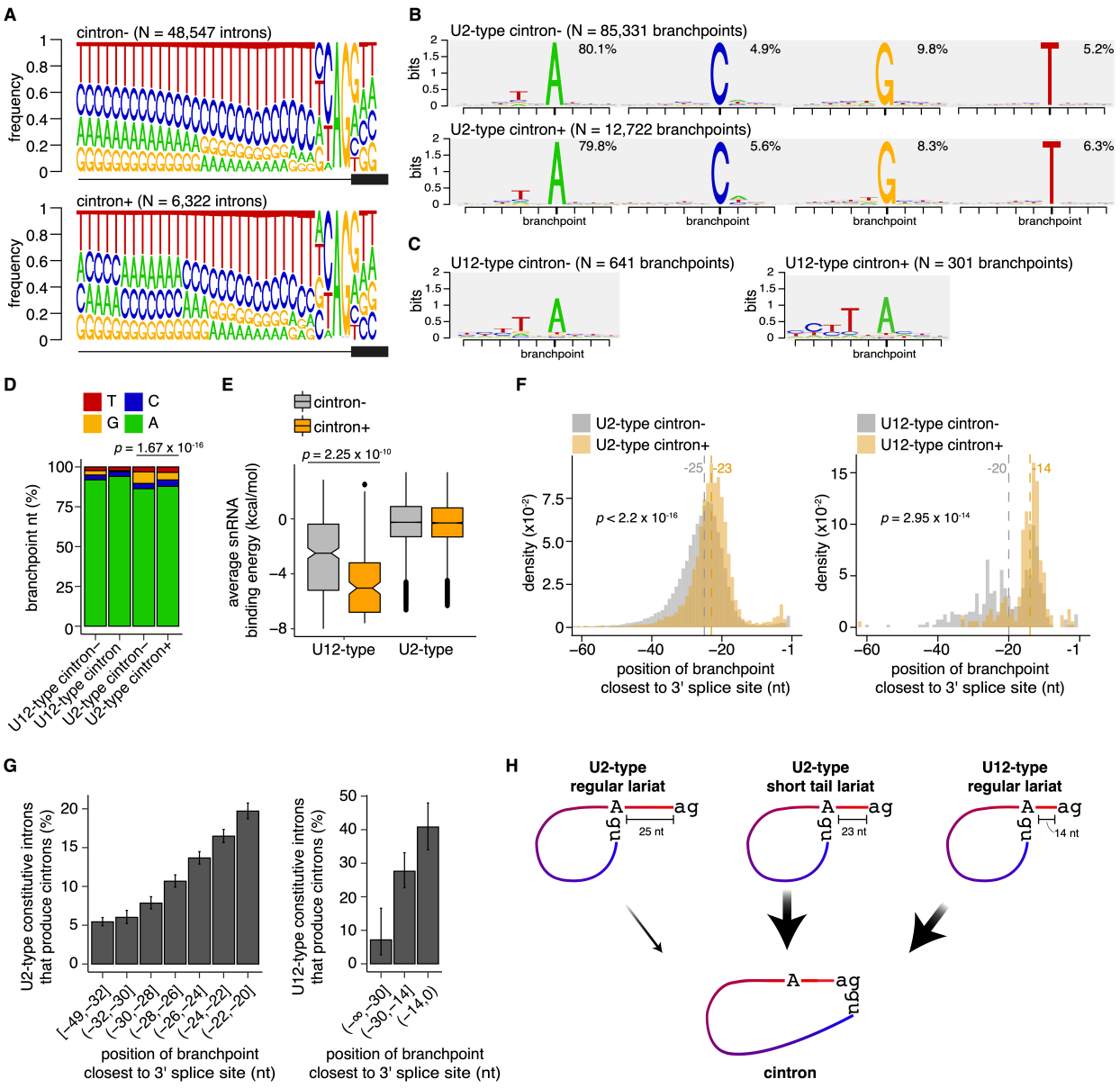


Figure 3. Lariat tail length is correlated with cintron abundance

(**A**) Sequence logos of the 3′ss sequences of cintron-producing (cintron+) or -null (cintron–) introns.

(**B**) Sequence logos of branchpoint sequence contexts. For U2-type introns, branchpoint sequences were stratified based on the branchpoint nucleotide used.

(**C**) As in (B), but illustrating U12-type introns. The branchpoint sequences were not stratified by branchpoint nucleotide.

(**D**) Branchpoint nucleotide usage frequencies for U2- and U12-type cintron+ and cintron– introns. The *p*-value was estimated via a multinomial proportion test.

(**E**) Estimated binding energies of branchpoint sequences to the either the U2 (AUGAUGUG) or U12 (AGGAAUG) snRNA sequences. The values displayed are per intron averages across annotated branchpoints, weighted by usage frequency The *p*-value was estimated via the Mann-Whitney *U* test.
(**F**) Histograms of positions of the most 3′ss-proximal branchpoints within the RefSeq constitutive U2- or U12-type introns. The vertical dashed lines illustrate the median branchpoint position. The −1 nt position corresponds to the last intronic nucleotide. The *p*-values were estimated via the Mann-Whitney *U* test.
(**G**) The fraction of U2- or U12-type introns that produce cintrons as a function of branchpoint position relative to the 3′ss. For each intron, the branchpoint closest to the 3′ss is shown. 80% of all annotated branchpoints within U2-type introns fall in the range of positions displayed (left). The 95% confidence intervals were estimated via a proportion test.
(**H**) Schematic illustrating that lariat tail length is correlated with cintron formation rate. The BP-3′ss distances shown correspond to the median for each lariat class, as shown in (**F**).

Figure 4

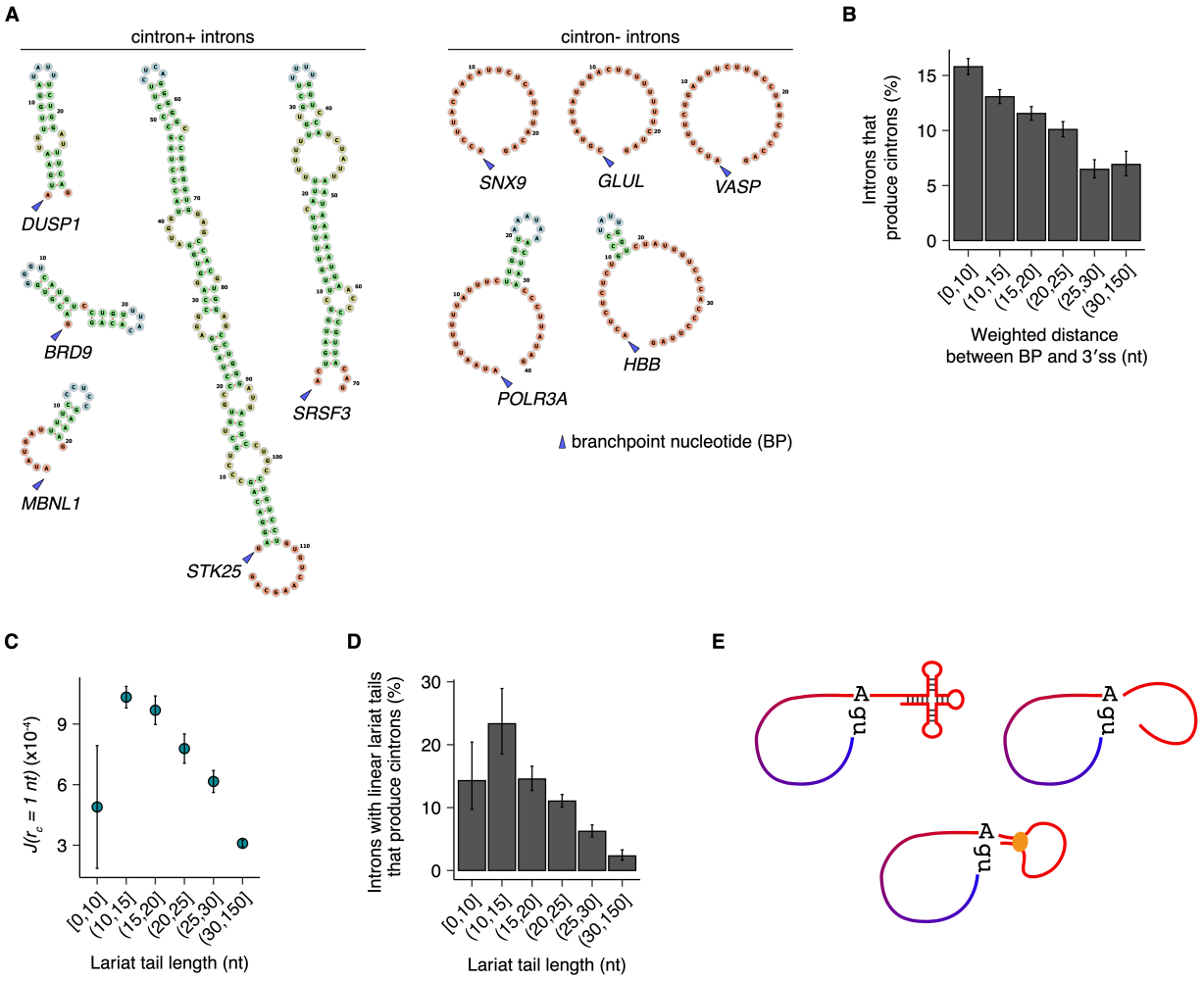


Figure 4. The distance between the 3′ss and the 5′ss-BP junction within the lariat conformation influences cintron formation rate

(**A**) MXfold2-predicted secondary structure of lariat tails for introns subjected to direct lariat and cintron sequencing. The tail sequences shown correspond to the most frequently used branchpoint. The branchpoint nucleotide (blue triangle) was excluded from binding and the lariat loop (not shown) was not modeled to bind with the lariat tail.

(**B**) The percentage of introns that produce cintrons as a function of the distance between the 3′ss and the 5′ss-BP junction after secondary structure formation of the tail. The distances shown are weighted by ensemble folding state probability and branchpoint usage frequency. The 95% confidence intervals were calculated from a proportion test.

(**C**) The Jacobson-Stockmayer cyclization factor (J-factor), as a function of lariat tail length. A capture radius (r_c) of 1 nucleotide length was used which corresponds to estimating the probability of BP-3′ss separation < 2 nucleotides.

(**D**) The proportion of introns that produce cintrons as a function of lariat tail length. RefSeq constitutive introns with no predicted lariat tail secondary structure are shown.

(**E**) Schematic of possible mechanisms for proximation of the 5′ss-BP linkage and the 3′ss: lariat tail secondary structure formation (upper left), lariat tail random spatial fluctuations (upper right), or RNA-binding protein-mediated juxtaposition (lower).

Figure 5

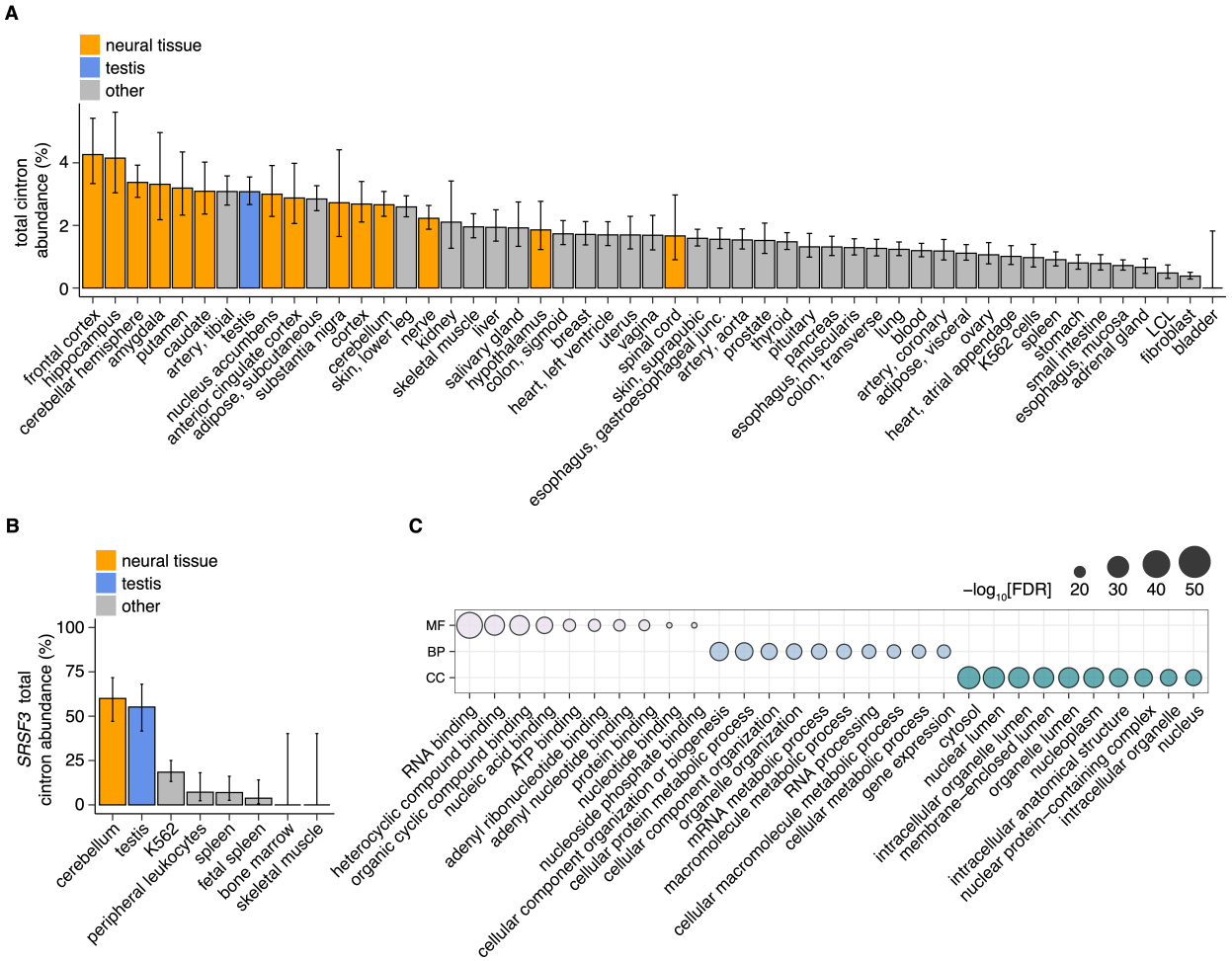


Figure 5. Cintron expression displays tissue-specificity

(**A**) Total cintron abundance, as the proportion of cintron-identifying reads from the lariat-cintron total, estimated from the Genotype-Tissue Expression (GTEx) Project.

(**B**) Total cintron abundance, as in (A), from the intron upstream of the *SRSF3* ultraconserved poison exon measured via direct lariat and cintron sequencing on total human RNA from various human tissues.

(**C**) Gene Ontology (GO) enrichment analysis of genes with cintron-producing introns. The top 10 GO terms from the molecular function (MF), biological process (BP), and cellular compartment (CC) domains are shown.

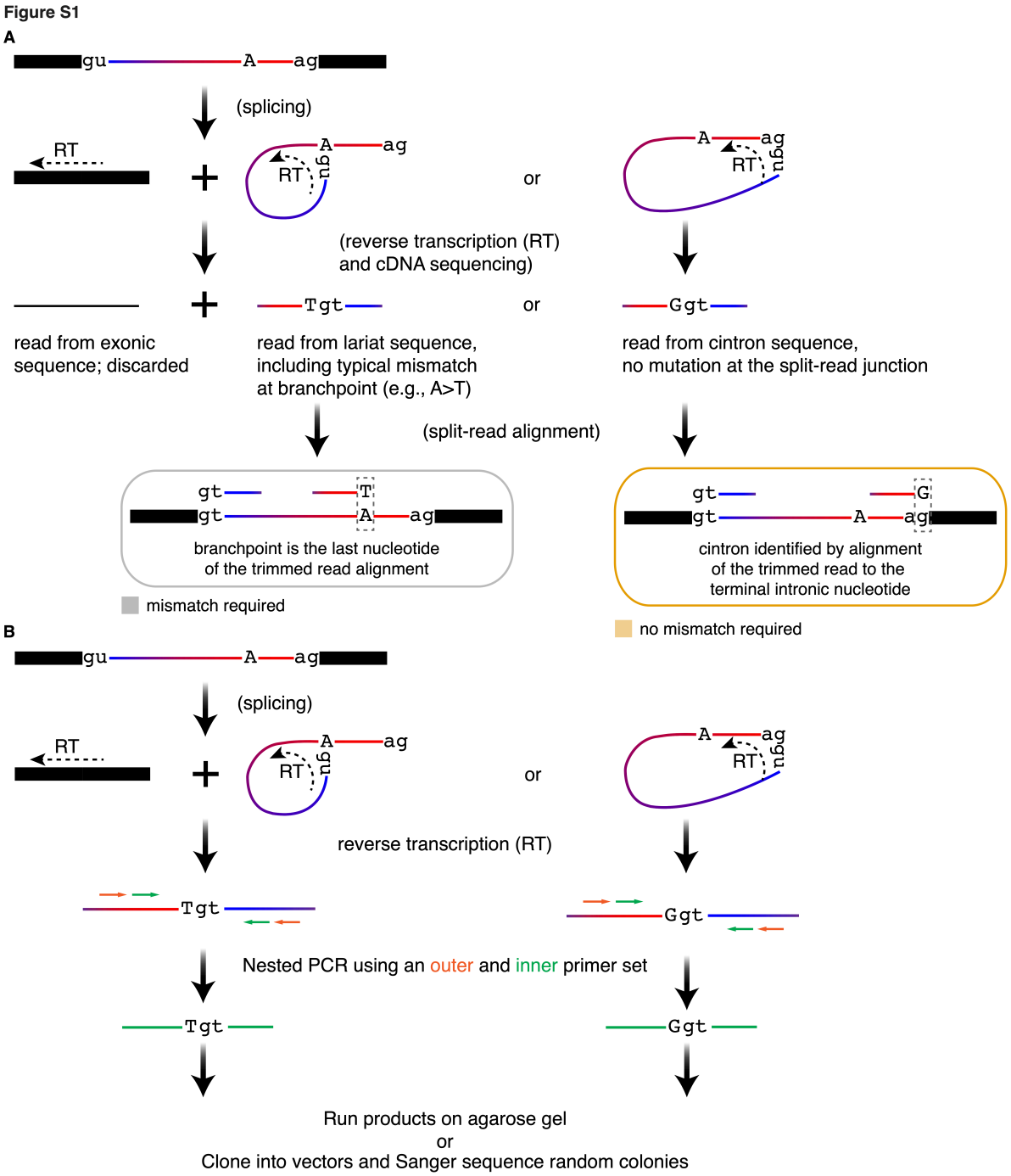


Figure S1. Strategy to detect lariats and cintrons

(**A**) The algorithm to detect branchpoints from RNA-seq data. To detect cintrons, the mismatch requirement at the junction nucleotide is removed, and alignment of the upstream portion of the read to the 3′ss is required.

(**B**) Experimental procedure for direct lariat and cintron sequencing.

3.6 TABLES

Gene	Intron	PCR	Direction	Sequence (5'->3')	Notes
SRSF3	3	outer	F	TCAAAATCTTGCCCCTTTTG	
SRSF3	3	outer	R	CAAAAGCCAACACTCAGCAC	
SRSF3	3	inner	F	ACAGCACACTGTTGCCCATC	
SRSF3	3	inner	R	CATACCCCAAATTACACCCAAC	
POLR3A	4	outer	F	TATTGGGAAACGGACCTCTC	
POLR3A	4	outer	R	GAGAAAAGCTGACTCCCGAAC	
POLR3A	4	inner	F	TATTGGGAAACGGACCTCTC	
POLR3A	4	inner	R	AACAGAAGACAGTGAGTGAAAAGG	
MBNL1	7	outer	F	AAGCCTGTTTGTGTCAATTTTC	
MBNL1	7	outer	R	GGAAATGGACTTGCCCAATAG	
MBNL1	7	inner	F	TGTCAATTTTCTTGATTTGATGG	
MBNL1	7	inner	R	ATTTTGAGGGGCTGTGAGG	
SNX9	10	outer	F	AGGGAGATAGAGTGGGAGCTG	
SNX9	10	outer	R	CTTCTGGCAGGCAGTTCTTC	
SNX9	10	inner	F	AATGAGCTGGTCCTGTTTGG	
SNX9	10	inner	R	AGGTTTCTGTCCCCTCACTG	
VASP	9	outer	F	GCACCCTTATAGGAGAGTCAGG	
VASP	9	outer	R	TAGTTCCTGTGGCTGGACTG	
VASP	9	inner	F	GCACCCTTATAGGAGAGTCAGG	
VASP	9	inner	R	GGCTGGACTGGGCACTCTAC	
GLUL	3	outer	F	GACTCGGTGATCCCAAGAAG	
GLUL	3	outer	R	TCCTCAAAGCAGAAGTTCAGG	
GLUL	3	inner	F	GACTCGGTGATCCCAAGAAG	
GLUL	3	inner	R	GTCTTCTCCCTCCCCTTCAC	
STK25	4	outer	F	AAAGCCAAATGCAGGAAGAG	
STK25	4	outer	R	CGGCCTCTCTAAGGTCAGTG	
STK25	4	inner	F	CTGGAGCTCGGAGTGGAG	
STK25	4	inner	R	GACGTTACAAGAGCCACATCC	
HBB	1	outer	F	GAAACTGGGCATGTGGAGAC	
HBB	1	outer	R	AAACCTGTCTTGTAACCTTGATACC	
HBB	1	inner	F	TGGAGACAGAGAAGACTCTTGG	
HBB	1	inner	R	AAACCTGTCTTGTAACCTTGATACC	
BRD9	14	outer	F	CGGTGCTAGGTCTTGTCTGATG	
BRD9	14	outer	R	GGAGCTATGGTCCAAAAACCTG	
BRD9	14	inner	F	CTGCCTTCCCTTCACTCACTG	
BRD9	14	inner	R	TGGGACAAAGACCAAATACCG	
DUSP1	2	outer	F	CTCCAATTGTAGGCTCTACGG	
DUSP1	2	outer	R	CTCCCTGGCACTACTCTTCC	
DUSP1	2	inner	F	CAGGCAAATGGGCTTAGTTC	
DUSP1	2	inner	R	CTCCCTGGCACTACTCTTCC	

FBXW4	2	circExon control	F	GGGATTCTGCTGAAGTGGAG	From Mercer, et al 2015
FBXW4	2	circExon control	R	TCCTTCACTGGGACACTGGT	From Mercer, et al 2015
FBXW4	8, 9	linear control	F	CGGAAATGTGTCATGGAGTG	From Mercer, et al 2015
FBXW4	8, 9	linear control	R	CAGTACACAGGGCTGCTGAG	From Mercer, et al 2015

Table S1. Primers for cintron sequencing. PCR primers used for direct lariat and cintron sequencing.

Dataset name	Reads (billion)	Accession number(s)	PMID(s)
2010/illumina.body_map_2	2.4	ERP000546 (ENA)	NA
2012/akimitsu.rna_stability	2.3	DRA000345, DRA000346, DRA000347, DRA000348, DRA000350, DRP000622 (DDBJ)	22369889, 23064110
2012/encode-gingeras.shortRNA_subcellular_fractions	13	GSE24565 (GEO)	22955620
2012/encode-gingeras.subcellular_fractions	5.1	GSE30567 (GEO)	22955620
2012/lopez-otin-quesada.chronic_lymphocytic_leukemia	2.4	EGAD00001000083 (EGA)	22158541, 23187290
2013/abdel-wahab.acute_myeloid_leukemia	1.4	NA	25965569
2013/aerts-cools.acute_lymphoblastic_leukemia	2.3	NA	24367274
2013/bradley-ramakrishnan.chronic_myelomonocytic_leukemia	1.1	NA	25965569
2013/bradley-ramakrishnan.spliceosomal_mutations	3.3	GSE58871,GSE65349 (GEO)	25267526, 25965569
2013/bradley-tapscott.fshd	1.4	GSE45883 (GEO)	24278031
2013/bradley.nonsense_mediated_decay	1.1	GSE58335, GSE61410 (GEO)	25385641
2013/carroll.B-lymphoblastic_leukemia	1.4	SRP009840 (SRA)	23377183
2013/ecker-ren.ucsd_epigenome_mapping	2.5	GSE16256 (GEO)	21289626
2013/geuvadis-dermitzakis.1000_genomes	11	E-GEUV-1 (ENA)	24037378
2013/gray.breast_cancer_celllines	4.2	GSE48213 (GEO)	24176112
2013/khaitovich-gelfand-chen.aging_brain	1.1	SRP005169 (SRA)	23340839
2013/paddison.glioblastoma	1.8	GSE75147 (GEO)	23651857
2013/sauvageau.acute_lymphoblastic_leukemia	2.3	GSE49601 (GEO)	24069164
2013/sauvageau.acute_myeloid_leukemia	28	GSE49642, GSE52656, GSE62190, GSE66917, GSE67039 (GEO)	24069164, 26237430, 26968532
2013/tapscott.fshd_patients	3.6	GSE56787 (GEO)	24861551
2013/TCGA.ACC	3.2	NCI Genomic Data Commons	
2013/TCGA.BLCA	18	NCI Genomic Data Commons	
2013/TCGA.BRCA	78	NCI Genomic Data Commons	

2013/TCGA.CESC	18	NCI Genomic Data Commons	
2013/TCGA.CHOL	2	NCI Genomic Data Commons	
2013/TCGA.COAD	20	NCI Genomic Data Commons	
2013/TCGA.DLBC	4.4	NCI Genomic Data Commons	
2013/TCGA.ESCA	74	NCI Genomic Data Commons	
2013/TCGA.GBM	38	NCI Genomic Data Commons	
2013/TCGA.HNSC	31	NCI Genomic Data Commons	
2013/TCGA.KICH	4.4	NCI Genomic Data Commons	
2013/TCGA.KIRC	43	NCI Genomic Data Commons	
2013/TCGA.KIRP	17	NCI Genomic Data Commons	
2013/TCGA.LAML	43	NCI Genomic Data Commons	
2013/TCGA.LGG	26	NCI Genomic Data Commons	
2013/TCGA.LIHC	21	NCI Genomic Data Commons	
2013/TCGA.LUAD	27	NCI Genomic Data Commons	
2013/TCGA.LUSC	30	NCI Genomic Data Commons	
2013/TCGA.MESO	7.2	NCI Genomic Data Commons	
2013/TCGA.OV	87	NCI Genomic Data Commons	
2013/TCGA.PAAD	11	NCI Genomic Data Commons	
2013/TCGA.PCPG	11	NCI Genomic Data Commons	
2013/TCGA.PRAD	30	NCI Genomic Data Commons	
2013/TCGA.READ	6.8	NCI Genomic Data Commons	
2013/TCGA.SARC	14	NCI Genomic Data Commons	
2013/TCGA.SKCM	24	NCI Genomic Data Commons	

2013/TCGA.STAD	180	NCI Genomic Data Commons	
2013/TCGA.TGCT	6.9	NCI Genomic Data Commons	
2013/TCGA.THCA	26	NCI Genomic Data Commons	
2013/TCGA.THYM	9.7	NCI Genomic Data Commons	
2013/TCGA.UCEC	22	NCI Genomic Data Commons	
2013/TCGA.UCS	2.8	NCI Genomic Data Commons	
2013/TCGA.UVM	7.4	NCI Genomic Data Commons	
2013/west.breast_cancer_stroma	3.8	GSE42948 (GEO)	24342436
2013/white-mcnerney.acute_myeloid_leukemia	1.9	SRP017262 (SRA)	23212519
2014/conboy.erythropoiesis	1.1	GSE53635 (GEO)	24442673
2014/gallagher.erythropoiesis	1.1	GSE53983 (GEO)	24637361
2014/kim.colorectal_cancer	4.1	GSE50760 (GEO)	25049118
2014/myers.breast_cancer	27	GSE581350 (GEO)	24929677
2014/su2c.prostate_cancer	25	phs000915.v1.p1 (dbGaP)	26000489
2014/target-meshinchi.acute_myeloid_leukemia	43	phs000218.v18.p7 (dbGaP)	NA
2014/varmus.u2af1_mutations	4.8	GSE80136 (GEO)	27776121
2015/boultwood.sf3b1_mutations	2.3	GSE63569 (GEO)	25428262
2015/buonamici.sf3b1_mutations	2.2	GSE72790 (GEO)	26565915
2015/koeffler.zrsr2_mutations	1.2	GSE63816 (GEO)	25586593
2015/mattick.branchpoint_discovery	2.2	GSE53328 (GEO)	25561518
2016/berglund-wang.myotonic_dystrophy	3.7	GSE86356 (GEO)	27681373
2016/cairns.embryogenesis	9.6	GSE85632 (GEO)	28459457
2016/chen-chen.acute_lymphoblastic_leukemia	12	Chinese Genotype-phenotype Archive	27428428
2016/fioretos.acute_lymphoblastic_leukemia	5.1	EGAD00001002112 (EGA)	27265895
2016/gtex.tissues	550	phs000424.v6.p1 (dbGaP)	25954002
2016/mano.acute_lymphoblastic_leukemia	13	JGAS00000000047 (JGA)	27019113
2016/stern.uveal_melanoma	3.9	NA	26842708
2015/garraway-schadendorf.melanoma_checkpoint_blockade	1.3	phs000452.v2.p1 (dbGaP)	26359337
2015/gilad-pritchard.hapmap_yoruba	5.2	GSE61742 (GEO)	25657249

2015/papapetrou.srsf2_mutations	1.5	SRP108379 (SRA)	29681544
2015/petrucelli.als	3.5	GSE67196 (GEO)	26192745
2015/snyder.1000_genomes	6.5	GSE65912 (GEO)	26297486
2016/bradley-porter.breast_cancer	3.1	NA	31434678
2016/hammerbacher.melanoma_checkpoint_blockade	1.4	NA	27956380
2016/lo.melanoma_checkpoint_blockade	1.9	GSE78220 (GEO)	26997480
2017/chinnaiyan.metastatic_cancer	76	phs000673.v3.p1 (dbGaP)	28783718
2017/weil.p-body	1.7	E-MTAB-5477,E-MTAB-4091 (ArrayExpress)	28965817
2017/yang-yeoh.acute_lymphoblastic_leukemia	14	EGAD00001002151 (EGA)	27903646
2018/boultwood.myelodysplasia	35	GSE114922 (GEO)	29930011
2018/perou.metastatic_breast_cancer	6.3	phs000676.v2.p2 (dbGaP)	29480819
2018/powles.urothelial_cancer_checkpoint_blockade	11	EGAD00001003977 (EGA)	29443960
2019/BeatAML.cohort_1	41	NA	30333627
2019/schadendorf-van_allen.melanoma_checkpoint_blockade	3	phs000452.v3.p1 (dbGaP)	31792460
2019/wilmott.melanoma_checkpoint_blockade	7.4	PRJEB23709 (SRA)	30753825

Table S2. RNA-seq datasets analyzed in this study. Published or publicly available datasets with >1 billion available reads that were analyzed in this study. Accession numbers and/or PMIDs are listed when available.

3.7 MATERIALS AND METHODS

3.7.1 *Method Details*

3.7.1.1 Genome annotations

We merged the UCSC knownGene (Meyer et al. 2013), Ensembl 71 (Flicek et al. 2013), and MISO v.2.0 (Katz et al. 2010) annotations for the UCSC hg19 (GRCh37) genome assembly. We further expanded this annotation by generating all possible combinations of annotated 5' and 3' splice sites, for each gene.

3.7.1.2 Gene expression and alternative splicing analysis

RSEM v.1.2.4 (Li and Dewey 2011) with Bowtie (Langmead et al. 2009), using the option "-v 2," was utilized to map all reads to the human transcriptome. RSEM generates gene expression estimates (TPM, transcripts per million) which were normalized via the trimmed mean of M values (TMM) method (Robinson, et al. 2010). The unaligned reads were mapped to the genome and the merged splice junction database using TopHat v.2.0.8b (Trapnell et al. 2009). To quantify isoform expression, the combined RSEM and TopHat alignments were inputted to MISO v.2.0.

3.7.1.3 Cintron detection

Our branchpoint detection strategy (Pineda & Bradley, 2018) was modified to detect cintron-identifying reads, as illustrated in Supplemental Figure S1A. In brief, inverted reads with unique split-read alignments to 5' and 3' splice sites within a single gene were queried for mutations. Positions with ≥75% mapped reads having no lesions in the 3' splice site sequence, including at the split-read junction nucleotide, are classified as cintrons.

3.7.1.4 Sequence motif and Gene Ontology analyses

Sequence logo plots were created using the GenomicRanges package from Bioconductor. GOseq (Young et al., 2010), was used to identify enriched Gene Ontology terms after comparing genes containing cintron+ introns against the set of genes lacking cintron+ introns. A false-discovery rate cutoff of 0.01 was used to demarcate statistically significant enrichment. The visualizations were created using the dplyr, and ggplot2 packages.

3.7.1.5 RNA extraction and cDNA synthesis

K562 total RNA was isolated from TRIzol-lysed cells (Thermo Fisher Scientific) then cleaned using the RNase-free DNase set (Qiagen). Human total RNA (peripheral blood mononuclear cells, cerebella, testes, spleens, and fetal spleens) was purchased from Takara Bio. The SuperScript III first strand synthesis system (Thermo Fisher Scientific) with random hexamer priming was used to generate cDNA from total RNA. These methods utilized the protocols specified by their respective manufacturers.

3.7.1.6 Sanger sequencing and analysis

Nested PCR primers, see **Supplemental Table 1** (**Table S1**), were obtained from Integrated DNA Technologies. A pair of forward primers were designed upstream of known branchpoint clusters. A pair of reverse primers were designed immediately downstream of the 5′ splice site. This design permits amplification of the lariat branchpoint-5′ splice site and cintron 5′ splice site-3′ splice site junctions, as exemplified in **Supplemental Figure S1B**. Two 30-cycle gradient PCR rounds were performed with Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific). Each round contained eight 50 µl replicates whose annealing temperatures ranged Tm ±3 °C. For the first round, K562 or human organ total RNA was used as a template. The PCR products were pooled, cleaned, and concentrated with the QIAquick PCR purification kit (Qiagen). The first-round

products served as templates for the second 30-cycle gradient PCR using the inner primer set. The second-round replicates were pooled and electrophoresed on a 2% agarose gel. Bands in the sizes ranges corresponding to the expected lengths of the lariat or cintron amplicons were excised, and the DNA was isolated with the MinElute gel extraction kit (Qiagen). The ZeroBlunt TOPO PCR cloning kit (Thermo Fisher Scientific) was used to clone the gel-extracted fragments into pCRBlunt II-TOPO vector (Thermo Fisher Scientific), which were transformed into TOP10 chemically competent Escherichia coli (Thermo Fisher Scientific). The transformants were plated on 50 µg/mL LB + kanamycin plates, and the vector inserts of randomly selected colonies were Sanger sequenced (Genewiz) after growth. Branchpoints and cintrons were identified based on mutations at the split-read junction nucleotide and lesion-less 5′ splice site-3′ splice site junction sequences, respectively.

3.7.1.7 RNase R digestion

In 25 µl duplicates, 1 µg of human cerebellum total RNA was combined with water, RNase R buffer (Lucigen), and 0 U, 5 U, or 30 U of RNase R (Lucigen). The reactions were incubated at 37 °C for either 30 or 60 minutes. The reactions were pooled and cleaned using the NEB Monarch RNA Cleanup kit (New England Biolabs) after incubation. RNA quality was assessed with a nanodrop. The primers used to amplify the *FBXW4* linear and circular RNA controls are enumerated in **Table S1**.

3.7.1.8 Lariat tail secondary structure analyses

The folding conformations, and their respective free energies, for each lariat tail sequence were generated using the RNAsubopt program of the ViennaRNA suite (Lorenz et al., 2011). The distance of the 3′ss to the BP for each secondary structure was computed by parsing the dot-bracket notation representing intramolecular base-paring. To generate a statistic summarizing the post-

folding distance for a single tail sequence over its ensemble of folding states, an average distance weighted by folding state probability was calculated. We assumed that the predicted secondary structures (Xi) are Boltzmann-distributed, with state probabilities calculated as $P(Xi) = \frac{e^{-E_i/RT}}{Z}$; where E_i is the free energy of the folding state, R is the gas constant, T is temperature (37 °C), and Z is the partition function (Ding and Lawrence, 2003). Due to branchpoint multiplicity, a single intron is frequently associated with multiple lariat tails. To get an intron-level summary, a branchpoint usage frequency-weighted average was calculated from the Boltzmann probability-weighted post-folding 3′ss-BP distances. Secondary structures in **Figure 4A** were created using MXfold2 (Sato et al., 2021). All lariat tail folding predictions were performed with the lariat loop BP nucleotide excluded from bonding. All analyses were conducted within the R Programming environment with tools from Bioconductor. The visualizations were created using the dplyr, and ggplot2 packages.

3.7.1.9 Lariat tail ring-closure probability calculation

Modeling the bending fluctuations of single-stranded RNA was done using the Worm-like Chain Model, as previously described (Guerin, 2017). For lariat tails with no predicted secondary structure, the Jacobson-Stockmayer cyclization factor (J-factor) was calculated: the probability that the branchpoint nucleotide and the 3′ splice site will be contained within a sphere with radius rc (capture radius). Experimental values for the persistence length (lp = 20.75 Å) and the length of a ribonucleotide monomer (derived from the 40-mer poly-U measurement of 196.4 Å) were acquired from the literature (Chen et al., 2012). The lariat tail was modeled as a free chain, ignoring the looped portion of the lariat. All analyses were conducted within the R Programming environment with tools from Bioconductor. Numerical evaluation of the complete elliptical

integrals of the first and second kind were done using the gsl package. The visualizations were created using the dplyr, and ggplot2 packages.

3.8 Acknowledgments

R.K.B. was supported in part by the NIH/NCI (R01 CA251138), NIH/NHLBI (R01 HL128239 and R01 HL151651) and the Blood Cancer Discoveries Grant program through the Leukemia & Lymphoma Society, Mark Foundation for Cancer Research, and Paul G. Allen Frontiers Group (8023-20). R.K.B is a Scholar of The Leukemia & Lymphoma Society (1344-18) and holds the McIlwain Family Endowed Chair in Data Science. The results published here are based in part upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx data used for the analyses described in this manuscript were obtained from dbGaP (accession number phs000424.v6.p1) on February 8, 2017.

3.9 Author Contributions

J.M.B.P., T.R.N., and R.K.B. performed the experiments, analyzed the data, and wrote the paper.

3.10 Competing Interests

The authors declare that they have no competing interests.

3.11 References

Been, M.D., and Cech, T.R. (1985). Sites of circularization of the Tetrahymena rRNA IVS are determined by sequence and influenced by position and secondary structure. Nucleic Acids Res *13*, 8389-8408. 10.1093/nar/13.23.8389.

Cech, T.R., Zaug, A.J., and Grabowski, P.J. (1981). In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell *27*, 487-496. 10.1016/0092-8674(81)90390-1.

Chakravarty, A.K., and Shuman, S. (2012). The sequential 2',3'-cyclic phosphodiesterase and 3'-phosphate/5'-OH ligation steps of the RtcB RNA splicing pathway are GTP-dependent. Nucleic Acids Res *40*, 8558-8567. 10.1093/nar/gks558.

Chakravarty, A.K., Subbotin, R., Chait, B.T., and Shuman, S. (2012). RNA ligase RtcB splices 3'-phosphate and 5'-OH ends via covalent RtcB-(histidinyl)-GMP and polynucleotide-(3')pp(5')G intermediates. Proc Natl Acad Sci U S A *109*, 6072-6077. 10.1073/pnas.1201207109.

Chen, H., Meisburger, S.P., Pabit, S.A., Sutton, J.L., Webb, W.W., and Pollack, L. (2012). Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. Proc Natl Acad Sci U S A *109*, 799-804. 10.1073/pnas.1119057109.

Chen, W., Moore, J., Ozadam, H., Shulha, H.P., Rhind, N., Weng, Z., and Moore, M.J. (2018). Transcriptome-wide Interrogation of the Functional Intronome by Spliceosome Profiling. Cell *173*, 1031-1044 e1013. 10.1016/j.cell.2018.03.062.

Dalgaard, J.Z., and Garrett, R.A. (1992). Protein-coding introns from the 23S rRNA-encoding gene form stable circles in the hyperthermophilic archaeon Pyrobaculum organotrophum. Gene *121*, 103-110. 10.1016/0378-1119(92)90167-n.

Ding, Y., and Lawrence, C.E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res *31*, 7280-7301. 10.1093/nar/gkg938.

Flynt, A.S., Greimann, J.C., Chung, W.J., Lima, C.D., and Lai, E.C. (2010). MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila. Mol Cell *38*, 900-907. 10.1016/j.molcel.2010.06.014.

Grabowski, P.J., Zaug, A.J., and Cech, T.R. (1981). The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena. Cell *23*, 467-476. 10.1016/0092-8674(81)90142-2.

Guerin, T. (2017). Analytical expressions for the closure probability of a stiff wormlike chain for finite capture radius. Phys Rev E *96*, 022501. 10.1103/PhysRevE.96.022501.

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. 10.1006/jmbi.1994.1377.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. Nat Rev Mol Cell Biol *7*, 529-539. 10.1038/nrm1964.

Houseley, J., and Tollervey, D. (2009). The many pathways of RNA degradation. Cell *136*, 763-776. 10.1016/j.cell.2009.01.019.

Kratky, O., and Porod, G. (1949). Diffuse small-angle scattering of X-rays in colloid systems. J Colloid Sci *4*, 35-70. 10.1016/0095-8522(49)90032-x.

Kristensen, L.S., Andersen, M.S., Stagsted, L.V.W., Ebbesen, K.K., Hansen, T.B., and Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet *20*, 675-691. 10.1038/s41576-019-0158-7.

Kristensen, L.S., Jakobsen, T., Hager, H., and Kjems, J. (2022). The emerging roles of circRNAs in cancer and oncology. Nat Rev Clin Oncol *19*, 188-206. 10.1038/s41571-021-00585-y.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell *31*, 147-157. 10.1016/0092-8674(82)90414-7.

Li, X., Yang, L., and Chen, L.L. (2018). The Biogenesis, Functions, and Challenges of Circular RNAs. Mol Cell *71*, 428-442. 10.1016/j.molcel.2018.06.034.

Li-Pook-Than, J., and Bonen, L. (2006). Multiple physical forms of excised group II intron RNAs in wheat mitochondria. Nucleic Acids Res *34*, 2782-2790. 10.1093/nar/gkl328.

Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26. 10.1186/1748-7188-6-26.

Maass, P.G., Glazar, P., Memczak, S., Dittmar, G., Hollfinger, I., Schreyer, L., Sauer, A.V., Toka, O., Aiuti, A., Luft, F.C., and Rajewsky, N. (2017). A map of human circular RNAs in clinically relevant tissues. J Mol Med (Berl) *95*, 1179-1189. 10.1007/s00109-017-1582-9.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333-338. 10.1038/nature11928.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res *25*, 290-303. 10.1101/gr.182899.114.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res *41*, D64-69. 10.1093/nar/gks1048.

Molina-Sanchez, M.D., Martinez-Abarca, F., and Toro, N. (2006). Excision of the Sinorhizobium meliloti group II intron RmInt1 as circles in vivo. J Biol Chem *281*, 28737-28744. 10.1074/jbc.M602695200.

Monat, C., and Cousineau, B. (2016). Circularization pathway of a bacterial group II intron. Nucleic Acids Res *44*, 1845-1853. 10.1093/nar/gkv1381.

Murray, H.L., Mikheeva, S., Coljee, V.W., Turczyk, B.M., Donahue, W.F., Bar-Shalom, A., and Jarrell, K.A. (2001). Excision of group II introns as circles. Mol Cell *8*, 201-211. 10.1016/s1097-2765(01)00300-8.

Nielsen, H., Fiskaa, T., Birgisdottir, A.B., Haugen, P., Einvik, C., and Johansen, S. (2003). The ability to form full-length intron RNA circles is a general property of nuclear group I introns. RNA *9*, 1464-1475. 10.1261/rna.5290903.

Nielsen, H., and Johansen, S.D. (2009). Group I introns: Moving in new directions. RNA Biol *6*, 375-383. 10.4161/rna.6.4.9334.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89-100. 10.1016/j.cell.2007.06.028.

Patop, I.L., Wust, S., and Kadener, S. (2019). Past, present, and future of circRNAs. EMBO J *38*, e100836. 10.15252/embj.2018100836.

Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev *32*, 577-591. 10.1101/gad.312058.118.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C.L., Luhrmann, R., Soll, D., and Martinez, J. (2011). HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science *331*, 760-764. 10.1126/science.1197847.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86. 10.1038/nature05983.

Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the Mammalian Brain Are

Highly Abundant, Conserved, and Dynamically Expressed. Mol Cell *58*, 870-885. 10.1016/j.molcel.2015.03.027.

Saini, H., Bicknell, A.A., Eddy, S.R., and Moore, M.J. (2019). Free circular introns with an unusual branchpoint in neuronal projections. Elife *8*. 10.7554/eLife.47809.

Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., and Brown, P.O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One *7*, e30733. 10.1371/journal.pone.0030733.

Sato, K., Akiyama, M., and Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun *12*, 941. 10.1038/s41467-021-21194-4.

Schmid, M., and Jensen, T.H. (2008). The exosome: a multipurpose RNA-decay machine. Trends Biochem Sci *33*, 501-510. 10.1016/j.tibs.2008.07.003.

Schmidt, K., and Butler, J.S. (2013). Nuclear RNA surveillance: role of TRAMP in controlling exosome specificity. Wiley Interdiscip Rev RNA *4*, 217-231. 10.1002/wrna.1155.

Szabo, L., Morey, R., Palpant, N.J., Wang, P.L., Afari, N., Jiang, C., Parast, M.M., Murry, C.E., Laurent, L.C., and Salzman, J. (2015). Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol *16*, 126. 10.1186/s13059-015-0690-5.

Taggart, A.J., Lin, C.L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Res *27*, 639-649. 10.1101/gr.202820.115.

Talhouarne, G.J.S., and Gall, J.G. (2018). Lariat intronic RNAs in the cytoplasm of vertebrate cells. Proc Natl Acad Sci U S A *115*, E7970-E7977. 10.1073/pnas.1808816115.

Vader, A., Johansen, S., and Nielsen, H. (2002). The group I-like ribozyme DiGIR1 mediates alternative processing of pre-rRNA transcripts in Didymium iridis. Eur J Biochem *269*, 5804-5812. 10.1046/j.1432-1033.2002.03283.x.

Vader, A., Nielsen, H., and Johansen, S. (1999). In vivo expression of the nucleolar group I intron-encoded I-dirI homing endonuclease involves the removal of a spliceosomal intron. EMBO J *18*, 1003-1013. 10.1093/emboj/18.4.1003.

Vicens, Q., and Cech, T.R. (2009). A natural ribozyme with 3',5' RNA ligase activity. Nat Chem Biol *5*, 97-99. 10.1038/nchembio.136.

Yuan, T., Stumpf, F.M., Schlor, L.A., Schmidt, O.P., Huber, L.B., Frese, M., Höllmüller, E., Scheffner, M., Stengel, F., Diederichs, K., and Marx, A. (2022). A human RNA ligase that operates via auto- and RNA-AMPylation. bioRxiv.

Zaug, A.J., Grabowski, P.J., and Cech, T.R. (1983). Autocatalytic cyclization of an excised intervening sequence RNA is a cleavage-ligation reaction. Nature *301*, 578-583. 10.1038/301578a0.

Zhang, Y., Zhang, X.O., Chen, T., Xiang, J.F., Yin, Q.F., Xing, Y.H., Zhu, S., Yang, L., and Chen, L.L. (2013). Circular intronic long noncoding RNAs. Mol Cell *51*, 792-806. 10.1016/j.molcel.2013.08.017.

Chapter 4. MODULATION OF RNA SPLICING ENHANCES RESPONSE TO BCL2 INHIBITION IN LEUKEMIA

Eric Wang*, **Jose Mario Bello Pineda*,** Won Jun Kim*, Sisi Chen, Jessie Bourcier, Maximilian Stahl, Simon J. Hogg, Jan Phillipp Bewersdorf, Cuijuan Han, Michael E. Singer, Daniel Cui, Caroline E. Erickson, Steven M. Tittley, Alexander V. Penson, Katherine Knorr, Robert F. Stanley, Jahan Rahman, Gnana Krishnamoorthy, James A. Fagin, Emily Creger, Elizabeth McMillan, Chi-Ching Mak, Matthew Jarvis, Carine Bossard, Darrin M. Beaupre, Robert K. Bradley, Omar Abdel-Wahab

*These authors contributed equally

As of November 2, 2022, a version of this chapter is in press in the journal *Cancer Cell*.

4.1 ABSTRACT

Therapy resistance is a major challenge in the treatment of cancer. Here, we performed CRISPR/Cas9 screens across a broad range of therapies used in acute myeloid leukemia to identify genomic determinants of drug response. Our screens uncovered a selective dependency on RNA splicing factors whose loss preferentially enhanced response to the *BCL2* inhibitor venetoclax. Loss of the splicing factor *RBM10* augmented response to venetoclax in leukemia yet was completely dispensable for normal hematopoiesis. Combined RBM10 and *BCL2* inhibition led to mis-splicing and inactivation of the inhibitor of apoptosis *XIAP* and downregulation of *BCL2A1*, an anti-apoptotic protein implicated in venetoclax resistance. A novel inhibitor of splicing kinase families CLKs and DYRKs led to aberrant splicing of key splicing and apoptotic factors that synergized with venetoclax and overcame resistance to *BCL2* inhibition. Our findings underscore the importance of splicing in modulating response to therapies and provide a strategy to improve venetoclax-based treatments.

4.2 INTRODUCTION

Acute myeloid leukemia (AML) is an aggressive hematologic malignancy marked by a dismal prognosis (Ferrara and Schiffer, 2013). For decades, the standard therapy for newly diagnosed AML has been intensive cytotoxic chemotherapy. Recently, new targeted therapies have been approved for AML, including inhibitors of *IDH1/2*, *FLT3*, and *BCL2* (Short et al., 2020). Despite the introduction of these novel agents, most patients ultimately relapse and acquire resistance to long-term, continuous drug exposure (Breems et al., 2005; Ganzel et al., 2018). Genetic mutations, such as in *TP53*, have been shown to contribute to poor prognosis in patients treated with chemotherapy or the *BCL2* inhibitor venetoclax (Nechiporuk et al., 2019; Zhang et al., 2020; Zuber et al., 2009). More recently, acquired *BAX* mutations have been shown to confer resistance to venetoclax in a subset of AML patients (Blombery et al., 2022). However, in the majority of cases, genetic lesions are not known to be the main underlying mechanism of AML relapse (Kandoth et al., 2013; Li et al., 2016), possibly implicating non-genetic mechanisms that allow persistent survival of leukemia cells upon exposure to drug therapy (Fennell et al., 2021). For instance, upregulation of anti-apoptotic proteins (Konopleva et al., 2016; Minn et al., 1995) and dysregulated mitochondrial metabolism (Chen et al., 2019; Jones et al., 2018; Jones et al., 2020) can alter responsiveness to venetoclax. Such findings have demonstrated that epigenetic plasticity and transcriptional variability can act as critical evolutionary drivers of clonal fitness and drug resistance in leukemia (Fong et al., 2015; Rathert et al., 2015).

The use of combinatorial therapies has been widely used to circumvent acquired drug resistance and is an approach with proven clinical efficacy for the treatment of several cancer types (Baselga et al., 2012; Rini et al., 2019). In AML, the combination of venetoclax with hypomethylating agents is now widely used and has significantly improved the response and

survival rates of patients (DiNardo et al., 2020; DiNardo et al., 2018). However, despite the success of venetoclax and hypomethylating agent combination therapy, this regimen is not curative. Furthermore, the majority of patients are unable to undergo curative allogeneic stem cell transplantation and ultimately become resistant to therapy (DiNardo *et al.*, 2020). As such, identifying and targeting drug resistance mechanisms in AML with new combinatorial treatment regimens is of critical importance.

Here, we utilized unbiased genetic screens to map drug/gene interactions for a variety of clinically approved therapies used in the treatment of AML. This effort highlighted a unique genetic relationship between response to venetoclax and the function of specific RNA splicing factors. While there is a well-established role for RNA splicing in the regulation of apoptosis (Schwerk and Schulze-Osthoff, 2005), clinically viable means to manipulate splicing to enhance cell death in cancer have been limited to date. As genetic proof of concept, we identified a number of splicing factors whose loss promotes cell death in the setting of venetoclax and are dispensable for normal hematopoiesis, suggesting a therapeutic index for augmenting venetoclax response by modulating RNA splicing. Moreover, we present a novel compound to modulate RNA splicing and enhance venetoclax response via inhibition of the splicing kinase families known as CLKs (CDC-like kinases) and DYRKs (dual-specificity tyrosine-regulated kinases).

4.3 RESULTS

4.3.1 *Mapping genomic determinants of AML drug response*

To explore drug-gene interactions that underpin response to AML therapies, we used a genome-wide library containing 77,441 single guide RNA (sgRNA) targeting 19,115 genes (Sanson et al., 2018). We transduced this library into the human AML cell line, MOLM-13 (an MLL-AF9 translocated cell line bearing a concomitant *FLT3*^{ITD} mutation) and after 8 days post-

transduction, cells were treated with a broad range of clinically approved AML drugs (venetoclax, 5-azacytidine, cytarabine, etoposide, midostaurin, and idarubicin) (**Fig. 1A**). Changes in sgRNA abundance were assessed at day 20 post-transduction by measuring the average fold change (drug/DMSO) of all sgRNAs targeting a given gene and top scoring candidates were classified as genes that sensitize (negative CRISPR score) or confer resistance (positive CRISPR score) to individual drugs.

We identified previously characterized genes shown to mediate resistance to these compounds, including sgRNAs targeting the pro-apoptotic factors, *BAX* and *PMAIP* (also known as *NOXA*), as well as *TP53* to confer venetoclax resistance (Nechiporuk *et al.*, 2019) (**Fig. 1B**). We also confirmed that inactivation of TOP2A, a target of etoposide, promoted survival of AML cells against etoposide exposure. Of note, sgRNAs targeting the uridine-cytidine kinase *UCK2* scored as the top positive hit in our 5-azacytidine screen and UCK has been previously implicated to confer resistance to hypomethylating agent (Gu et al., 2021; Sripayap et al., 2014).

To identify new combinatorial strategies that enhance existing AML therapies, we explored genes whose sgRNAs were significantly depleted upon drug exposure. We performed Gene Ontology (GO) enrichment analysis on the top scoring negative hits from each CRISPR screen and uncovered significant terms associated with RNA splicing and regulation of mRNAs, linked to venetoclax sensitization (**Fig. 1C**). Consistently, we observed a significantly wider distribution (higher variance) of CRISPR scores for sgRNAs targeting RNA processing genes in the setting of venetoclax treatment compared to other drugs (**Fig. S1A**). These data suggest a unique relationship between perturbation of RNA processing and response to venetoclax compared to other commonly used AML therapies. Previous reports have indicated the importance of leukemia cells exploiting alternative splicing and post-transcriptional mechanisms to promote tumor growth and therapy

resistance (Han et al., 2022; Wang et al., 2019; Wang et al., 2021; Witkowski et al., 2022; Zhou et al., 2020). Moreover, clinical observations in AML patients have also demonstrated correlations between spliceosome mutations and alterations in response to venetoclax (Lachowiez et al., 2021; Zhang et al., 2020).

To further investigate the functional impact of RNA splicing factors in modulating drug response, we applied a previously developed CRISPR library targeting functional domains of 492 RNA-binding proteins (RBPs) consisting of 2,855 sgRNAs (Wang *et al.*, 2019) to enhance CRISPR/Cas9 negative selection by targeting functional protein domains (Shi et al., 2015) (**Fig. 1D**). Consistent with our initial findings from the genome-wide screen, we identified that loss-of-function of several RNA splicing factors enhanced sensitivity or resistance to venetoclax treatment (**Fig. 1E-F**). We further validated the top scoring sensitizers such as *RBM10*, *SRSF11*, *SRSF8*, *HNRNPD*, *HNRNPAB*, and *HNRNPF* whose inactivation led to preferential sensitivity in AML cells treated with venetoclax, which was not seen with other tested therapeutics (**Fig. 1G** and **Fig. S1B-D**).

4.3.2 *Loss of RBM10 sensitizes leukemia cells to venetoclax*

Among the top gene candidates whose loss sensitized cells to venetoclax was *RBM10*, whose loss-of-function exclusively enhanced venetoclax efficacy in AML amongst other drugs screened (**Fig. 1F-G** and **Fig. S2A**). We further explored publicly available genome-wide CRISPR screens performed in a broad range of human cancer cell lines which revealed that RBM10 loss is least essential in leukemia cell lines compared other cancer subtypes (**Fig. S2B**). However, in the presence of venetoclax, *RBM10* deletion strikingly conferred preferential lethality and anti-leukemic effects in human AML cell lines, across a variety of molecular subtypes (**Fig. 2A-C** and **Fig. S2C**). Of note, *RBM10* deletion even augmented *BCL2* inhibition in *TP53*-mutated AML cell

lines (THP-1, and U937) (Sugimoto et al., 1992), which have been previously described as venetoclax resistant (Nechiporuk *et al.*, 2019; Zhang *et al.*, 2020) (**Fig. S2D**).

We next assessed the impact of *RBM10* deletion on the response of human AML cells to venetoclax *in vivo*. To achieve this, we transplanted MOLM-13 cells stably expressing firefly luciferase and anti-RBM10 sgRNAs or the non-targeting control (sgRosa) into (NOD)/severe combined immunodeficiency (SCID) IL2Rgamma^{null} (NSG) mice. Upon disease onset, mice were treated with venetoclax (100 mg/kg/day) or vehicle control (**Fig. 2D**). Consistent with our *in vitro* findings, *RBM10* deletion reduced leukemia burden and extended survival in the setting of venetoclax treatment (**Fig. 2E-F** and **Fig. S2E**). Indel analysis of prolonged *RBM10* sgRNA editing by next-generation sequencing showed an outgrowth of cells containing in-frame *RBM10* mutations, implicating that mice succumb to an outgrowth of sgRNA-expressing cells that retain RBM10 functionality (**Fig. S2F**). Overall, these findings provide genetic evidence that loss of RBM10 has a synthetic lethal relationship with *BCL2* inhibition in AML.

Many RNA splicing factors are known to be pan-essential for cell survival (Hart et al., 2015). To evaluate the therapeutic potential of RBM10 modulation as a therapeutic candidate for venetoclax-based therapies, we generated an Rbm10 conditional knockout (cKO) mouse by inserting *lox*P sites flanking exon 3 of *Rbm10* (*Rbm10*$^{fl/fl}$; **Fig. 2G**) and crossing with interferon-induced *Mx1*-driven Cre recombinase mice. Following intraperitoneal polyinosinic;polycytidylic acid (pIpC) injections, *Rbm10* cKO mice were confirmed to excise exon 3 of *Rbm10* leading to an early frameshift and loss of Rbm10 protein in bone marrow cells (**Fig. 2H** and **Fig. S2G-I**). We next assessed stem cell functionality using *in vitro* colony-replating assays which demonstrated that *Rbm10* deletion in hematopoietic precursors did not impair colony formation (**Fig. 2I**). In parallel, bone marrow-derived cells from CD45.2^{+} *Rbm10* floxed mice were transplanted in a

competitive manner along with competitor *Rbm10* wild-type CD45.1^{+} cells and treated with pIpC after stable reconstitution of hematopoiesis. There was no significant effect of Rbm10 deletion on absolute numbers or frequency of peripheral blood and bone marrow cells (**Fig. 2J-K** and **Figure S2J-K**). These data demonstrate that Rbm10 is dispensable for normal hematopoiesis.

4.3.3 *Dual inhibition of RBM10 and BCL2 promotes XIAP mis-splicing*

We next sought to understand the mechanistic basis for the relationship between RBM10 loss and enhanced response to venetoclax. The effects of *RBM10* KO on venetoclax response were rescued by expressing an *RBM10* cDNA impervious to anti-*RBM10* sgRNAs (due to mismatches between cDNA sequence and the *RBM10* sgRNAs; **Fig. 3A**). However, expression of *RBM10* lacking its second RNA recognition motif 2 (RRM2) or C2H2-type zinc finger (C2H2 ZnF) (Collins et al., 2017) failed to rescue response to venetoclax (**Fig. 3A**).

The above data indicate the importance of *RBM10*'s RNA binding domains on venetoclax response. We therefore further assessed the direct impact of *RBM10*-RNA interactions on pre-mRNA binding and splicing, which have not been explored in hematopoietic cells previously. We performed anti-*RBM10* enhanced UV cross-linking immunoprecipitation (eCLIP) (Van Nostrand et al., 2016) in MOLM-13 AML cells (**Fig. S3A**). This approach identified approximately 29,000 significant sequence clusters bound by *RBM10*, which corresponded to ~5,000 annotated transcripts. Approximately 90% of *RBM10* binding sites mapped to intronic sites, with a preferential occupancy of distal (further than 500 nucleotides (nt) from the splice site region) (77.1%) and proximal (within 500 nt of splice site region) intronic (8%) sequences near 5' and 3' splice sites throughout the transcriptome (**Fig. 3B** and **Fig. S3B**).

Next, we evaluated the transcriptional and splicing changes in *RBM10*-deleted AML cells treated with venetoclax or DMSO, compared to non-targeting sgRosa, by RNA sequencing (RNA-

seq). We measured isoform usage frequencies across seven main types of alternative splicing events [skipped (or retained) cassette exons (SE), alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), mutually exclusive exons (MXE), tandem 3' UTRs (TUTR), and retained (RI) and constitutive introns (CI)] to quantify splicing changes across treatments (**Fig. S3C-D**). *RBM10* KO primarily led to changes in cassette exon splicing (**Fig. 3C**), suggesting that *RBM10* most commonly regulates exon usage in AML cells. In comparison, RBM10 deletion in the presence of venetoclax amplified the degree of aberrant splicing involving constitutive introns and cassette exons. Most notably, we observed an increase in exon exclusion events in the combination treatment versus *RBM10* deletion alone (n=342) (**Fig. 3D**).

We further investigated the link between *RBM10* binding and differential splicing observed in combined *RBM10* KO and venetoclax. These analyses revealed *RBM10* binding signal in the 5′ region of the upstream intron of repressed cassette exons following combination treatment, suggestive of a role of *RBM10* binding in this region in promoting exon exclusion (**Fig. 3E**). Interestingly, we found that the inhibitor of apoptosis protein (IAP) family member, *XIAP*, displayed increased exclusion of the first coding exon in venetoclax-treated *RBM10* KO AML cells, which also had significant *RBM10* binding at this region (**Fig. 3F** and **Fig. S3E**). *XIAP*, also known as *BIRC4*, binds and sequesters pro-apoptotic caspases through direct protein-protein interactions with its BIR domains to prevent caspase homodimerization thereby inactivating apoptosis (Huang et al., 2001; Riedl et al., 2001; Shiozaki et al., 2003; Srinivasula et al., 2001). Based on our findings, we hypothesize that activation of apoptosis is a consequence of skipping the first coding exon of *XIAP*. The resulting mRNA lacks *XIAP*'s canonical start codon as well as the sequence encoding the majority of its *BIR1-3* domains, strongly suggesting that this splicing change results in loss of functional *XIAP* production (**Fig. 3G**). We also functionally evaluated the

mis-spliced isoform of *XIAP* event induced by *RBM10* KO and venetoclax treatment (which we refer to as XIAP Dexon 1) by ectopically expressing full-length *XIAP* (FL) or *XIAP* Dexon 1 linked to a GFP reporter in MOLM-13 cells (**Fig. 3H**). Consistent with the function of IAP proteins, we found *XIAP* FL overexpression allowed survival of AML cells after venetoclax treatment, whereas *XIAP* Dexon 1 resulted in increased apoptosis (**Fig. 3I-J** and **Fig. S3F**). Overall, these results demonstrate that *XIAP* Dexon 1 cannot rescue cell death induced by venetoclax treatment and *RBM10* deletion (**Fig. S3G**). Importantly, prior work has demonstrated that inhibition of *XIAP* synergized with venetoclax (Hashimoto et al., 2021), highlighting the importance of *XIAP* levels in *BCL2* inhibitor sensitivity.

Gene expression analysis of venetoclax-treated *RBM10* KO AML cells revealed downregulated expression of *BCL2A1,* which encodes an anti-apoptotic factor whose expression is correlated with venetoclax resistance in AML patients (Zhang *et al.*, 2020) (**Fig. 3K** and **Fig. S3H-I**). Consistent with these data, overexpression of *BCL2A1* cDNA was able to fully rescue the anti-leukemic effects seen with the combined loss of *RBM10* and *BCL2* inhibition (**Fig. 3L**). Moreover, we did not observe significant *RBM10* eCLIP peaks or splicing alteration of *BCL2A1* mRNA which suggests that upstream factors may regulate *BCL2A1* transcript. Overall, these data provide mechanistic evidence that the combined loss of *RBM10* and BCL2 leads to altered splicing and expression of mRNAs encoding key apoptotic genes.

4.3.4 *Pharmacologic inhibition of splicing kinases synergizes with venetoclax*

Utilizing our CRISPR screens to identify pharmacologically intervenable splicing factors to augment venetoclax response, we found that inactivation of several serine/arginine (SR)-rich proteins (*SRSF2*, *SRSF3*, *SRSF8*, and *SRSF11*) sensitized AML cells to venetoclax (**Fig. 1D**). The family of SR splicing factors are essential for alternative pre-mRNA splicing and their activity is

tightly regulated by post-translational modifications placed by serine/threonine kinases (Aubol et al., 2016; Colwill et al., 1996; Gui et al., 1994). For example, CLKs phosphorylate Arginine-Serine (RS) domains in SR proteins and regulate pre-mRNA splicing (Aubol et al., 2016; Prasad et al., 1999). Moreover, *DYRK1A* has been reported to regulate alternative splicing via phosphorylation of *SF3B1* (de Graaf et al., 2006; Qian et al., 2011; Shi et al., 2008). In addition, analysis of publicly available genome-wide CRISPR screens from DepMap (Meyers et al., 2017) revealed *BCL2* as one of the top co-dependencies with *DYRK1A* loss (**Fig. S4A**).

These findings support the rationale to inhibit splicing-dependent kinases as a combinatorial strategy with venetoclax treatment. To pursue therapeutic inhibition of splicing-dependent kinases, SM09419, a pan-CLK pan-DYRK inhibitor was developed via rational design and iterative medicinal chemistry to achieve drug-like and favorable pharmacokinetic profiles (**Fig. 4A-D**). We confirmed the selectivity of SM09419 to target CLK kinases *CLK1-4* as well as *DYRK1A-B* and *DYRK2* using in cell NanoBRET target engagement assays (**Fig. 4C** and **Fig. S4B**). Accordingly, SM09419 treatment resulted in dose-dependent reduction of CLK activity and SR protein phosphorylation in AML cells (**Fig. 4E** and **Fig. S4C**). Next, we assessed the combinatorial effects of SM09419 with a panel of drugs (venetoclax, 5-azacytidine, cytarabine, and midostaurin) in human AML cell lines. We observed a synergistic effect exclusively when combining SM09419 and venetoclax in MOLM-13 parental and venetoclax-resistant cells but not with other drugs (**Fig. 4F-I** and **Fig. S4D-E**). Despite robust anti-leukemic effects of SM09419 *in vitro*, SM09419 (25 mg/kg) treatment in wild-type C57BL/6 mice was well tolerated *in vivo* with no signs of hematologic toxicities (based on serial blood counts, in vitro hematopoietic progenitor cell assays, and detailed analysis of hematopoietic cell composition in blood and bone marrow) or

liver or kidney dysfunction, thus providing a rationale for pharmacologic inhibition of CLK/DYRK in combination with venetoclax in AML (**Fig. S5A-H**).

To understand the mechanistic basis for the synergy of SM09419 and venetoclax combination, we performed RNA-seq on MOLM-13 human AML cells treated with SM09419 alone or in combination with venetoclax. Splicing analyses showed that SM09419 alone, or in combination with venetoclax, mainly resulted in changes in the processing of constitutive/retained introns and cassette exons (**Fig. 5A**). CLK/DYRK inhibition affects cassette exon recognition in a sequence-specific manner, as evidenced by the enrichment of pyrimidines in exons preferentially excluded upon SM09419 treatment (**Fig. 5B**). While venetoclax monotherapy had no significant effects on RNA splicing, treatment with SM09419 or the combination resulted in striking reductions in RNA splicing efficiency as manifested by cassette exon skipping and intron retention (**Fig. 5A, Fig. 5C**). Of note, these splicing shifts resulted in substantial increases in levels of mRNAs that contain premature termination codons and are therefore predicted substrates for degradation by nonsense-mediated decay (NMD).

In order to understand how the effects of SM09419 relates to deletion of RBM10, we next performed a systematic comparison of the splicing changes and gene expression across both conditions in the same MOLM-13 cells. Both *RBM10* deletion and SM09419 treatment cause splicing changes which promote nonsense-mediated decay (NMD)-inducing transcripts. However, the magnitude of NMD-inducing splicing events is greater with SM09419 treatment (a result consistent with the fact that *RBM10* deletion in the absence of any drug treatment is well-tolerated in MOLM-13 cells) (**Fig. 5D**). Nonetheless, a number of mRNA isoforms were shared across *RBM10* deletion versus SM09419 treatment in the absence and presence of concomitant venetoclax treatment. Interestingly, one concordant effect was the same mis-splicing event in *XIAP*

seen with *RBM10* deletion (**Fig. 5E-F**). Finally, both RBM10 deletion and SM09419 treatment prominently downregulate *TNFAIP3* (also known as A20) (**Figure S6A**). *TNFAIP3* is a well described regulator of NF-kB signaling (Duwel et al., 2009) and its inhibition may explain the *BCL2A1* down-regulation in *RBM10* KO cells exposed to venetoclax. Transcriptomic analysis of SM09419-treated AML cells also demonstrated downregulation of *MYB* and *MYC* mRNA levels, which are essential oncogenic factors in AML (**Fig. S6B**) (Zuber et al., 2011a; Zuber et al., 2011b).

4.3.5 *SM09419 induces splicing alterations of key survival genes in AML*

Further characterization of SM09419-associated splicing changes revealed increased intron retention within the transcripts of a number of RNA splicing factors (*SRSF5*, *U2AF2*, *RBM17*, and *RBM5*) which also led to decreased protein expression in two independent human AML cell lines (**Fig. 5G-H** and **Fig. S6C-E**). Interestingly, several of these same splicing factors were also identified by our CRISPR screens as genes whose inactivation enhanced venetoclax efficacy and therefore explains the synergistic effects when combining SM09419 and venetoclax (**Fig. 5E** and **Fig. 5I**). Furthermore, we found that SM09419-treated AML cells led to downregulation of several key apoptotic proteins, such as *MCL-1*, which have been shown to be upregulated in hematologic neoplasms and confers resistance to *BCL2* inhibitors (Konopleva *et al.*, 2016) (**Fig. 5H** and **Fig. S6E**). Finally, SM09419 treatment promoted inclusion of an exon with an in-frame stop codon (a "poison exon," whose inclusion renders the transcript NMD-sensitive) in the receptor tyrosine kinase *FLT3* (**Fig. 5J**). As such, there was reduced *FLT3* mRNA and protein expression in SM09419 treated cells (which is especially pertinent given the known dependence of this *FLT3* mutant AML cell line on *FLT3* expression) (**Fig. 5H** and **Fig. S6E**). Importantly, the above results on the impact of SM09419 treatment on *XIAP*, *FLT3*, and *MCL-1* levels were confirmed in an additional AML cell line. Both venetoclax resistant and sensitive KG-1a cells were similarly

susceptible to SM09419 treatment and experienced comparable dose-dependent reductions in *XIAP*, *FLT3*, and *MCL-1* (**Fig. S6D-E**).

Additional predicted NMD-inducing splicing events upon SM09419 treatment with commensurate reduction in mRNA expression in *SMYD2* (a lysine methyltransferase recognized as a therapeutic target in AML (Zuber *et al.*, 2011a), *DHODH* (a metabolic enzyme and recent AML therapeutic target) (Sykes et al., 2016), *ATAD3A* (a metabolic enzyme whose expression has been included in leukemia stem cell signatures) (Kim et al., 2010), the *MYC* target gene *CDC16* (Somervaille et al., 2009), and the additional RNA processing genes *SRPK3, TRA2A,* and *DDX51* (**Figure S6F**). Overall, these data identify that SM09419 downregulates expression of key RNA splicing factors as well as important apoptotic factors and *FLT3* via impaired splicing to enhance response to venetoclax in AML while having minimal impact on normal hematopoiesis.

4.3.6 *SM09419 overcomes venetoclax-based therapy resistance*

We next tested the efficacy of SM09419 across a spectrum of human AML cell lines. SM09419 treatment resulted in broad anti-leukemic effects with potent inhibitory activity across AML subtypes, including cell lines that were highly resistant to venetoclax treatment (**Fig. 6A**). Based on these data, we evaluated the ability of SM09419 to overcome venetoclax resistance. We developed three independent venetoclax-resistant MOLM-13 cell lines following continuous exposure to venetoclax for 3 weeks. Dose-response curves after drug selection confirmed that venetoclax-resistant cell lines displayed a high inhibitory effect concentration (IC_{50} > 99 nM) approximately six times greater than parental cells (IC_{50} = ~15 nM) (**Fig. 6B**). Whole-exome sequencing (WES) and targeted capture sequencing (MSKCC-IMPACT) did not reveal any known genomic alterations that may cause venetoclax resistance. SM09419 as single agent led to approximately equally potent inhibitory activity against venetoclax-resistant AML cells as

parental, venetoclax sensitive cells (**Fig. 6B**). Moreover, addition of venetoclax and SM09419 led to synergistic effects in venetoclax-resistant MOLM-13 cells (**Fig. 4I** and **Fig. S4D**). Consistent with our previous findings in MOLM-13 parental cells, we observed downregulation of essential apoptotic proteins (*XIAP*, *MCL-1*), splicing factors (*RBM5*, *U2AF2*), and the tyrosine kinase *FLT3* in venetoclax-resistant MOLM-13 and KG-1a cells (**Fig. 5H** and **Fig. S6E).** We further extended these findings to patient-derived xenograft (PDX) models of AML from patients with *de novo* resistance to venetoclax combination regimens (5-azacytidine or low-dose cytarabine) (**Fig. 6C-D**). Following xenotransplantation from two individual venetoclax-resistant patients into NSGS mice, we detected disease engraftment with $\geq$10% human $hCD45^+$ $hCD34^+$ $hCD38^+$ cells and exposed mice to SM09419 (25 mg/kg) or vehicle administered orally and daily for 3 weeks. SM09419 resulted in significant reduction of hCD45 AML cells in the peripheral blood and bone marrow of mice treated with SM09419 when compared to vehicle control (**Fig. 6D-F**). Moreover, *ex vivo* culturing of these AML patient samples demonstrated single-agent potency of SM09419 as well as synergistic effects when combined with venetoclax (**Fig. 6G-H**). Collectively, these findings demonstrate the *in vivo* efficacy of SM09419 to overcome resistance to venetoclax-based therapies.

4.4 DISCUSSION

Here, we performed comprehensive mapping of drug-gene interactions that dictate response to a broad range of AML therapies. This effort uncovered genetic strategies which enhance the effects of these existing AML drugs which may ultimately lead to new combinatorial strategies to improve patient outcomes. We focused on those genetic events which augment response to venetoclax given the clinical need to develop novel venetoclax-based combinatorial treatment regimens for

AML. Overall, our findings establish a functional link between splicing modulation and therapeutic efficacy of *BCL2* inhibition in AML.

Given the established role of RNA splicing in regulating the expression and function of key proteins involved in cell death signaling and apoptosis, a number of prior studies have attempted to pharmacologically perturb splicing to promote response to *BCL2* inhibition. For example, one prior study demonstrated that E7107, a potent SF3b inhibitor, can synergize with venetoclax in B-cell malignancies and solid tumors (Aird et al., 2019; Ten Hacken et al., 2018). However, toxicities associated with E7107 have led to its suspension from clinical use (Eskens et al., 2013; Hong et al., 2014), and the clinical efficacy of more recent SF3b inhibitors (such as the drug H3B-8800) (Seiler et al., 2018) for high-risk myeloid neoplasms remain unclear. Here, we provide rationale for a new clinical modality to modulate RNA splicing through pharmacological inhibition of splicing-dependent kinases. Specifically, our data suggest that inhibition of CLKs and DYRKs in combination with venetoclax or as a single agent represent a new therapeutic strategy to circumvent resistance to venetoclax.

CLK and DYRK kinases are two highly related families of kinases within the CMGC (cyclin-dependent kinase [CDK], mitogen-activated protein kinase [MAPK], glycogen synthase kinase [GSK3], CLK) group of the eukaryotic kinome. Each CLK and DYRK are dual specificity kinases which phosphorylate serine/threonine and tyrosine residues (Lindberg and Meijer, 2021; Martin Moyano et al., 2020). A variety of structurally diverse CLK inhibitors have been developed and most of these also inhibit DYRK kinases to varying degrees (Martin Moyano et al., 2020). Despite the fact that CLKs and DYRKs perform multi-site phosphorylation of a number of substrates, it is clear that perturbing the activity of *CLK1-4* or *DYRK1A/B* globally impacts RNA splicing via altering splicing factor protein phosphorylation. *DYRK1A* resides in nuclear speckles

and overexpression of *DYRK1A* induces redistribution of SR proteins from nuclear speckles to active sites of transcription/splicing in a manner that depends on its kinase activity (Alvarez et al., 2003). Similarly, *CLK1* has been shown to regulate the cellular localization and splicing impact of SRSF1 via phosphorylation of Serine-Proline dipeptides at multiple sites on *SRSF1* (Aubol et al., 2013).

Interestingly, prior data have identified that dephosphorylation of RNA splicing factors occurs during apoptosis and that the ensuing change in splicing may be necessary for cells to execute apoptosis (Kamachi et al., 2002). For example, SR proteins are targets for a number of apoptosis agonists and splicing factor kinases are inactivated during cell death by caspase-mediated proteolysis (Kamachi et al., 2002). This includes caspases 8, 9, and 3/6 cleaving *SRPK1*, *SPRK2*, and topoisomerase respectively, thereby altering splicing during apoptosis. Moreover, *FAS* activation results in dephosphorylation of SR proteins via induction of *PP1* phosphatase (Chalfant et al., 2001). Thus, pharmacologic inhibition of phosphorylation of splicing factors may enhance response to venetoclax by mimicking the impact of apoptosis signaling cascade on RNA splicing.

SM09419 has pharmacologic properties that are very similar to Cirtuvivint, one of the first CLK/DYRK ATP-competitive inhibitors that has entered first-in-human and phase 1b clinical trials in patients with advanced solid tumors (NCT03355066 and NCT05084859) (Scott et al., 2022; Tolcher et al., 2021). In the Cirtuvivint first-in-human study, pharmacodynamic evidence for proof of mechanism in human whole blood was reported at well tolerated doses. Importantly, infrequent grade 3 adverse hematologic events have been observed with a grade 3 anemia rate of <15% and even lower frequency for neutropenia or thrombocytopenia. In the dose escalation portion of this trial and in the combination study, Cirtuvivint has shown early evidence of anti-

tumor activity with declines in PSA in prostate cancer subjects, tumor shrinkage in several tumor types, and prolonged stable disease (treatment reaching cycle 6 and beyond). Similarly, in the phase I trial of the pan-CLK inhibitor CTX-712 that evaluated subjects with solid tumors and hematologic malignancies, two complete remissions in refractory AML patients along with two partial responses in ovarian cancers subjects were reported with single agent therapy (Shimizu et al., 2022). While it is too early to make conclusions about efficacy from these early data, hematologic recovery is required for CR in AML which indicates that CLK/DYRK inhibition is feasible with manageable hematologic toxicity. Using the SM09419 compound herein, we provide preclinical data demonstrating the utility of CLK/DYRK inhibition as a single agent to overcome resistance to venetoclax-based therapies.

Beyond chemical modulation of RNA splicing, our genetic studies also highlighted a number of additional therapeutic targets to rationally enhance response to venetoclax and/or overcome venetoclax resistance. For instance, we demonstrated that loss of RBM10 enhances efficacy of venetoclax. *RBM10* is a known splicing factor that promotes exon inclusion (Wang et al., 2013). While loss-of-function *RBM10* mutations have been described in certain solid tumors (Cancer Genome Atlas Research, 2014; Giannakis et al., 2016; Witkiewicz et al., 2015) and in the genetic disease TARP syndrome (Gripp et al., 2011), we found that loss of *Rbm10* does not alter hematopoiesis. These data suggest context-specific roles for *RBM10* and nominate *RBM10* as a therapeutic vulnerability in combination with *BCL2* inhibitors.

Importantly, loss of *RBM10* was associated with reduced expression of the anti-apoptotic *BCL2* homologue *BCL2A1* as well as alternative splicing of *XIAP*, the most well-characterized IAP protein. Amongst *BCL2* family members, expression of *BCL2A1* has been most consistently associated with venetoclax resistance in a variety of leukemias. Upregulation of *BCL2A1* has been

reported to be associated with resistance to venetoclax in both the BeatAML as well as Leucegene cohorts of AML patients as well as AML preclinical models (Bisaillon et al., 2020; Zhang et al., 2020). In addition, *BCL2A1* mRNA is heavily expressed (>10-fold more than *BCL2*) in monocytes, which is thought to at least partially explain the relationship between monocytic leukemia differentiation and impaired response to venetoclax (Zhang et al., 2020). Consistent with these findings, genetic downregulation of *BCL2A1* restores venetoclax sensitivity in AML models with venetoclax resistance (Zhang et al., 2020). While there is no clinical means to chemically inhibit *RBM10*, our data underscores the need to develop potent small molecule inhibitors or peptide aptamers of *BCL2A1* in the treatment of AML.

Lastly, our data demonstrates that co-targeting of *RBM10* and *BCL2* or pharmacologic inhibition of CLK/DYRK converge on the mis-splicing of the IAP-antagonist protein, *XIAP* as a mechanism to enhance venetoclax efficacy. These findings suggest the potential benefit of chemical campaigns to develop mimetics of IAP-antagonist proteins to negatively regulate *XIAP* as seen with *RBM10* loss.

4.5 FIGURES

Figure 1

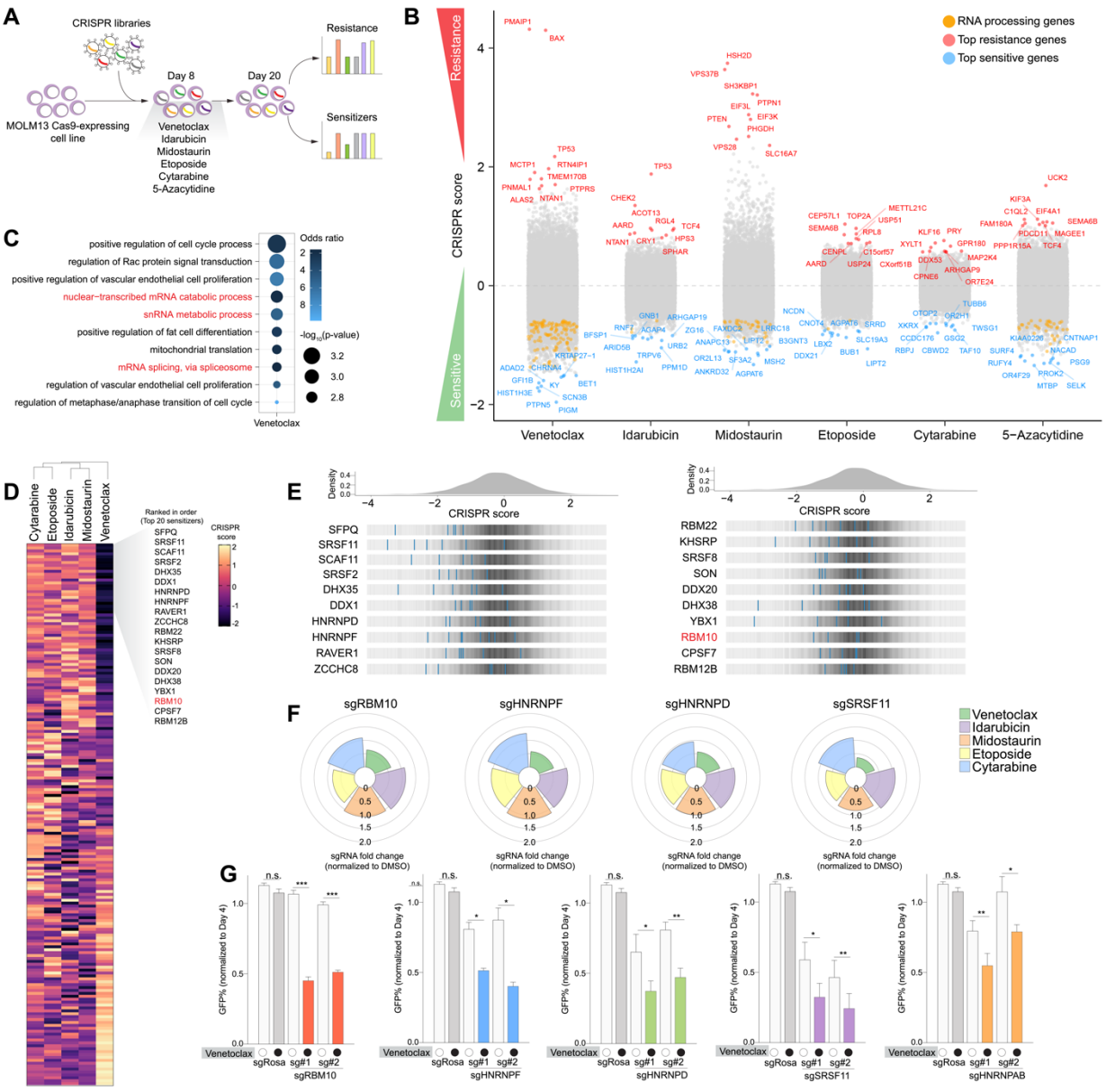


Figure 1. Mapping genomic determinants of AML drug response and synthetic lethal relationship between RNA splicing factors and venetoclax sensitivity.

(A) Schematic of genome-wide CRISPR screens in MOLM-13 AML cells treated with a panel of clinically approved AML drugs.

(B) Manhattan plot depicting top 10 genes that sensitizes (blue) or confer resistance (red) in individual CRISPR drug screens. Orange dots represent RNA processing genes. CRISPR score represents the log2 (fold-change) values of sgRNAs normalized to DMSO.
(C) Gene ontology (GO) enrichment analysis of top sensitizers in the venetoclax screen.
(D) Clustered heatmap of results of the RNA-binding protein-focused CRISPR drug screens in MOLM-13 AML cells treated with drugs. CRISPR score represents log2 fold change of sgRNAs normalized to DMSO.
(E) Histogram of CRISPR scores for all sgRNAs in the venetoclax screen in **(D)**. Values represent the log2 (fold-change) values of sgRNAs normalized to DMSO. The blue lines represent individual sgRNAs targeting the indicated genes among the top splicing factor candidates.
(F) Polar plots of top synergistic splicing factors identified in **(D)** treated with various AML drugs. The height of the wedge corresponds to the sgRNA fold change normalized to DMSO.
(G) Competition-based assay in MOLM-13 cells 10 days post-transduction with top 2 sgRNAs targeting each splicing factor or non-targeting sgRosa control (n=3, mean+SEM) treated with 50 nM venetoclax. Statistical analysis was performed using unpaired Student's t test by Prism GraphPad (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, n.s., not significant). See also **Figure S1**.

Figure 2

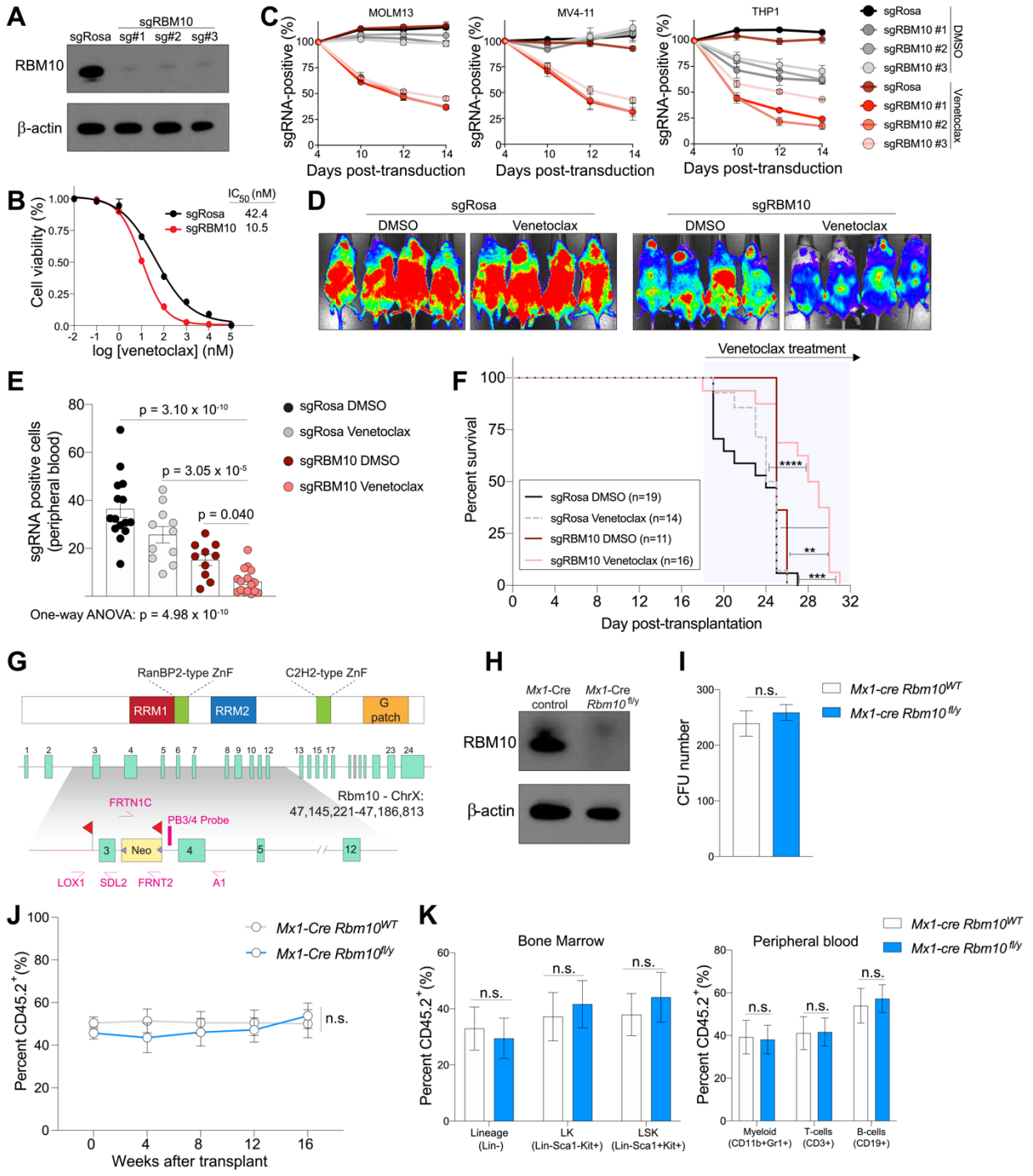


Figure 2. RBM10 loss enhances BCL2 inhibition in AML cells but is dispensable for normal hematopoiesis.

(A) Western blot of CRISPR-mediated knockout of *RBM10* in MOLM-13 cells.

(B) Dose-response curves of sgRBM10 or sgRosa treated with indicated venetoclax concentrations at 48 hours. IC50 values were calculated from technical triplicates per experiment, error bars represent SEM.
(C) Competition proliferation assays of sgRBM10 or non-targeting sgRosa in human AML cell line expressing Cas9 and treated with 50 nM of venetoclax or DMSO.
(D) Bioluminescent imaging of mice transplanted with MOLM-13 cells transduced with sgRBM10 or sgRosa and treated daily with venetoclax (100 mg/kg) or vehicle control. Representative images of 4 mice per condition is shown. Images were taken 4 days post-treatment.
(E) Flow cytometry analysis of GFP-positive sgRNA-expressing MOLM-13 cells in peripheral blood at day 6 post-treatment. Statistical analysis was performed using One-way ANOVA with post-hoc testing as indicated.
(F) Kaplan-Meier survival curves of mice transplanted with MOLM-13 cells transduced with sgRBM10 or sgRosa and treated daily with venetoclax (100 mg/kg) or vehicle. The p values were determined using a log-rank Mantel-Cox test (**$p < 0.01$, ***$p < 0.001$, n.s., not significant).
(G) Schematic depiction of the targeting strategy to generate Rbm10 cKO mice. The Rbm10 allele was deleted by targeting exon 3 that resulted in a frameshift following excision. Two LoxP sites flanking exon 3 and an Frt-flanked neomycin selection cassette were inserted in the downstream intron.
(H) Western blot of Rbm10 in bone marrow mononuclear cells from Mx1-cre Rbm10fl/y (Rbm10 cKO) or Mx1-cre control 7 days after polyinosinic-polycytidylic acid (pIpC) treatment.
(I) Total number of colony-forming units (CFU) from bone marrow cells of Mx1-cre Rbm10fl/y (Rbm10 cKO) or Mx1-cre control mice following 7 days of culture. The p values were determined by unpaired student t test. n.s., not significant.
(J) Percentage of CD45.2+ cells in peripheral blood over the course of 4 months competitive transplantation.
(K) Percentage of CD45.2+ of hematopoietic stem and progenitor cells in the bone marrow (left) and mature immune cells in the peripheral blood (right). See also **Figure S2**.

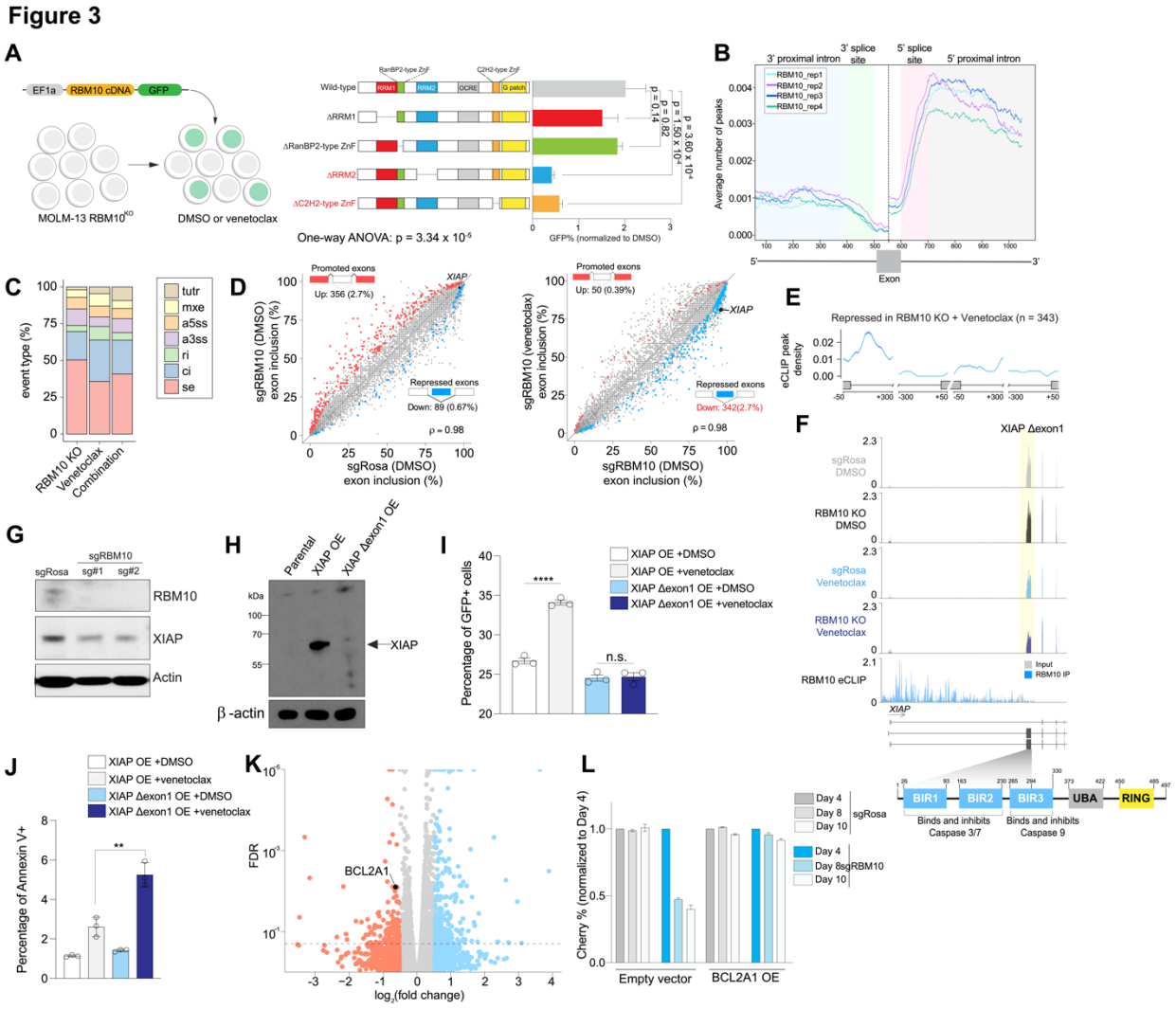


Figure 3. Impact of RBM10 on RNA binding, RNA splicing, and response to venetoclax.

(A) Competition-based assay of RBM10 KO in MOLM-13 cells and transduced with RBM10 cDNA wild-type (WT) or individual mutant (lacking RNA-binding domains) RBM10 cDNA and treated with venetoclax (50 nM) or DMSO at 48 hours. The p-values were determined by One-way ANOVA with post-hoc testing.

(B) Metaintron plots of average number of RBM10 peaks mapped to intronic regions flanking exons in MOLM-13 cells. This plot is exon-centered (500-600 bp) on the x-axis. Enhanced crosslinking and immunoprecipitation (eCLIP) was performed in 4 replicates.

(C) Percentage of treatment-responsive (*RBM10* KO, venetoclax, or *RBM10* KO and venetoclax) differentially spliced event types: cassette exons (SE), alternative 5' ss exon (A5E), alternative 3' ss exon (A3E), mutually exclusive exons (MXE), retained intron (RI), constitutive intron (CI), and tandem 3' UTR (TUTR).

(D) Scatter plot of cassette exons (SE) promoted (red circles) or repressed (blue circles) in MOLM-13 cells transduced with sgRosa or sgRBM10 treated with DMSO or venetoclax. ρ denotes Spearman's rank correlation coefficient.
(E) *RBM10* splicing map generated by integrating *RBM10* KO splicing changes from RNA-seq and *RBM10* eCLIP binding sites.
(F) RNA-seq and eCLIP (bottom) coverage plots of XIAP Δexon1 in MOLM-13 cells with *RBM10* KO or non-targeting sgRNAs treated with DMSO or venetoclax. Yellow shadow depicts exon exclusion event in *RBM10* KO venetoclax-treated MOLM-13 cells overlapped with functional protein domains of *XIAP*.
(G) Western blotting of *XIAP* after 50 nM venetoclax treatment of MOLM-13 cells with sgRosa or sgRBM10 for 48 hrs.
(H) Western blotting of *XIAP* protein levels after ectopic overexpression of *XIAP* full-length (FL) or *XIAP* Δexon1.
(I) Competition-based assay of *XIAP* full-length (FL) or *XIAP* Δexon1 linked to GFP reporter after 24 hrs of venetoclax treatment.
(J) Annexin V staining of *XIAP* full-length (FL) or XIAP Δexon1 after 24 hrs of venetoclax treatment. Statistical analysis was performed using unpaired Student's t test by Prism GraphPad (*$p < 0.05$, **$p < 0.01$, n.s., not significant).
(K) Volcano plot of differentially expressed genes (DEGs) upon RBM10KO venetoclax-treated MOLM-13 cells compared to DMSO. (L) Competition-based assay measuring Cherry-expressing sgRBM10 or sgRosa cells transduced with overexpression (OE) of BCL2A1 cDNA or empty vector GFP-positive cells in MOLM-13 cells treated with 50 nM venetoclax for 48 hours. See also **Figure S3**.

Figure 4

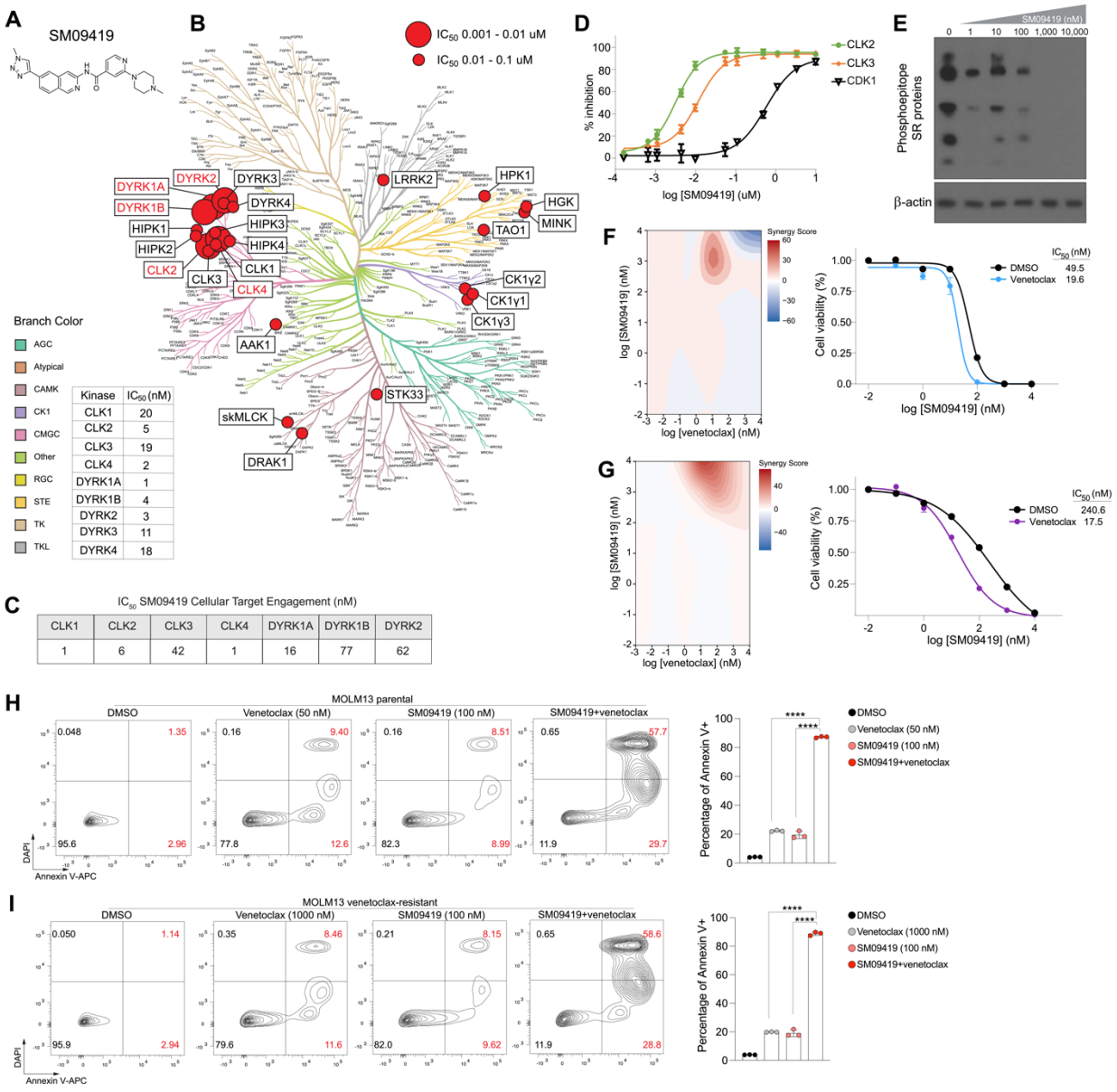


Figure 4. Pharmacologic inhibition of splicing-dependent kinases synergizes with venetoclax.
(A) Structure of SM09419 selectivity.
(B) Kinase dendrogram of SM09419. Kinases with IC50 values of 0.01 to 0.1 µM are indicated by small red circle, whereas larger red circles represent more potent IC50 values with 0.001 to 0.01 µM.
(C) NanoBRET target engagement assay of *CLK1-4*, *DYRK1A/B*, and *DYRK2* upon 24 hrs of SM09419 treatment.
(D) Inhibition of CLK kinases (*CLK2* and *CLK3*) and *CDK1* kinase. IC50 values were determined from dose response curves.

(E) Western blot of phosphorylated SR proteins treated with increasing concentration of SM09419 for 48 hrs in MOLM-13 cells.
(F) 2D synergy plots using Zero interaction potency (ZIP) model (left) and dose-response curves (right) of SM09419 and venetoclax combination at various concentration treated for 48 hrs in MOLM-13 and **(G)** KG-1 cells (n=3, mean+SEM). The presence of synergy was determined using the SynergyFinder computational package and the ZIP synergy index in which red signifies synergism and blue is antagonism. A positive synergy score is the percent more cell death than expected. IC50 values were calculated from technical triplicates per experiment.
(H) Annexin V staining (left) and quantification (right) of MOLM-13 parental and **(I)** venetoclax-resistant cell lines treated with SM09419, venetoclax, or the combination at 48 hrs post-treatment. Statistical analysis was performed using unpaired Student's t test by Prism GraphPad (****$p < 0.0001$). See also **Figure S4-5**.

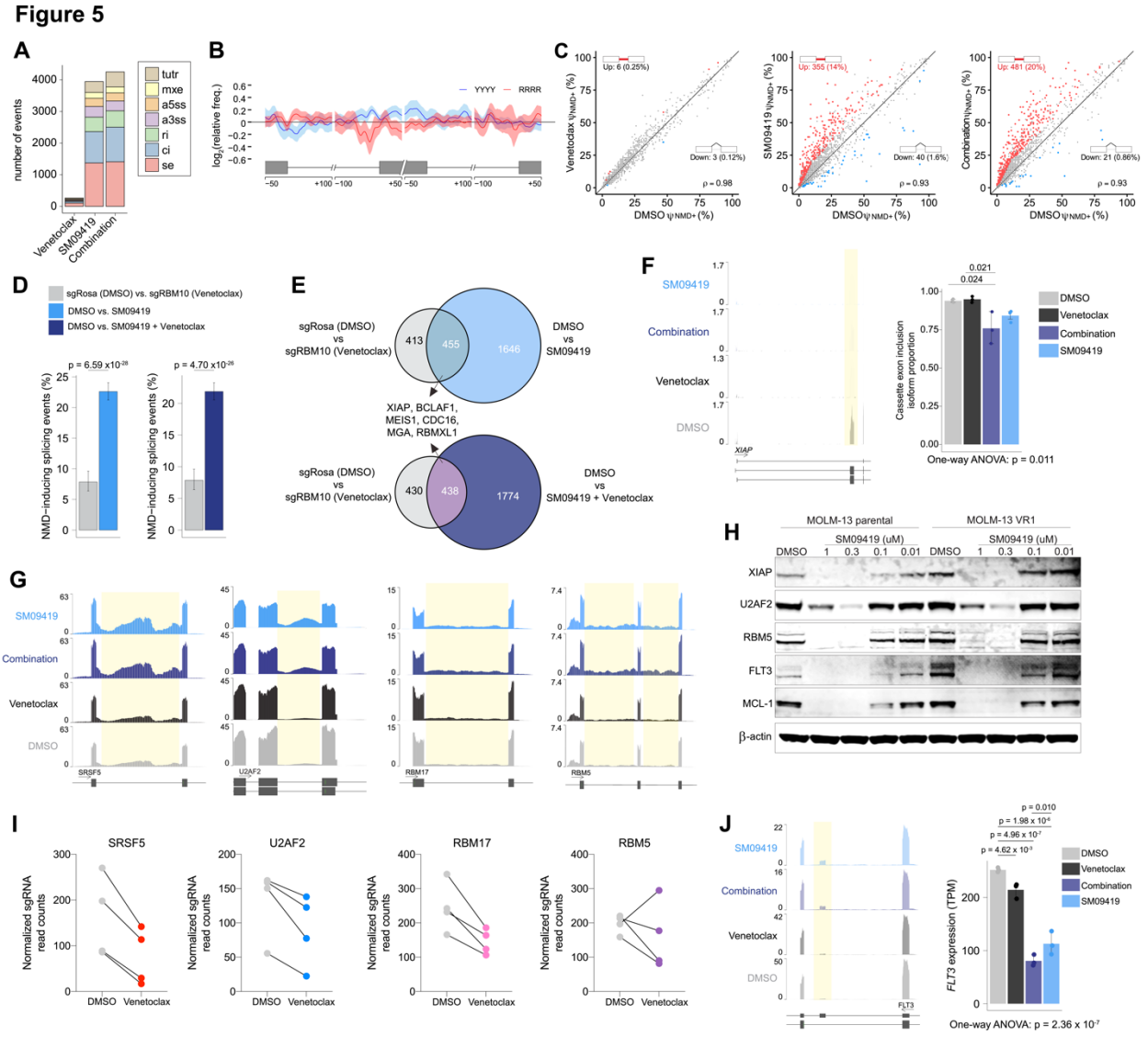


Figure 5. SM09419 promotes mis-splicing of key oncogenic pathways in AML.

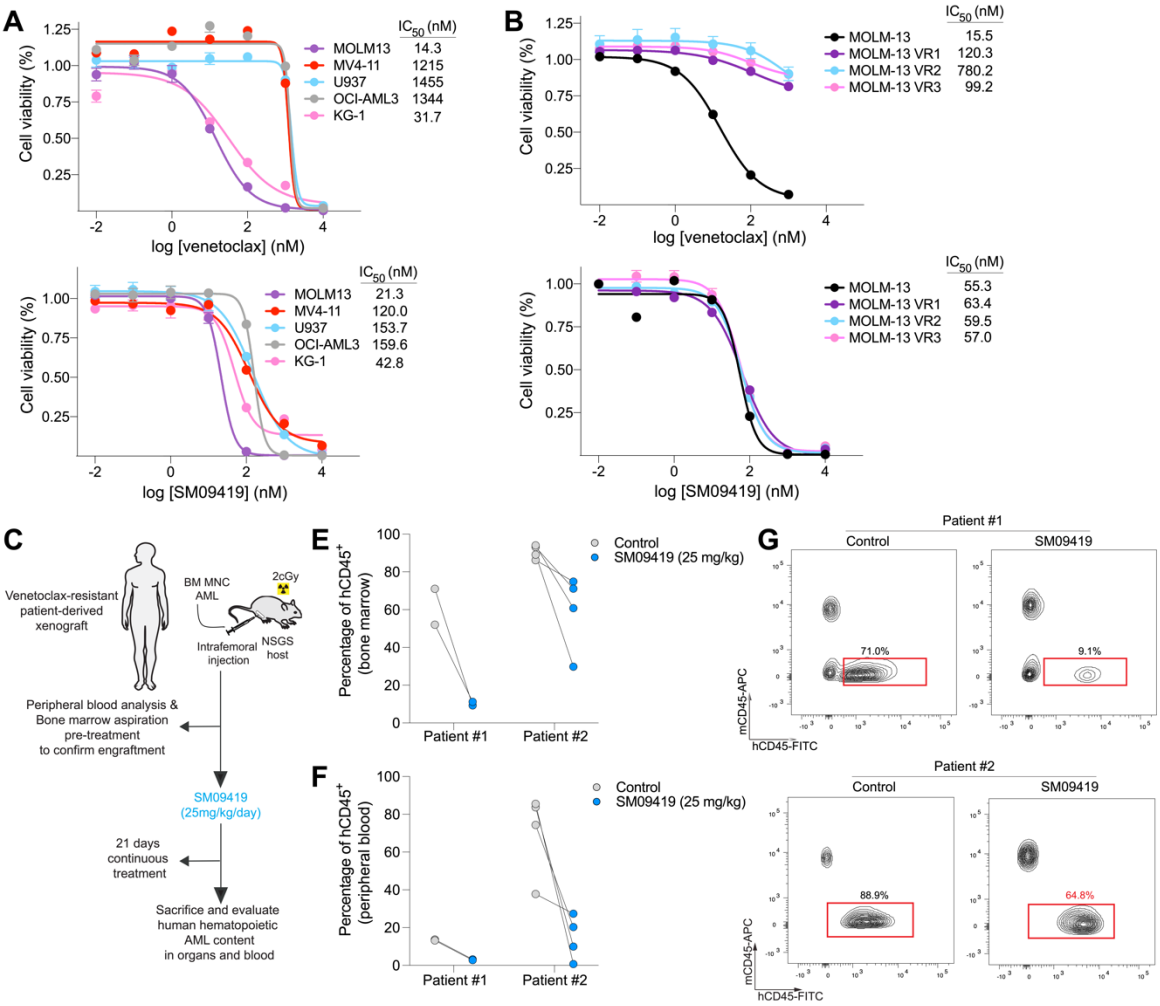
(A) Total number of splicing changes observed after SM09419 (100 nM), venetoclax (10 nM), or combination of SM09419 (100 nM) and venetoclax (10 nM) treatment for 48 hours. Cassette exons (SE), alternative 5' ss exon (A5E), alternative 3' ss exon (A3E), mutually exclusive exons (MXE), retained intron (RI), constitutive intron (CI), and tandem 3' UTR (TUTR).

(B) Spatial distribution of pyrimidines-rich (YYYY) and purine-rich (RRRR) motifs comparing sequence enrichment of excluded exons (n = 674) against included exons (n = 370) in SM0419-treated (100 nM) MOLM-13 cells.

(C) Scatter plot of NMD-inducing retained intron (RI) events (red circles) in MOLM-13 cells treated with venetoclax (left), SM09419 (middle) or the combination of venetoclax and SM09419 (right).

(D) Percentage of NMD-inducing events in *RBM10* KO venetoclax (compared to non-targeting sgRosa) and SM09419, or SM09419+venetoclax (compared to DMSO) RNA-seq.
(E) Venn diagram of NMD-inducing events in *RBM10* KO venetoclax (compared to non-targeting sgRosa) and SM09419, or SM09419+venetoclax (compared to DMSO).
(F) RNA-seq coverage plot (left) and mean PSI of *XIAP* cassette exon inclusion isoform.
(G) RNA-seq coverage plots of the splicing factors *SRSF5*, *U2AF2*, *RBM17*, and *RBM5* in MOLM-13 cells. Yellow regions represent retained intron events in each of the genes.
(H) Western blotting of *XIAP*, *U2AF2*, *RBM5*, *FLT3*, *MCL-1*, and actin in MOLM-13 parental or venetoclax-resistant (VR1) cells treated with varying concentration of SM09419 for 24 hrs.
(I) Normalized sgRNA counts of top splicing factors from RNA-binding protein CRISPR screen that synergized with venetoclax treatment in MOLM-13 cells.
(J) RNA-seq coverage plots (left) and gene expression (right) plots for *FLT3* mRNA. p-values were determined by One-way ANOVA with post-hoc testing as indicated. See also **Figure S6**.

Figure 6



D

Patient	Prior therapies	Reponse	Start of venetoclax	End of venetoclax	Sample date	Genetic alterations
1	VEN+Azacytidine	Primary refractory	01/18/2019	02/27/2019	06/05/2020	RUNX1, PTPN11, FLT3, SF3B1, STAG2
2	VEN+Low-dose Cytarabine	Primary refractory	01/03/2017	04/06/2017	04/10/2017	RUNX1, DNMT3A, IDH2

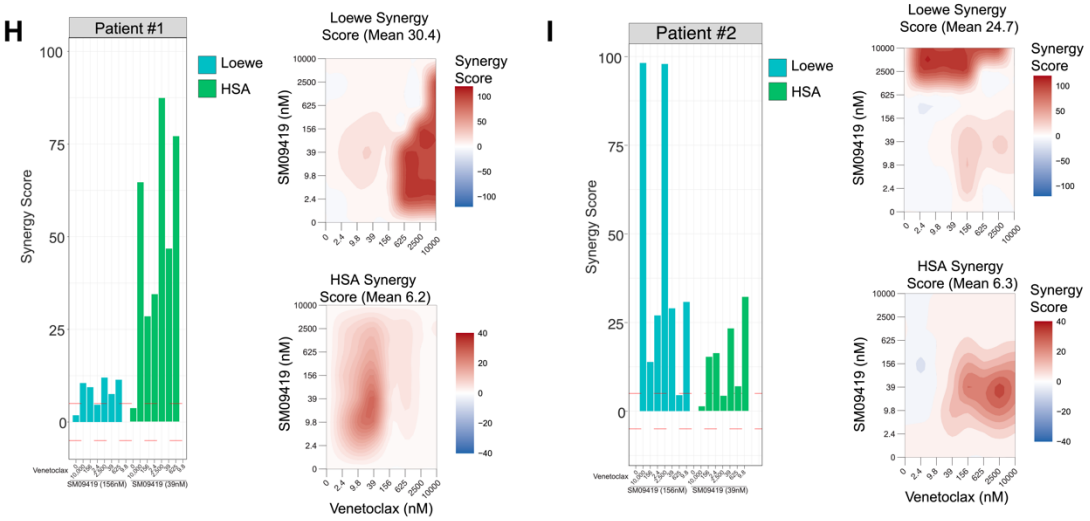


Figure 6. SM09419 circumvents therapeutic resistance to venetoclax.
(A) Dose-response curves of human AML cell lines treated with various concentrations of venetoclax (top) or SM09419 (bottom). IC50 values were calculated from technical triplicates per experiment, error bars represent SEM.
(B) Dose-response curves of venetoclax-resistant MOLM-13 cells treated with different concentrations of venetoclax (top) and SM09419 (bottom).
(C) Schematic of patient-derived xenograft (PDXs) generation and treated daily with SM09419 (25 mg/kg, QD, PO) or vehicle.
(D) Diagnosis, treatment regimen and genetic characteristics of AML patient-derived xenograft samples.
(E) Percentage of human *CD45*+ (h*CD45*+) cells in bone marrow and **(F)** peripheral blood of PDXs following 3-weeks of SM09419 treatment.
(G) Representative flow-cytometry plots of h*CD45*+ and mouse *CD45*+ (m*CD45*+) in bone marrow from PDXs treated daily with 25 mg/kg SM09419 after 3-weeks.
(H) Synergy scores (Loewe and HSA) (left) and 2D synergy plots (right) from ex vivo cultured patient #1 and **(I)** patient #2 samples treated with venetoclax, SM09419 or the combination after 48 hours.

Figure S1

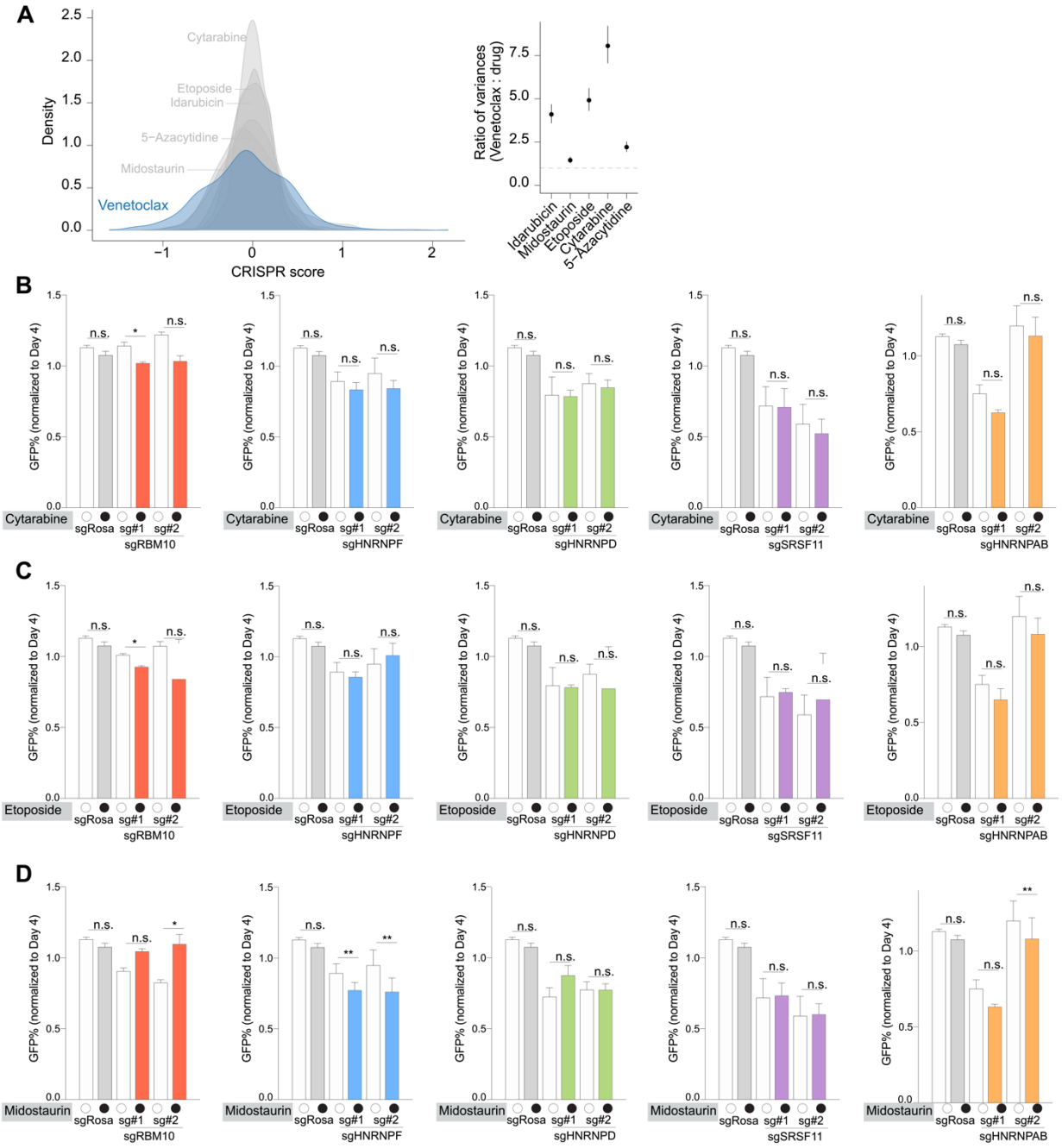


Figure S1. Targeting RNA splicing factors sensitizes AML cells to venetoclax.

(A) Distribution of CRISPR scores of sgRNAs targeting RNA processing genes (left) and variance comparisons (against venetoclax; right) from the genome-wide drug screens. The variance ratios and 95% confidence intervals were estimated via an F-test.

(B) Competition-based assay in MOLM-13 cells 10 days post-transduction with top 2 sgRNAs targeting each splicing factor or non-targeting sgRosa control (n=3, mean+SEM) treated with 50

nM cytarabine, **(C)** 400 nM etoposide, or **(D)** 25 nM midostaurin. Statistical analysis was performed using unpaired Student's t-test by Prism GraphPad (*$p < 0.05$, **$p < 0.01$, n.s., not significant).

Figure S2

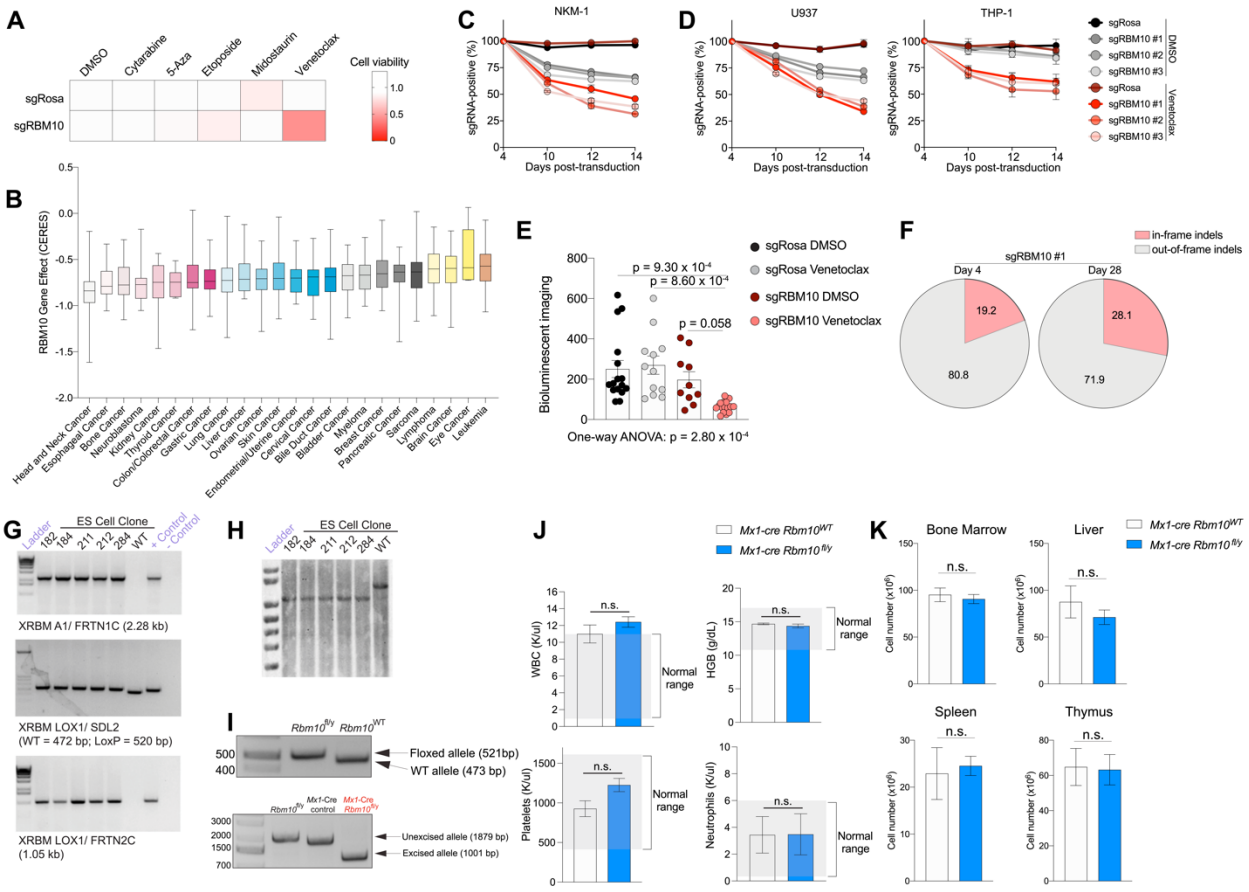


Figure S2. RBM10 ablation sensitizes AML cells to death from venetoclax but RBM10 is not required for normal hematopoiesis.

(A) Competition-based assay in MOLM-13 cells transduced with sgRBM10 treated with the indicated drugs (same dose used in CRISPR screens and indicated in the Methods section). Values are normalized to DMSO control.

(B) *RBM10* dependency score (CERES) across cancer types from DepMap database. Higher negative values are indicative that gene is essential.

(C) Competition proliferation assay of sgRBM10 or non-targeting sgRosa in *NKM-1* and **(D)** *TP53*-mutated cell lines (U937 and THP-1) treated with venetoclax or DMSO.

(E) Quantification of bioluminescent images from mice transplanted with RBM10 KO or sgRosa MOLM-13 cells at day 4 post-venetoclax (100 mg/kg) daily treatment (n=3, mean+SEM).

(F) CRISPResso indel analysis of *RBM10* sgRNAs in MOLM-13 cells at the indicated time-points.

(G) Genotyping of embryonic stem cell clones containing floxed *Rbm10* allele using three distinct genotyping primers.

(H) Southern blot confirmation of floxed ES cell clones using the Southern probe labeled as "PB3/4."

(I) Validation of Rbm10 floxed alleles and excision of exon 3 of *Rbm10* using genomic PCR.
(J) Peripheral blood counts and **(K)** total cell numbers of tissues from primary non-competitive transplantation of *Rbm10* cKO (n=4) and Mx1-cre control (n=4) after 1 month of pIpC treatment. Grey represents normal levels of blood counts. Statistical analysis was performed using an unpaired Student's t-test.

Figure S3

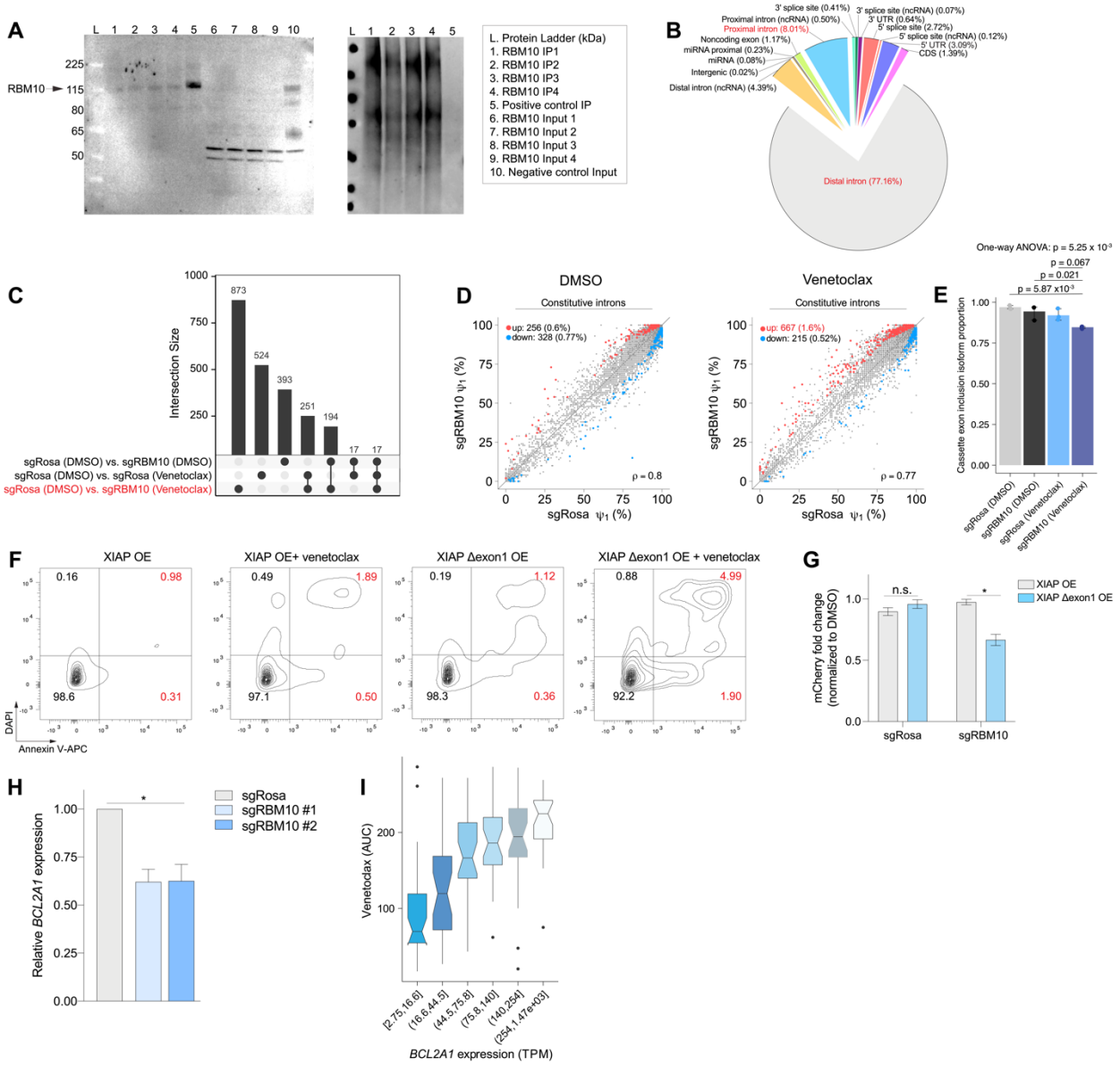


Figure S3. Characterization of RBM10 on RNA splicing and binding in AML cells.

(A) Immunoblotting of endogenous *RBM10* (left) and RNA visualization (right) of immunoprecipitation in MOLM-13 cells.

(B) Genomic distribution of *RBM10* eCLIP binding sites in MOLM-13 cells.

(C) UpSet plot of overlapping splicing events in RBM10 KO and sgRosa treated with DMSO or venetoclax.

(D) Scatter plots of constitutive introns [in MOLM-13 cells transduced with sgRosa (x-axis) or sgRBM10 (y-axis)] in DMSO (left) or venetoclax treatment (right).

(E) Percent spliced in (PSI) values of the cassette exon (exon 1) inclusion isoform of *XIAP* Δexon1.

(F) Annexin V staining of ectopic overexpression of *XIAP* full-length or *XIAP* Δexon1 treated with DMSO or venetoclax for 24 hours and **(G)** upon *RBM10* KO compared to non-targeting sgRosa. **(H)** RT-qPCR of *BCL2A1* mRNA expression in MV4-11 human AML cells with two independent *RBM10* sgRNAs and treated with 50 nM venetoclax for 48 hrs (n=3, mean+SEM). Statistical analysis was performed using un- paired Student's t-test by Prism GraphPad (*$p < 0.05$).
(I) Correlation of *BCL2A1* expression and venetoclax resistance (AUC) from BeatAML AML patients.

Figure S4

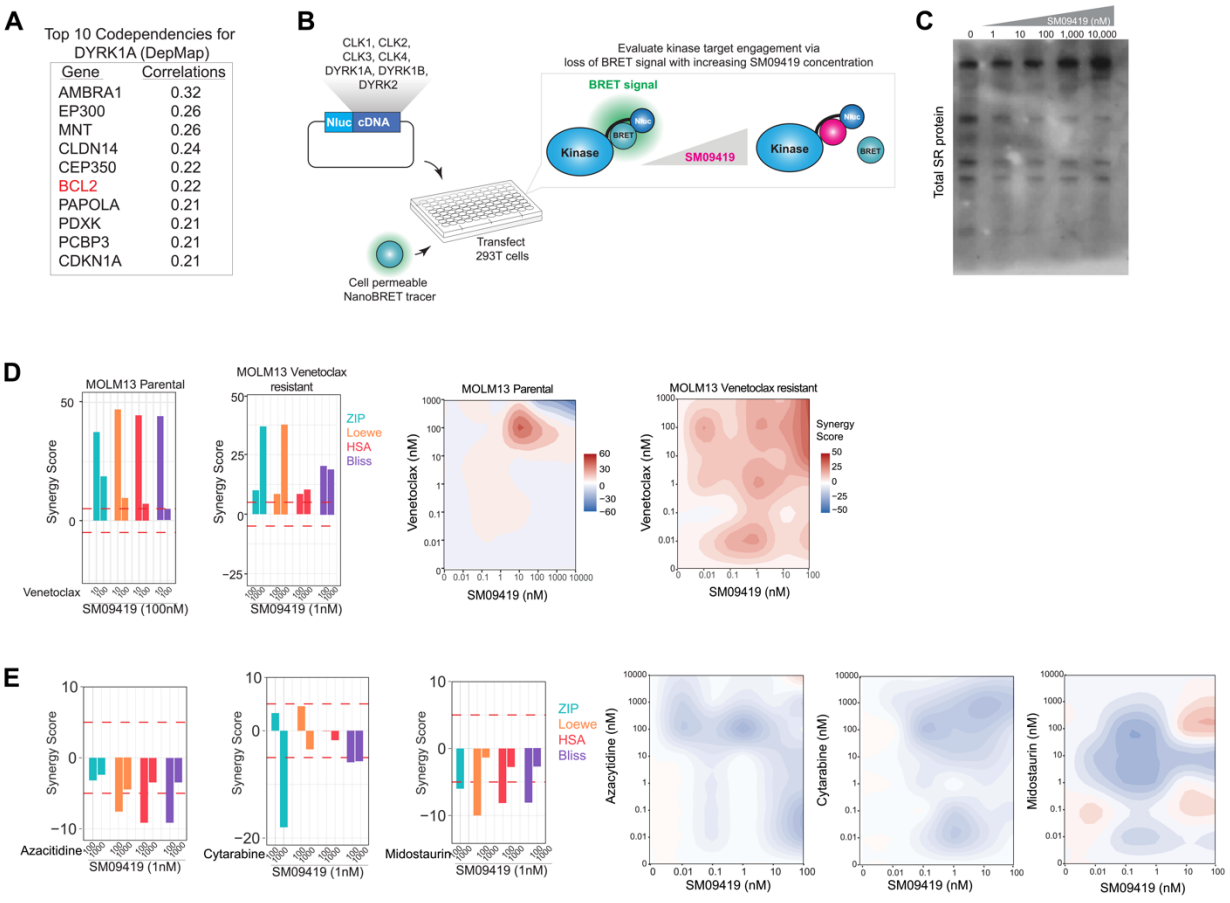


Figure S4. SM09419 is a highly specific CLK/DYRK inhibitor and synergizes with venetoclax.

(A) DepMap co-dependency CRISPR screen analysis of *DYRK1A*.

(B) NanoBRET target engagement assay of *CLK1-4*, *DYRK1A/B*, and *DYRK2* after treatment with varying concentrations of SM09419.

(C) Western blot of total SR protein levels in MOLM-13 cells.

(D) Four synergy scores (ZIP, Loewe, HSA, and Bliss) (left) and 2D synergy plots (right) of MOLM-13 parental or venetoclax-resistant cells treated with venetoclax, SM09419, or the combination after 48 hours.

(E) Same as in **(D)** except treated with 5-azacytidine, cytarabine, midostaurin, or the combination of each.

Figure S5

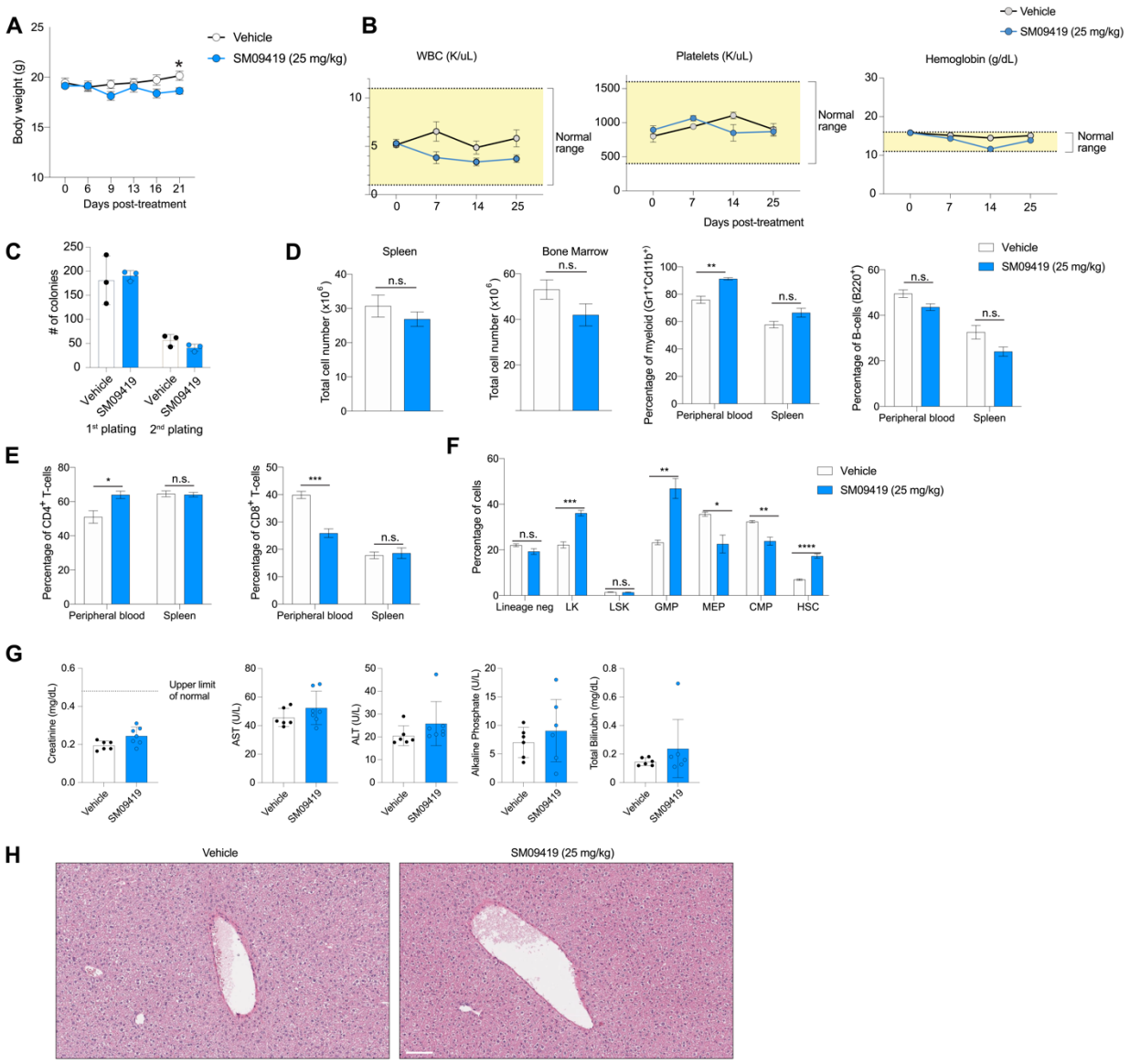


Figure S5. SM09419 is well-tolerable and does not alter normal hematopoiesis.

(A) Analysis of normal C57BL/6 mice body weight and **(B)** complete blood counts (CBCs) at the indicated time- points after daily treatment of SM09419 (25 mg/kg) or vehicle. Yellow area represents the normal ranges for blood counts.

(C) Total number of colony-forming units (CFUs) using methylcelllulose assays with normal C57BL/6 treated with SM09419 daily for 3-weeks. Colonies were assessed at day 7 after plating.

(D) Total cell numbers of tissues from normal C57BL/6 mice treated daily with SM09419 (25 mg/kg) or vehicle for 3 weeks.

(E) Flow cytometry analysis of T-cells from spleen and peripheral blood after 3 weeks of SM09419 daily treatment.

(F) Flow cytometric analysis of hematopoietic stem and progenitor cells (HSPCs) in bone marrow from mice treated daily with SM09419 for 3 weeks.
(G) Assessment of kidney function (creatinine test) and liver function (AST, ALT, Alkaline phosphate, total Bilirubin) after treatment with daily SM09419 for 3 weeks.
(H) Hematoxylin and eosin (H&E) staining of liver after treatment with daily SM09419 for 3 weeks (bar: 500μM). Statistical analysis was performed using unpaired Student's t-test by Prism GraphPad (*$p < 0.05$, n.s., not significant).

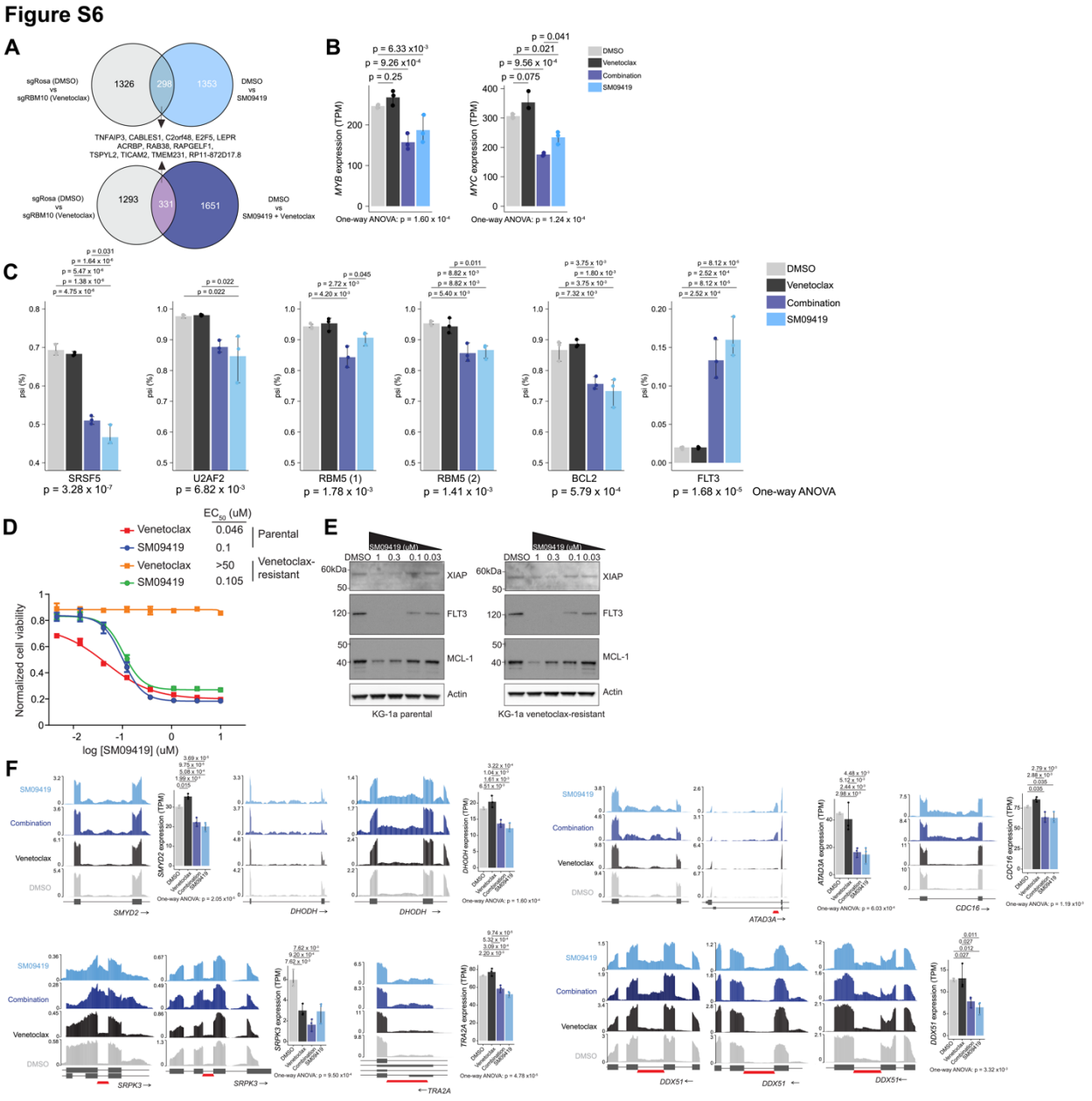


Figure S6. SM09419-responsive transcriptome and splicing changes in AML.

(A) Venn diagram of differentially genes expressed in *RBM10* KO treated with venetoclax (compared to sgRosa) and SM09419 monotherapy or venetoclax-combined (compared to DMSO) from RNA-seq in MOLM-13 cells.

(B) Gene expression for *MYC* and *MYB* mRNA from MOLM-13 RNA-seq.

(C) Mean PSI values of three replicates of each group treated with venetoclax or combination of SM09419 and venetoclax.

(D) Dose-response curves of KG-1a parental and venetoclax-resistant cell lines treated with venetoclax or SM09419 (normalized to DMSO) after 24 hours.

(E) Western blotting of *XIAP*, *FLT3*, *MCL-1*, and Actin in KG-1a parental or venetoclax-resistant cells after 24 hours of treatment.
(F) Mean PSI values and gene expression of *SMYD2*, *DHODH*, *ATAD3A*, *CDC16*, *SRPK3*, *TRA2A*, and *DDX51* of each group treated with venetoclax or combination of SM09419 and venetoclax. p-values were determined by One-way ANOVA with post-hoc testing as indicated.

4.6 MATERIALS AND METHODS

4.6.1 *Experimental Model and Subject Details*

4.6.1.1 Cell Lines and Cell Culture

All human leukemia cell lines were cultured in recommended media, typically RPMI medium with 20% FBS and 1% penicillin/streptomycin. TF-1 human AML cell line was cultured in RPMI 20% FBS, 1% penicillin/streptomycin and 2 ng/ml GM-CSF. HEK293T cells were grown in DMEM medium with 10% FBS and 1% penicillin streptomycin. Cell lines transduced with retroviral Cas9 blasticidin (Addgene plasmid no. 59262) were selected with blasticidin (Fisher) 48 hours after transduction. All transfections were performed in HEK293T cells using Polyethylenimine (PEI) reagent at 4:2:3 ratios of plasmid: pVSVG: pPax2 in OPTI-MEM solution. Viral supernatant was collected 48 hrs and 72 hrs post-transfection. Spin infections were performed at room temperature at 1,800 RPM for 30 mins with polybrene reagent (1:2000 dilution) (Fisher Scientific).

4.6.1.2 Animals

8-10 weeks-old C57BL/6, NSG-S and Mx1-Cre mice were purchased from Jackson Laboratory. 8 weeks-old NOD scid gamma female mice were obtained from Jackson Laboratory. Mice were bred and maintained in individual ventilated cages and fed with autoclaved food and water at Memorial Sloan Kettering Animal Facility. All animal procedures were completed in accordance with the Guidelines for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committees at MSKCC. All mouse experiments were performed in accordance with a protocol approved by the MSKCC Institutional Animal Care and Use Committees (13-04-003). Kaplan-Meier survival curves were compared using the Wilcoxon Rank-Sum test via GraphPad Prism.

4.6.1.3 Human patient samples

Studies were approved by the Institutional Review Boards of Memorial Sloan Kettering Cancer Center and conducted in accordance with the Declaration of Helsinki protocol. Primary human de-identified AML samples derived from whole peripheral blood or BM mononuclear cells were utilized. Mutational genotyping of each sample was performed by the MSKCC IMPACT assay as described previously (Cheng et al., 2015; Zehir et al., 2017). Cord blood was acquired from NY Blood Bank. Informed consent was obtained from all subjects to obtain the patient specimens used in the studies described. Specimens were obtained as part of the Memorial Sloan-Kettering Cancer Center Institutional Review Board approved clinical protocol #16-171 to which all subjects consented. O.A-W is a participating investigator on this protocol.

4.6.2 *Method Details*

4.6.2.1 CRISPR Screen

250 million MOLM13 Cas9-expressing cells were transduced with the Brunello sgRNA library (Doench et al., 2016) at a low multiplicity of infection (~0.3) to obtain at least 500 cells per sgRNA (500X). Spin infections were performed at room temperature at 1,500 RCF for 90 mins with polybrene reagent (1:2000 dilution) (Fisher Scientific). On Day 4 post-transduction, GFP percentage was assessed to determine infection efficiency and sgRNA coverage (~300-500X). Remaining 300-500X cells were placed back into culture after each passage until 20 days post-transduction. At day 8 post-transduction, pooled sgRNA cells were treated with either DMSO (1%), cytarabine (50 nM), 5-azacytidine (3 uM), etoposide (400 nM), idarubicin (5 nM), midostaurin (25 nM) or venetoclax (25 nM). Genomic DNA (gDNA) extraction using NucleoSpin Blood XL, Maxi kit for DNA from blood (Takara) according to manufacturer's protocol. For pooled CRISPR screen analysis, sgRNAs were normalized using the formula (sgRNA read

count/total read count) x CPM+1. Subsequently, normalized reads were then used to calculate $\log_2$ fold change (normalized read count drug treatment/normalized read count DMSO). CRISPR library amplifications were performed according to published study (Doench *et al.*, 2016). Competition assays were performed using MOLM-13 cells transduced with sgRNA or cDNA constructs and mixed with parental cells at fixed ratios followed by 4 days of treatment with either vehicle (DMSO) or venetoclax, and GFP percentages were analyzed using BD LSR Fortessa FlowCytometer. The RNA processing factor (genes in the "RNA processing" gene ontology term, GO:0006396) sgRNA $\log_2$ fold change distributions in cytarabine, 5-azacytidine, etoposide, idarubicin, and midostaurin were compared to venetoclax. Specifically, a two-sided F-test for equality of variances was used to assess if the drug:venetoclax ratios significantly deviated from 1. Variance ratios and the 95% confidence intervals were estimated using the stats R package.

4.6.2.2 CRISPR indel analysis

To quantify the spectrum of indel mutations with RBM10 sgRNAs, we transduced MOLM-13 cells with sgRBM10 or sgRosa (non-targeting), followed by cell sorting of GFP+/sgRNA+ populations at day 4 and day 28 post-infection. Cells were then harvested for gDNA and PCR amplicon (~200 bp) was designed to flank the sgRNA recognition sequence. 200 ng of gDNA was amplified using 2x Phusion Master Mix. Sequencing libraries were prepared from amplicons with an average size of 200 bp. The reported concentration was 3-7 ng/μL, and 50 μL were used as input for the KAPA Hyper Library Preparation Kit (Kapa Biosystems KK8504) according to the manufacturer's instructions with 8 cycles of PCR. Barcoded libraries were pooled in equal volumes and run on MiSeq in a PE150 run, using the MiSeq Reagent Micro Kit v2 (300 Cycles) (Illumina). The average number of read pairs per sample was 203,000. Indel analysis was performed using CRISPResso (http://crispresso.pinellolab.partners.org/)

4.6.2.3 RNA-sequencing library preparation and sequencing

For cell line RNA sequencing (RNA-seq), RNA was extracted from MOLM13 cells using the Qiagen RNeasy extraction kit, according to the manufacturer's instructions. A minimum of 500 ng of high-quality RNA (as determined by Agilent Bioanalyzer) per replicate was used as input for library preparation. Poly(A)-selected, strand-specific (dUTP method) Illumina libraries were prepared by the Integrated Genomics Operation (IGO) at Memorial Sloan Kettering with a modified TruSeq protocol and sequenced on the Illumina HiSeq 2000 to obtain ~50-60M 2x101 bp paired-end reads per sample.

4.6.2.4 eCLIP library preparation

eCLIP studies were performed in duplicates by Eclipse Bioinnovations Inc (San Diego, www.eclipsebio.com) according to the published single-end enhanced CLIP protocol with the following modifications. For Rbm10 immunoprecipitation, 10% of IP samples and 1% of input samples were run on NuPAGE 4-12% Bis-Tris protein gels, transferred to PVDF membrane, probed with 1:1,000 of RBM10 antibody and 1:10,000 TrueBlot Anti Rabbit IgG (HRP) and imaged with C300 Imager for 1 minute on normal settings using Azure Radiance ECL. Only the region from ~100 kDa to 180 kDa (protein size to 80 kDa above) was isolated during eCLIP. For RNA visualization, 10% of IP samples were run on NuPAGE 4-12% Bis-Tris protein gels, transferred to nitrocellulose membrane, visualized using the Chemiluminescent Nucleic Acid Detection Kit (cat. no. 89880) from Thermo Fisher Scientific and imaged with C300 Imager for 30 seconds on normal settings. For eCLIP preparation, 10 million MOLM-13 cells were UV crosslinked at 400 mJoules/cm2 with 254 nm radiation, and snap frozen. Cells were then lysed and treated with RNase I to fragment RNA as previously described. RBM10 antibody (A301-006A, Bethyl) was then pre-coupled to Protein G Dynabeads (Thermo Fisher), added to lysate, and

incubated overnight at 4 deg C. Prior to immunoprecipitation, 2% of the sample was taken as the paired input sample, with the remainder magnetically separated and washed with lysis buffer only (as the standard high-salt eCLIP wash buffer gave poor immunoprecipitation yield). eCLIP was performed by excising the area from ~100 kDa to ~180 kDa. RNA adapter ligation, IP-western, reverse transcription, DNA adapter ligation, and PCR amplification were performed as previously described.

4.6.2.5 Whole-exome sequencing and targeted capture sequencing

For MSKCC-IMPACT, after PicoGreen quantification and quality control by Agilent BioAnalyzer, 100 ng of DNA were used to prepare libraries using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) with 8 cycles of PCR. 80-190 ng of each barcoded library were captured by hybridization in pools of 6-14 samples using the IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) assay (Cheng *et al.*, 2015) (Nimblegen SeqCap), designed to capture all protein-coding exons and select introns of 505 commonly implicated oncogenes, tumor suppressor genes, and members of pathways deemed actionable by targeted therapies. Captured pools were sequenced on a NovaSeq 6000 in a PE100 run using the NovaSeq 6000 S4 Reagent Kit (200 Cycles) (Illumina) producing an average of 540X coverage per sample. For exome capture and sequencing, after PicoGreen quantification and quality control by Agilent BioAnalyzer, 100 ng of DNA were used to prepare libraries using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) with 8 cycles of PCR. After sample barcoding, 500 ng of library were captured by hybridization using the xGen Exome Research Panel v2.0 (IDT) according to the manufacturer's protocol. PCR amplification of the post-capture libraries was carried out for 12 cycles. Samples were run on a NovaSeq 6000 in a PE100 run, using the NovaSeq 6000 S4 Reagent Kit (200 Cycles) (Illumina). Samples were covered to an average of 251X.

4.6.2.6 Western blotting

MOLM-13 Cas9-expressing cells were transduced with sgRNAs and harvested for protein on day 6 post-transduction. For SM09419, MOLM-13 cells were treated with varying concentrations of SM09419, and protein was harvested 48 hours post-treatment. Lysate protein concentration was measured with the BCA reagent and 10-30 mcg was loaded per lane onto 4-12% NuPAGE™ Bis-Tris protein gels. After transfer, PVDF membranes were probed with anti-RBM10 (Bethyl Laboratories), anti-Phosphoepitope SR proteins Antibody (clone 1H4, Millipore Sigma), total SR protein (Santa Cruz), anti-XIAP (Cell signaling), anti-MCL-1 (Cell signaling), anti-RBM5 (Abcam), anti-FLT3 clone 8F2 (Cell signaling), anti-U2AF2/U2AF65 (Abcam) and anti-BCL-2 (Abcam) at 1:1,000 and visualized by standard methods.

4.6.2.7 Colony-forming assays

Total bone marrow from Mx1-Cre WT and Mx1-Cre Rbm10fl/y mice were harvested and seeded at a density of 20,000 cells per replicate into cytokine-supplemented methylcellulose medium (MethoCult M3434, Stemcell Technologies). For SM09419 experiments, total bone marrow from C57BL/6 treated with 25 mg/kg SM09419 for 3 weeks were harvested and seeded as described above. Colonies propagated in culture were scored at day 7.

4.6.2.8 Annexin V assay

Apoptotic analysis was determined using APC Annexin V (BD Bioscience) and performed according to manufacturer's specifications and co-stained with 4',6-Diamidino-2-Phenylindole, Dihydrochloride (DAPI) for DNA content. Cells were analyzed by flow cytometry and FlowJo software.

4.6.2.9 Generation of Rbm10 conditional knockout mice

The *Rbm10* allele was deleted by targeting exon 4. Two LoxP sites flanking exon 3 and a Frt flanked neomycin selection cassette were inserted in the downstream intron. Ten micrograms of the targeting vector were linearized and then transfected by electroporation of HF4 (129/SvEv x C57Bl/6) (FLP Hybrid) embryonic stem cells. After selection with G418 antibiotic, surviving clones were expanded for PCR analysis to identify recombinant ES clones. The Neo cassette in targeting vector has been removed during ES clone expansion. Screening primer A1 was designed downstream of the short homology arm (SA) outside the 3' region used to generate the targeting construct. Clones 182, 184, 211, 212, and 284 were expanded and reconfirmed for SA integration. A PCR was performed on clones 182, 184, 211, 212, and 284 to detect presence of the distal LoxP site using the LOX1 and SDL2 primers. This reaction amplifies a wild-type product 472 bp in size. The presence of a second PCR product 48 bp greater than the wild-type product indicates a positive LoxP PCR. Confirmation of distal LoxP retention was performed by PCR using the LOX1 and FRTN2C primers. This reaction produces a product 1.05 kb in size. Sequencing was performed on purified PCR DNA to confirm presence of the distal LoxP cassette using the SDL2 primer. Secondary confirmation of positive clones identified by PCR was performed by Southern Blotting analysis. DNA was digested with Apa I, and electrophoretically separated on a 0.8% agarose gel. After transfer to a nylon membrane, the digested DNA was hybridized with a probe targeted against the 5' external region. DNA from HF4 mouse ES cells was used as a wild-type control. Positive clones were further confirmed by Southern Blotting analysis using an internal probe. DNA was digested with BamH I, and electrophoretically separated on a 0.8% agarose gel. After transfer to a nylon membrane, the digested DNA was hybridized with a probe targeted against the 3' internal region. DNA from HF4 mouse ES cells was used as a wild-type control. Primer set NDEL1 and NDEL2 was used to screen mice for the deletion of the Neo cassette. The PCR product

for the wild-type is 322 bp. After Neo deletion, one set of LoxP-FRT sites remains (147 bp). A second band with a size of 469 bp indicates Neo deletion. A PCR was performed to detect presence of the distal LoxP site using the LOX1 and SDL2 primers. This reaction amplifies a wild-type product 473 bp in size. The presence of a second PCR product 48 bp greater than the wild-type product indicates a positive LoxP PCR. Tail DNA samples from positive mice were amplified with primers NEOGT and A1. NEOGT is located inside the Neo cassette and A1 is located downstream of the short homology arm, outside the region used to create the targeting construct. NEOGT / A1 amplifies a fragment of 2.32 kb in length.

4.6.2.10 Bone marrow (BM) transplantation

Freshly dissected femora and tibiae were isolated from *Mx1*-cre WT and *Mx1*-cre *Rbm10*$^{fl/y}$, CD45.2^{+} mice. BM was flushed with a 3-cc insulin syringe into PBS supplemented with 3% fetal bovine serum. The BM was spun at 0.5 g by centrifugation and RBCs were lysed in ammonium chloride-potassium bicarbonate lysis buffer for 5 min. After centrifugation, cells were resuspended in PBS plus 3 % FBS, passed through a cell strainer, and counted. Finally, 0.5 million total BM cells of *Mx1*-cre WT and *Mx1*-cre *Rbm10*$^{fl/y}$ CD45.2^{+} mice were mixed with 0.5 million WT CD45.1^{+} support BM and transplanted via tail vein injection into lethally irradiated (two times 450 cGy) CD45.1^{+} recipient mice. Chimerism was measured by FACS from the peripheral blood 4 weeks after transplant. Chimerism was followed via FACS in the peripheral blood every 4 weeks (week 0, 4, 6, 8,12, and 16 after polyI:polyC injection). For noncompetitive transplantation experiments, 1 million total BM cells of *Mx1*-cre WT and *Mx1*-cre *Rbm10*$^{fl/y}$ CD45.2^{+} mice were injected into lethally irradiated (two times 450 cGy) CD45.1^{+} recipient mice.

4.6.2.11 Drug treatment IC50 measurements

Cell lines were plated in 96 well plates and exposed to the indicated compounds at various concentration ranges with a minimum of three technical replicates per concentration per cell line. Cell viability was measured with the CellTiter Glo reagent (Promega) as per manufacturer's instructions. Absolute viability values were converted to percentage viability versus DMSO control treatment, and then non-linear fit of log(inhibitor) versus response (three parameters) was performed in GraphPad Prism v7.0 to obtain an IC_{50} values. Two-dimensional heatmaps of Synergy Scores from Bliss synergy models were generated based on Demidenko et al., 2019 (Demidenko and Miller, 2019).

4.6.2.12 QPCR measurement of BCL2A1 gene expression

RNA was extracted from the indicated cell lines and reverse transcribed into cDNA using the Verso cDNA synthesis Kit (ThermoFisher Scientific). Measurement of *BCL2A1* gene expression was performed using primers amplifying *BCL2A1* CDS region and designed by primer3 (https://bioinfo.ut.ee/primer3-0.4.0/) with *ACTB* as the housekeeping gene. Relative expression levels across cell lines were calculated using the Delta-delta Ct method as per standard procedures.

4.6.2.13 cDNA overexpression

BCL2A1, *RBM10* wild-type and RBM10 domain mutants as well as *XIAP* full-length and □exon 1 were codon optimized and synthesized as gene blocks by Integrated DNA Technologies (IDT) and was sublconed cloned into lentiviral Puro-IRES-GFP construct using NEBuilder Hifi DNA assembly. MOLM-13 *RBM10*-KO cells were transduced with either *BCL2A1*, *RBM10* wild-type, or *RBM10* mutant constructs and treated with venetoclax.

4.6.2.14 Animal experiments

For *in vivo* Cas9 experiments, MOLM-13 Cas9-expressing cells were transduced with sgRosa (negative control) or sgRBM10 constructs. At day 2 post-transduction, sgRNA positive cells (GFP^+) were sorted by FACS. 100,000 leukemia-sgRNA expressing cells were intravenously injected into each sub-lethal irradiated (5.5 Gy) 8 weeks-old NOD scid gamma mice mice. For venetoclax trials, a 100 mg/ml venetoclax (Sigma Aldrich) stock was diluted in a carrier containing 10% ethanol, 30% polyethyleneglycol-400 (Sigma), and 60% phosal 50 propylene glycol (Lipoid) to obtain a final concentration of 100 mg/kg. Upon disease onset as measured by bioluminescent imaging, we performed oral gavage once daily with either 100 mg/kg venetoclax or vehicle (1% DMSO). All whole-body bioluminescent imaging was performed by intraperitoneally injection of Luciferin (Goldbio) at a 50 mg/kg concentration and imaging was performed after 5 mins using an IVIS imager. Bioluminescent signals (radiance) were quantified using Living Image software with standard regions of interests (ROI) rectangles.

4.6.2.15 Kinase assays

IC_{50} values for *CLK2*, *CLK3*, *DYRK1A* and *CDK1* were determined by transferring test compounds to 1536-well plates (Echo 550, LabCyte) and by optimizing and performing Z'-LYTE™ kinase assays per the manufacturer's instructions (Thermo Fisher). In addition, a full kinome screen (464 kinases) with 1 μM SM09419 was performed by Thermo Fisher Select Screen service. The IC_{50} for each kinase demonstrating >80% inhibition was then determined. Kinase tree dendrogram was generated using Coral (Metz et al., 2018).

4.6.2.16 NanoBRET target engagement assay

Cellular target engagement assays were performed using NanoBRET in 293T cells expressing *CLK1*, *CLK2*, *CLK3*, *CLK4*, *DYRK1A*, *DYRK1B*, and *DYRK2* in-frame with a nanoluciferase

(NanoLuc) tag. A cell permeable NanoBRET fluorescent tracer was then added to the cells which reversibly binds the target-NanoLuc Fusion protein in live cells to result in a BRET signal. SM09419 or vehicle were then added to each cell over a dose range and the degree of drug-target protein binding was assessed via loss of NanoBRET signal. An IC_{50} value indicating SM09419-protein binding was then identified via 10-point dose response curves.

4.6.2.17 Patient-derived xenograft experiments

Frozen human peripheral blood mononuclear cells (PBMCs) from two individual PDX models were rapidly thawed and transferred into 50 ml conical tubes. 20 mL pre-warmed RPMI 1640 (Corning) was added dropwise to tubes. After centrifuging at 300 x g at 4 degrees Celsius, cell pellet was resuspended in PBS (Corning). 4 million cells were intrafemorally injected per mouse. Blood was collected by retro-orbital bleeding using heparinized microhematocrit capillary tubes (Thermo Fisher Scientific) and a flow cytometry panel consisting of mCD45/ hCD45/hCD3/hCD11b/hB220 were used to discriminate human from mouse cells and human myeloid vs T-cell engraftment. Upon disease onset as measured by hCD45-positive cells by flow cytometry, we performed oral gavage once daily with either 25 mg/kg SM09419 or vehicle (5% polyvinylpyrrolidone).

4.6.3 *Quantification and Statistical Analysis*

4.6.3.1 Genome-wide differential gene expression analysis

FASTQ files were first trimmed using Trim_galore (v0.6.4) to remove sequencing adapters and low quality (Q<15) reads. Trimmed sequencing reads were aligned to the human Hg19 reference genome (GENCODE, GRCh37.p13) using STAR (v2.7.5) (Dobin et al., 2013). SAM files were subsequently converted to BAM files, sorted, and indexed using samtools (v1.9). BAM files were used to generate bigwig files using bamCoverage (part of the Deeptools package; v3.3.1). Read

counting across genomic features was performed using featureCounts (part of the subread package; v1.5.0) (Liao et al., 2014).

4.6.3.1 Gene expression estimation and alternative splicing analysis

Annotations from UCSC knownGene (Meyer et al., 2013), Ensembl 71 (Flicek et al., 2013), and MISO v2.0(Katz et al., 2010) were combined to create a genome annotation for the human UCSC hg19 (GRCh37) assembly. We mapped all reads to the transcriptome via RSEM v1.2.4(Li and Dewey, 2011), using the Bowtie alignment option "-v 2"(Langmead et al., 2009). RSEM produces gene-level estimates of expression in units of transcripts per million (TPM). All gene expression estimates were normalized via the trimmed mean of M values (TMM) method (Robinson and Oshlack, 2010) . Reads which failed to align were mapped to the genome with TopHat v2.0.8b (Trapnell et al., 2009), as well to an expanded annotation created by computing all possible combinations of annotated 5′ and 3′ splice sites per gene. We quantified isoform expression with MISO v2.0 (Katz et al., 2010), using the combined RSEM and TopHat alignments as input. We used the two-sided t-test to test differential isoform expression between sample groups. Differentially spliced events were defined as those with at least 20 isoform-identifying reads in each sample, a minimum absolute difference of 10% in isoform expression, and a p-value < 0.05. All analyses were conducted within the R Programming environment with tools from Bioconductor (Huber et al., 2015). The visualizations were created using the dplyr, ggplot2, and UpSetR (Conway et al., 2017) packages.

4.6.3.1 Purine/Pyrimidine Motif Enrichment Analysis

Differentially spliced cassette exon events following SM09419 treatment were identified. The enrichment of purines/pyrimidines in excluded, relative to included, cassette exons was measured within exonic regions and immediately adjacent intronic sequences. The 95% confidence interval

was estimated with bootstrapping (1000 resampling iterations). The motif enrichment analysis was conducted within the R Programming environment with GenomicRanges from Bioconductor (Huber et al., 2015).

4.6.3.1 eCLIP data analysis

The eCLIP data was processed similarly as described previously (Van Nostrand et al., 2016) and is outlined shortly in the following. First, adapter sequences were trimmed from both reads of all read-pairs using cutadapt version 1.14. Then, all remaining reads longer than 16 bases were aligned against the human reference genome sequence hg19/GRCh37 using STAR version 2.5.0c. Only uniquely mapped reads were kept. Read-pair duplicates by position were removed using picard tools version 2.6.0. To identify binding sites, we first ran a custom script to identify clusters of overlapping reads that had a read-depth of at least 10 reads. Then, we calculated significant enrichments for all such identified clusters by comparing IP-samples versus input-samples using edgeR. More specifically, we ran bamutils count version 0.5.7 to counted stranded reads within all identified clusters for all samples. Using this output, we calculated differential coverage between IP-vs-input for each cluster with edgeR after normalizing for total sequencing depth per replicate (resulting in counts per million/CPM per cluster). Final binding sites were called by applying logFC > 2 and FDR < 0.05 thresholds between IP-vs-input. Identification of RBM10 binding positions in events alternatively spliced following RBM10 KO relied on the htseq-clip suite (https://htseq-clip.readthedocs.io), and the DEWSeq (Huppertz et al., 2022) and GenomicRanges Bioconductor packages (Lawrence et al., 2013). In brief, the GRCh38.v40 GENCODE annotation was processed into 50 nucleotide (nt) genomic sliding windows, with step size of 20 nt, using htseq-clip. From the STAR-aligned eCLIP BAM files, htseq-clip was used to identify crosslink positions and count their abundance in each window. The htseq-clip counts matrix was used as

input to DEWSeq for normalization and identification of IP-vs-input significantly enriched windows (adjusted p-value < 0.05, logFC > 2) in protein-coding regions. The p-values were FDR-adjusted via Independent Hypothesis Weighting (Ignatiadis et al., 2016) and overlapping significantly enriched windows were combined. The positions of enriched windows in alternatively spliced regions identified from our RNA-seq analyses were determined using the GenomicRanges package.

4.6.3.1 Gene Ontology analysis

Gene set enrichment was performed using the fgsea R package (1.4.0) using the KEGG, GO and MsigDB specific signatures according to the manual.

4.6.3.1 Statistical analysis

Kaplan-Meier survival curve p-values were performed using Log rank Mantel-COX test. For statistical comparison, we performed unpaired Student's t test. Statistical analyses were performed using Prism 7 software (GraphPad). Data with statistical significance are as indicated, *p< 0.05, **p< 0.01, ***p< 0.001.

4.6.3.1 Data and Software Availability

Gene Expression Omnibus: all newly generated RNA-seq and eCLIP data were deposited under accession number GSE199161.

4.7 ACKNOWLEDGMENTS

O.A.-W. and R.K.B. were supported in part by the Edward P. Evans Foundation, NIH/NCI (R01 CA251138), and NIH/NHLBI (R01 HL128239). O.A.-W. was supported in part by the NIH/NCI (R01 CA242020; P50 CA254838-01), and The Leukemia & Lymphoma Society. R.K.B. was supported in part by NIH/NHLBI (R01 HL151651) and the Blood Cancer Discoveries Grant

program through the Leukemia & Lymphoma Society, Mark Foundation for Cancer Research, and Paul G. Allen Frontiers Group (8023-20). R.K.B is a Scholar of The Leukemia & Lymphoma Society (1344-18) and holds the McIlwain Family Endowed Chair in Data Science. W.J.K. was supported by a Medical Scientist Training Program grant from the National Institute of General Medical Sciences of the National Institutes of Health under award number: T32GM007739 to the Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program. Computational studies were supported in part by FHCRC's Scientific Computing Infrastructure (ORIP S10 OD028685). We acknowledge the use of the Integrated Genomics Operation Core, funded by the NCI Cancer Center Support Grant (CCSG, P30 CA08748), Cycle for Survival, and the Marie-Josée and Henry R. Kravis Center for Molecular Oncology.

4.8 Author Contributions

E.W., J.M.B.P., E.M., C.B., R.K.B., and O.A.-W. designed the study. E.W., E.C., C.B., and C.-C.M. performed *in vitro* experiments. E.W., J.B., W.J.K., S.C., M.E.S., D.C., C.E.E., K.K., R.S., S.T., and C.B., performed *in vivo* experiments. M.S. and J.P.B. identified and provided patient materials. J.M.B.P., S.J.H., E.M., and R.K.B. performed computational analyses. G.K., J.A.F. and O.A.-W. generated the *Rbm10* conditional knockout animal model. W.J.K., S.C., S.J.H., C.H., K.K., R.F.S., E.M., C.B., M.J., and D.M.B. assisted with experimental design and data interpretation. E.W., J.M.B.P., E.M., C.B., R.K.B., and O.A.-W. wrote the manuscript with input from all authors.

4.9 Competing Interests

E.M., E.C., M.J., C.B., C.-C.M., and D.M.B. are employees of Biosplice Therapeutics. O.A.-W. has served as a consultant for H3B Biomedicine, Foundation Medicine Inc., Merck, Prelude

Therapeutics, and Janssen, and is on the Scientific Advisory Board of Envisagenics Inc., AIChemy, Harmonic Discovery Inc., and Pfizer Boulder; O.A.-W. has received prior research funding from H3B Biomedicine, Nurix Therapeutics, and LOXO Oncology unrelated to the current manuscript. The remaining authors declare no competing interests.

4.10 REFERENCES

Aird, D., Teng, T., Huang, C.L., Pazolli, E., Banka, D., Cheung-Ong, K., Eifert, C., Furman, C., Wu, Z.J., Seiler, M., et al. (2019). Sensitivity to splicing modulation of BCL2 family genes defines cancer therapeutic strategies for splicing modulators. Nat Commun *10*, 137. 10.1038/s41467-018-08150-5.

Alvarez, M., Estivill, X., and de la Luna, S. (2003). DYRK1A accumulates in splicing speckles through a novel targeting signal and induces speckle disassembly. J Cell Sci *116*, 3099-3107. 10.1242/jcs.00618.

Aubol, B.E., Plocinik, R.M., Hagopian, J.C., Ma, C.T., McGlone, M.L., Bandyopadhyay, R., Fu, X.D., and Adams, J.A. (2013). Partitioning RS domain phosphorylation in an SR protein through the CLK and SRPK protein kinases. J Mol Biol *425*, 2894-2909. 10.1016/j.jmb.2013.05.013.

Aubol, B.E., Wu, G., Keshwani, M.M., Movassat, M., Fattet, L., Hertel, K.J., Fu, X.D., and Adams, J.A. (2016). Release of SR Proteins from CLK1 by SRPK1: A Symbiotic Kinase System for Phosphorylation Control of Pre-mRNA Splicing. Mol Cell *63*, 218-228. 10.1016/j.molcel.2016.05.034.

Baselga, J., Cortes, J., Kim, S.B., Im, S.A., Hegg, R., Im, Y.H., Roman, L., Pedrini, J.L., Pienkowski, T., Knott, A., et al. (2012). Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. N Engl J Med *366*, 109-119. 10.1056/NEJMoa1113216.

Bisaillon, R., Moison, C., Thiollier, C., Krosl, J., Bordeleau, M.E., Lehnertz, B., Lavallee, V.P., MacRae, T., Mayotte, N., Labelle, C., et al. (2020). Genetic characterization of ABT-199 sensitivity in human AML. Leukemia *34*, 63-74. 10.1038/s41375-019-0485-x.

Blombery, P., Lew, T.E., Dengler, M.A., Thompson, E.R., Lin, V.S., Chen, X., Nguyen, T., Panigrahi, A., Handunnetti, S.M., Carney, D.A., et al. (2022). Clonal hematopoiesis, myeloid disorders and BAX-mutated myelopoiesis in patients receiving venetoclax for CLL. Blood *139*, 1198-1207. 10.1182/blood.2021012775.

Breems, D.A., Van Putten, W.L., Huijgens, P.C., Ossenkoppele, G.J., Verhoef, G.E., Verdonck, L.F., Vellenga, E., De Greef, G.E., Jacky, E., Van der Lelie, J., et al. (2005). Prognostic index for adult patients with acute myeloid leukemia in first relapse. J Clin Oncol *23*, 1969-1978. 10.1200/JCO.2005.06.027.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543-550. 10.1038/nature13385.

Chalfant, C.E., Ogretmen, B., Galadari, S., Kroesen, B.J., Pettus, B.J., and Hannun, Y.A. (2001). FAS activation induces dephosphorylation of SR proteins; dependence on the de novo generation of ceramide and activation of protein phosphatase 1. J Biol Chem *276*, 44848-44855. 10.1074/jbc.M106291200.

Chen, X., Glytsou, C., Zhou, H., Narang, S., Reyna, D.E., Lopez, A., Sakellaropoulos, T., Gong, Y., Kloetgen, A., Yap, Y.S., et al. (2019). Targeting Mitochondrial Structure Sensitizes Acute Myeloid Leukemia to Venetoclax Treatment. Cancer Discov *9*, 890-909. 10.1158/2159-8290.CD-19-0117.

Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., et al. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn *17*, 251-264. 10.1016/j.jmoldx.2014.12.006.

Collins, K.M., Kainov, Y.A., Christodolou, E., Ray, D., Morris, Q., Hughes, T., Taylor, I.A., Makeyev, E.V., and Ramos, A. (2017). An RRM-ZnF RNA recognition module targets RBM10 to exonic sequences to promote exon exclusion. Nucleic Acids Res *45*, 6761-6774. 10.1093/nar/gkx225.

Colwill, K., Feng, L.L., Yeakley, J.M., Gish, G.D., Caceres, J.F., Pawson, T., and Fu, X.D. (1996). SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. J Biol Chem *271*, 24569-24575. 10.1074/jbc.271.40.24569.

Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938-2940. 10.1093/bioinformatics/btx364.

de Graaf, K., Czajkowska, H., Rottmann, S., Packman, L.C., Lilischkis, R., Luscher, B., and Becker, W. (2006). The protein kinase DYRK1A phosphorylates the splicing factor SF3b1/SAP155 at Thr434, a novel in vivo phosphorylation site. BMC Biochem *7*, 7. 10.1186/1471-2091-7-7.

Demidenko, E., and Miller, T.W. (2019). Statistical determination of synergy based on Bliss definition of drugs independence. PLoS One *14*, e0224137. 10.1371/journal.pone.0224137.

DiNardo, C.D., Jonas, B.A., Pullarkat, V., Thirman, M.J., Garcia, J.S., Wei, A.H., Konopleva, M., Dohner, H., Letai, A., Fenaux, P., et al. (2020). Azacitidine and Venetoclax in Previously Untreated Acute Myeloid Leukemia. N Engl J Med *383*, 617-629. 10.1056/NEJMoa2012971.

DiNardo, C.D., Pratz, K.W., Letai, A., Jonas, B.A., Wei, A.H., Thirman, M., Arellano, M., Frattini, M.G., Kantarjian, H., Popovic, R., et al. (2018). Safety and preliminary efficacy of venetoclax with decitabine or azacitidine in elderly patients with previously untreated acute myeloid leukaemia: a non-randomised, open-label, phase 1b study. Lancet Oncol *19*, 216-228. 10.1016/S1470-2045(18)30010-X.

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

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol *34*, 184-191. 10.1038/nbt.3437.

Duwel, M., Welteke, V., Oeckinghaus, A., Baens, M., Kloo, B., Ferch, U., Darnay, B.G., Ruland, J., Marynen, P., and Krappmann, D. (2009). A20 negatively regulates T cell receptor signaling to NF-kappaB by cleaving Malt1 ubiquitin chains. J Immunol *182*, 7718-7728. 10.4049/jimmunol.0803313.

Eskens, F.A., Ramos, F.J., Burger, H., O'Brien, J.P., Piera, A., de Jonge, M.J., Mizui, Y., Wiemer, E.A., Carreras, M.J., Baselga, J., and Tabernero, J. (2013). Phase I pharmacokinetic and pharmacodynamic study of the first-in-class spliceosome inhibitor E7107 in patients with advanced solid tumors. Clin Cancer Res *19*, 6296-6304. 10.1158/1078-0432.CCR-13-0485.

Fennell, K.A., Vassiliadis, D., Lam, E.Y.N., Martelotto, L.G., Balic, J.J., Hollizeck, S., Weber, T.S., Semple, T., Wang, Q., Miles, D.C., et al. (2021). Non-genetic determinants of malignant clonal fitness at single-cell resolution. Nature. 10.1038/s41586-021-04206-7.

Ferrara, F., and Schiffer, C.A. (2013). Acute myeloid leukaemia in adults. Lancet *381*, 484-495. 10.1016/S0140-6736(12)61727-9.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic Acids Res *41*, D48-55. 10.1093/nar/gks1236.

Fong, C.Y., Gilan, O., Lam, E.Y., Rubin, A.F., Ftouni, S., Tyler, D., Stanley, K., Sinha, D., Yeh, P., Morison, J., et al. (2015). BET inhibitor resistance emerges from leukaemia stem cells. Nature *525*, 538-542. 10.1038/nature14888.

Ganzel, C., Sun, Z., Cripe, L.D., Fernandez, H.F., Douer, D., Rowe, J.M., Paietta, E.M., Ketterling, R., O'Connell, M.J., Wiernik, P.H., et al. (2018). Very poor long-term survival in past and more recent studies for relapsed AML patients: The ECOG-ACRIN experience. Am J Hematol *93*, 1074-1081. 10.1002/ajh.25162.

Giannakis, M., Mu, X.J., Shukla, S.A., Qian, Z.R., Cohen, O., Nishihara, R., Bahl, S., Cao, Y., Amin-Mansour, A., Yamauchi, M., et al. (2016). Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Rep *15*, 857-865. 10.1016/j.celrep.2016.03.075.

Gripp, K.W., Hopkins, E., Johnston, J.J., Krause, C., Dobyns, W.B., and Biesecker, L.G. (2011). Long-term survival in TARP syndrome and confirmation of RBM10 as the disease-causing gene. Am J Med Genet A *155A*, 2516-2520. 10.1002/ajmg.a.34190.

Gu, X., Tohme, R., Tomlinson, B., Sakre, N., Hasipek, M., Durkin, L., Schuerger, C., Grabowski, D., Zidan, A.M., Radivoyevitch, T., et al. (2021). Decitabine- and 5-azacytidine resistance emerges from adaptive responses of the pyrimidine metabolism network. Leukemia *35*, 1023-1036. 10.1038/s41375-020-1003-x.

Gui, J.F., Tronchere, H., Chandler, S.D., and Fu, X.D. (1994). Purification and characterization of a kinase specific for the serine- and arginine-rich pre-mRNA splicing factors. Proc Natl Acad Sci U S A *91*, 10824-10828. 10.1073/pnas.91.23.10824.

Han, C., Khodadadi-Jamayran, A., Lorch, A.H., Jin, Q., Serafin, V., Zhu, P., Politanska, Y., Sun, L., Gutierrez-Diaz, B.T., Pryzhkova, M.V., et al. (2022). SF3B1 homeostasis is critical for survival and therapeutic response in T cell leukemia. Sci Adv *8*, eabj8357. 10.1126/sciadv.abj8357.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515-1526. 10.1016/j.cell.2015.11.015.

Hashimoto, M., Saito, Y., Nakagawa, R., Ogahara, I., Takagi, S., Takata, S., Amitani, H., Endo, M., Yuki, H., Ramilowski, J.A., et al. (2021). Combined inhibition of XIAP and BCL2

drives maximal therapeutic efficacy in genetically diverse aggressive acute myeloid leukemia. Nature Cancer *2*, 340-356. 10.1038/s43018-021-00177-w.

Hong, D.S., Kurzrock, R., Naing, A., Wheler, J.J., Falchook, G.S., Schiffman, J.S., Faulkner, N., Pilat, M.J., O'Brien, J., and LoRusso, P. (2014). A phase I, open-label, single-arm, dose-escalation study of E7107, a precursor messenger ribonucleic acid (pre-mRNA) splicesome inhibitor administered intravenously on days 1 and 8 every 21 days to patients with solid tumors. Invest New Drugs *32*, 436-444. 10.1007/s10637-013-0046-5.

Huang, Y., Park, Y.C., Rich, R.L., Segal, D., Myszka, D.G., and Wu, H. (2001). Structural basis of caspase inhibition by XIAP: differential roles of the linker versus the BIR domain. Cell *104*, 781-790.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods *12*, 115-121. 10.1038/nmeth.3252.

Huppertz, I., Perez-Perri, J.I., Mantas, P., Sekaran, T., Schwarzl, T., Russo, F., Ferring-Appel, D., Koskova, Z., Dimitrova-Paternoga, L., Kafkia, E., et al. (2022). Riboregulation of Enolase 1 activity controls glycolysis and embryonic stem cell differentiation. Mol Cell *82*, 2666-2680 e2611. 10.1016/j.molcel.2022.05.019.

Ignatiadis, N., Klaus, B., Zaugg, J.B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods *13*, 577-580. 10.1038/nmeth.3885.

Jones, C.L., Stevens, B.M., D'Alessandro, A., Reisz, J.A., Culp-Hill, R., Nemkov, T., Pei, S., Khan, N., Adane, B., Ye, H., et al. (2018). Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. Cancer Cell *34*, 724-740 e724. 10.1016/j.ccell.2018.10.005.

Jones, C.L., Stevens, B.M., Pollyea, D.A., Culp-Hill, R., Reisz, J.A., Nemkov, T., Gehrke, S., Gamboni, F., Krug, A., Winters, A., et al. (2020). Nicotinamide Metabolism Mediates Resistance to Venetoclax in Relapsed Acute Myeloid Leukemia Stem Cells. Cell Stem Cell *27*, 748-764 e744. 10.1016/j.stem.2020.07.021.

Kamachi, M., Le, T.M., Kim, S.J., Geiger, M.E., Anderson, P., and Utz, P.J. (2002). Human autoimmune sera as molecular probes for the identification of an autoantigen kinase signaling pathway. J Exp Med *196*, 1213-1225. 10.1084/jem.20021167.

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. 10.1038/nature12634.

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

Kim, J., Woo, A.J., Chu, J., Snow, J.W., Fujiwara, Y., Kim, C.G., Cantor, A.B., and Orkin, S.H. (2010). A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell *143*, 313-324. 10.1016/j.cell.2010.09.010.

Konopleva, M., Pollyea, D.A., Potluri, J., Chyla, B., Hogdal, L., Busman, T., McKeegan, E., Salem, A.H., Zhu, M., Ricker, J.L., et al. (2016). Efficacy and Biological Correlates of Response in a Phase II Study of Venetoclax Monotherapy in Patients with Acute Myelogenous Leukemia. Cancer Discov *6*, 1106-1117. 10.1158/2159-8290.CD-16-0313.

Lachowiez, C.A., Loghavi, S., Furudate, K., Montalban-Bravo, G., Maiti, A., Kadia, T., Daver, N., Borthakur, G., Pemmaraju, N., Sasaki, K., et al. (2021). Impact of splicing mutations in acute myeloid leukemia treated with hypomethylating agents combined with venetoclax. Blood Adv *5*, 2173-2183. 10.1182/bloodadvances.2020004173.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25. 10.1186/gb-2009-10-3-r25.

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

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

Li, S., Garrett-Bakelman, F.E., Chung, S.S., Sanders, M.A., Hricik, T., Rapaport, F., Patel, J., Dillon, R., Vijay, P., Brown, A.L., et al. (2016). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat Med *22*, 792-799. 10.1038/nm.4125.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lindberg, M.F., and Meijer, L. (2021). Dual-Specificity, Tyrosine Phosphorylation-Regulated Kinases (DYRKs) and cdc2-Like Kinases (CLKs) in Human Disease, an Overview. Int J Mol Sci *22*. 10.3390/ijms22116047.

Martin Moyano, P., Nemec, V., and Paruch, K. (2020). Cdc-Like Kinases (CLKs): Biology, Chemical Probes, and Therapeutic Potential. Int J Mol Sci *21*. 10.3390/ijms21207549.

Metz, K.S., Deoudes, E.M., Berginski, M.E., Jimenez-Ruiz, I., Aksoy, B.A., Hammerbacher, J., Gomez, S.M., and Phanstiel, D.H. (2018). Coral: Clear and Customizable Visualization of Human Kinome Data. Cell Syst *7*, 347-350 e341. 10.1016/j.cels.2018.07.001.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res *41*, D64-69. 10.1093/nar/gks1048.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet *49*, 1779-1784. 10.1038/ng.3984.

Minn, A.J., Rudin, C.M., Boise, L.H., and Thompson, C.B. (1995). Expression of bcl-xL can confer a multidrug resistance phenotype. Blood *86*, 1903-1910.

Nechiporuk, T., Kurtz, S.E., Nikolova, O., Liu, T., Jones, C.L., D'Alessandro, A., Culp-Hill, R., d'Almeida, A., Joshi, S.K., Rosenberg, M., et al. (2019). The TP53 Apoptotic Network Is a Primary Mediator of Resistance to BCL2 Inhibition in AML Cells. Cancer Discov *9*, 910-925. 10.1158/2159-8290.CD-19-0125.

Prasad, J., Colwill, K., Pawson, T., and Manley, J.L. (1999). The protein kinase Clk/Sty directly modulates SR protein activity: both hyper- and hypophosphorylation inhibit splicing. Mol Cell Biol *19*, 6991-7000. 10.1128/MCB.19.10.6991.

Qian, W., Liang, H., Shi, J., Jin, N., Grundke-Iqbal, I., Iqbal, K., Gong, C.X., and Liu, F. (2011). Regulation of the alternative splicing of tau exon 10 by SC35 and Dyrk1A. Nucleic Acids Res *39*, 6161-6171. 10.1093/nar/gkr195.

Rathert, P., Roth, M., Neumann, T., Muerdter, F., Roe, J.S., Muhar, M., Deswal, S., Cerny-Reiterer, S., Peter, B., Jude, J., et al. (2015). Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. Nature *525*, 543-547. 10.1038/nature14898.

Riedl, S.J., Renatus, M., Schwarzenbacher, R., Zhou, Q., Sun, C., Fesik, S.W., Liddington, R.C., and Salvesen, G.S. (2001). Structural basis for the inhibition of caspase-3 by XIAP. Cell *104*, 791-800. 10.1016/s0092-8674(01)00274-4.

Rini, B.I., Plimack, E.R., Stus, V., Gafanov, R., Hawkins, R., Nosov, D., Pouliot, F., Alekseev, B., Soulieres, D., Melichar, B., et al. (2019). Pembrolizumab plus Axitinib versus Sunitinib

for Advanced Renal-Cell Carcinoma. N Engl J Med *380*, 1116-1127. 10.1056/NEJMoa1816714.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25. 10.1186/gb-2010-11-3-r25.

Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., and Doench, J.G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun *9*, 5416. 10.1038/s41467-018-07901-8.

Schwerk, C., and Schulze-Osthoff, K. (2005). Regulation of apoptosis by alternative pre-mRNA splicing. Mol Cell *19*, 1-13. 10.1016/j.molcel.2005.05.026.

Scott, A., Call, J.A., Chandana, S., Borazanci, E., Falchook, G.S., Bordoni, R., Richey, S., Starodub, A., Chung, V., Lakhani, N.J., et al. (2022). 451O Preliminary evidence of clinical activity from phase I and Ib trials of the CLK/DYRK inhibitor cirtuvivint (CIRT) in subjects with advanced solid tumors. Annals of Oncology *33*, S742-S743. 10.1016/j.annonc.2022.07.580.

Seiler, M., Yoshimi, A., Darman, R., Chan, B., Keaney, G., Thomas, M., Agrawal, A.A., Caleb, B., Csibi, A., Sean, E., et al. (2018). H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. Nat Med *24*, 497-504. 10.1038/nm.4493.

Shi, J., Wang, E., Milazzo, J.P., Wang, Z., Kinney, J.B., and Vakoc, C.R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol *33*, 661-667. 10.1038/nbt.3235.

Shi, J., Zhang, T., Zhou, C., Chohan, M.O., Gu, X., Wegiel, J., Zhou, J., Hwang, Y.W., Iqbal, K., Grundke-Iqbal, I., et al. (2008). Increased dosage of Dyrk1A alters alternative splicing factor (ASF)-regulated alternative splicing of tau in Down syndrome. J Biol Chem *283*, 28660-28669. 10.1074/jbc.M802645200.

Shimizu, T., Yonemori, K., Koyama, T., Katsuya, Y., Sato, J., Fukuhara, N., Yokoyama, H., Iida, H., Ando, K., Fukuhara, S., et al. (2022). A first-in-human phase I study of CTX-712 in patients with advanced, relapsed or refractory malignant tumors. Journal of Clinical Oncology *40*, 3080-3080. 10.1200/JCO.2022.40.16_suppl.3080.

Shiozaki, E.N., Chai, J., Rigotti, D.J., Riedl, S.J., Li, P., Srinivasula, S.M., Alnemri, E.S., Fairman, R., and Shi, Y. (2003). Mechanism of XIAP-mediated inhibition of caspase-9. Mol Cell *11*, 519-527. 10.1016/s1097-2765(03)00054-6.

Short, N.J., Konopleva, M., Kadia, T.M., Borthakur, G., Ravandi, F., DiNardo, C.D., and Daver, N. (2020). Advances in the Treatment of Acute Myeloid Leukemia: New Drugs and New Challenges. Cancer Discov *10*, 506-525. 10.1158/2159-8290.CD-19-1011.

Somervaille, T.C., Matheny, C.J., Spencer, G.J., Iwasaki, M., Rinn, J.L., Witten, D.M., Chang, H.Y., Shurtleff, S.A., Downing, J.R., and Cleary, M.L. (2009). Hierarchical maintenance of MLL myeloid leukemia stem cells employs a transcriptional program shared with embryonic rather than adult stem cells. Cell Stem Cell *4*, 129-140. 10.1016/j.stem.2008.11.015.

Srinivasula, S.M., Hegde, R., Saleh, A., Datta, P., Shiozaki, E., Chai, J., Lee, R.A., Robbins, P.D., Fernandes-Alnemri, T., Shi, Y., and Alnemri, E.S. (2001). A conserved XIAP-interaction motif in caspase-9 and Smac/DIABLO regulates caspase activity and apoptosis. Nature *410*, 112-116. 10.1038/35065125.

Sripayap, P., Nagai, T., Uesawa, M., Kobayashi, H., Tsukahara, T., Ohmine, K., Muroi, K., and Ozawa, K. (2014). Mechanisms of resistance to azacitidine in human leukemia cell lines. Exp Hematol *42*, 294-306 e292. 10.1016/j.exphem.2013.12.004.

Sugimoto, K., Toyoshima, H., Sakai, R., Miyagawa, K., Hagiwara, K., Ishikawa, F., Takaku, F., Yazaki, Y., and Hirai, H. (1992). Frequent mutations in the p53 gene in human myeloid leukemia cell lines. Blood *79*, 2378-2383.

Sykes, D.B., Kfoury, Y.S., Mercier, F.E., Wawer, M.J., Law, J.M., Haynes, M.K., Lewis, T.A., Schajnovitz, A., Jain, E., Lee, D., et al. (2016). Inhibition of Dihydroorotate Dehydrogenase Overcomes Differentiation Blockade in Acute Myeloid Leukemia. Cell *167*, 171-186 e115. 10.1016/j.cell.2016.08.057.

Ten Hacken, E., Valentin, R., Regis, F.F.D., Sun, J., Yin, S., Werner, L., Deng, J., Gruber, M., Wong, J., Zheng, M., et al. (2018). Splicing modulation sensitizes chronic lymphocytic leukemia cells to venetoclax by remodeling mitochondrial apoptotic dependencies. JCI Insight *3*. 10.1172/jci.insight.121438.

Tolcher, A., Babiker, H.M., Chung, V., Kim, E., Moser, J., Karim, R., Vandross, A., Sommerhalder, D., Scott, A.J., Fakih, M., et al. (2021). Abstract CT112: Initial results from a Phase 1 trial of a first-in-class pan-CDC-like kinase inhibitor (SM08502) with proof of mechanism in subjects with advanced solid tumors. Cancer Research *81*, CT112-CT112. 10.1158/1538-7445.Am2021-ct112.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111. 10.1093/bioinformatics/btp120.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods *13*, 508-514. 10.1038/nmeth.3810.

Wang, E., Lu, S.X., Pastore, A., Chen, X., Imig, J., Chun-Wei Lee, S., Hockemeyer, K., Ghebrechristos, Y.E., Yoshimi, A., Inoue, D., et al. (2019). Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. Cancer Cell *35*, 369-384 e367. 10.1016/j.ccell.2019.01.010.

Wang, E., Zhou, H., Nadorp, B., Cayanan, G., Chen, X., Yeaton, A.H., Nomikou, S., Witkowski, M.T., Narang, S., Kloetgen, A., et al. (2021). Surface antigen-guided CRISPR screens identify regulators of myeloid leukemia differentiation. Cell Stem Cell *28*, 718-731 e716. 10.1016/j.stem.2020.12.005.

Wang, Y., Gogol-Doring, A., Hu, H., Frohler, S., Ma, Y., Jens, M., Maaskola, J., Murakawa, Y., Quedenau, C., Landthaler, M., et al. (2013). Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation. EMBO Mol Med *5*, 1431-1442. 10.1002/emmm.201302663.

Witkiewicz, A.K., McMillan, E.A., Balaji, U., Baek, G., Lin, W.C., Mansour, J., Mollaee, M., Wagner, K.U., Koduru, P., Yopp, A., et al. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun *6*, 6744. 10.1038/ncomms7744.

Witkowski, M.T., Lee, S., Wang, E., Lee, A.K., Talbot, A., Ma, C., Tsopoulidis, N., Brumbaugh, J., Zhao, Y., Roberts, K.G., et al. (2022). NUDT21 limits CD19 levels through alternative mRNA polyadenylation in B cell acute lymphoblastic leukemia. Nat Immunol. 10.1038/s41590-022-01314-y.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med *23*, 703-713. 10.1038/nm.4333.

Zhang, H., Nakauchi, Y., Kohnke, T., Stafford, M., Bottomly, D., Thomas, R., Wilmot, B., McWeeney, S.K., Majeti, R., and Tyner, J.W. (2020). Integrated analysis of patient samples identifies biomarkers for venetoclax efficacy and combination strategies in acute myeloid leukemia. Nat Cancer *1*, 826-839. 10.1038/s43018-020-0103-x.

Zhou, Y., Han, C., Wang, E., Lorch, A.H., Serafin, V., Cho, B.K., Guttierrez Diaz, B.T., Calvo, J., Fang, C., Khodadadi-Jamayran, A., et al. (2020). Posttranslational regulation of the exon

skipping machinery controls aberrant splicing in leukemia. Cancer Discov. 10.1158/2159-8290.CD-19-1436.

Zuber, J., Radtke, I., Pardee, T.S., Zhao, Z., Rappaport, A.R., Luo, W., McCurrach, M.E., Yang, M.M., Dolan, M.E., Kogan, S.C., et al. (2009). Mouse models of human AML accurately predict chemotherapy response. Genes Dev *23*, 877-889. 10.1101/gad.1771409.

Zuber, J., Rappaport, A.R., Luo, W., Wang, E., Chen, C., Vaseva, A.V., Shi, J., Weissmueller, S., Fellmann, C., Taylor, M.J., et al. (2011a). An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. Genes Dev *25*, 1628-1640. 10.1101/gad.17269211.

Zuber, J., Shi, J., Wang, E., Rappaport, A.R., Herrmann, H., Sison, E.A., Magoon, D., Qi, J., Blatt, K., Wunderlich, M., et al. (2011b). RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. Nature *478*, 524-528. 10.1038/nature10334.

Chapter 5. DUX4 IS A COMMON DRIVER OF IMMUNE EVASION AND IMMUNOTHERAPY FAILURE IN METASTATIC CANCERS

Jose Mario Bello Pineda and Robert K. Bradley

5.1 ABSTRACT

Cancer immune evasion underlies checkpoint immunotherapy failure in most metastatic cancer patients. Our previous findings in a subset of primary cancers identified *DUX4*, a transcription factor of the pre-implantation embryo typically silent in somatic tissues, as an underappreciated suppressor of tumor cell-intrinsic interferon-γ signaling. We extended our analyses to several advanced cancer cohorts treated with immune checkpoint inhibition (ICI) therapy to show that the frequency of *DUX4* reactivation is exaggerated in metastasis compared to the early-stage equivalent. We evaluated a large cohort of advanced urothelial carcinoma patients treated with *PD-L1* blockade and show that metastatic *DUX4* expression is correlated with downregulation of antigen processing and presentation, signatures of increased immune cell exclusion, and decreased objective response to ICI. *DUX4* expression was also linked to significantly reduced overall survival after adjusting for other genetic, clinical, and demographic patient covariates. Our data indicate that *DUX4* reactivation is a common mechanism for checkpoint immunotherapy resistance in metastatic cancers.

5.2 INTRODUCTION

Immune checkpoint inhibition (ICI) therapy utilizes immunomodulatory monoclonal antibodies to stimulate patient anti-tumor immune responses. Blockade of the T cell co-inhibitory receptors,

such as *CTLA4* and the *PD-1*/*PD-L1* axis, has achieved major success in the treatment of diverse metastatic cancers compared to first-line chemotherapy (Doki et al., 2022; Hellmann et al., 2019; Klein et al., 2020; Larkin et al., 2019; Motzer et al., 2020; Stein et al., 2022). However, majority of advanced cancer patients fail to respond to ICI due to de-novo or acquired resistance, the bases of which remain incompletely understood.

Diverse mechanisms modulate sensitivity and resistance to immune checkpoint inhibition (Kalbasi & Ribas, 2020). These mechanisms include defects in MHC-I-mediated antigen presentation due to loss of *B2M* or HLA (Grasso et al., 2018; Lee et al., 2020; McGranahan et al., 2016; Sade-Feldman et al., 2017; Sucker et al., 2014; Wolf et al., 2019), *PTEN* and *LSD1* inactivation which sensitize tumor cells to type I interferon signaling (S. Li et al., 2016; Peng et al., 2016; Sheng et al., 2018), T cell dysfunction (Jiang et al., 2018), presence of specific T cell populations in the tumor microenvironment (Gide et al., 2019), and active WNT–β-catenin signaling (Spranger et al., 2015). *MAPK* signaling in *BRAF*-mutated melanomas (Ascierto et al., 2019; Ebert et al., 2016; Riaz et al., 2017; Ribas et al., 2019; Sullivan et al., 2019) and *CDK4*/*CDK6* activity (Deng et al., 2018; Goel et al., 2017; Jerby-Arnon et al., 2018; Schaer et al., 2018) have also been implicated in reduced ICI efficacy, and combination treatment with a MAPK/CDK inhibitor improves response to checkpoint blockade.

Tumor cell-intrinsic interferon-gamma (IFN-γ) signaling is particularly important in anti-tumor immunity. This pathway induces expression of genes involved in MHC-I-mediated antigen processing and presentation which include the *TAP1*/*TAP2* transporter, components of the immunoproteasome, HLA proteins, and *B2M*, (Alspach et al., 2019). Thus, suppression of IFN-γ activity promotes tumor immune evasion and decreased *CD8*+ T cell activation. Indeed, decreased ICI efficacy was observed in patients with tumors harboring inactivating mutations in IFN-γ

pathway genes such as *JAK1* and *JAK2* (Gao et al., 2016; Nguyen et al., 2021; Sucker et al., 2017; Zaretsky et al., 2016). Similarly, a recent study reported a splicing-augmenting mutation in *JAK3*, linked to decreased *JAK3* expression levels, as a potential mechanism of resistance in a metastatic melanoma patient treated with anti-*PD-1* and anti-*CTLA4* combination therapy (Newell et al., 2022).

Recent work from our group showed that the embryonic transcription factor *DUX4* was expressed in a small subset of primary tumors and promoted downregulation of tumor IFN-γ activity (Chew et al., 2019). In its native context, *DUX4* initializes human zygotic genome activation. *DUX4* transcription peaks at the 4-cell stage of the cleavage embryo before immediate silencing via epigenetic repression of the D4Z4 repeat array (de Iaco et al., 2017; Hendrickson et al., 2017; Himeda & Jones, 2019; Sugie et al., 2020; Whiddon et al., 2017). Aside from select sites of immune privilege, *DUX4* typically remains somatically inactive (Das & Chadwick, 2016; Snider et al., 2010). *DUX4* reactivation in these cancers suppresses MHC-I-mediated antigen presentation (Chew et al., 2019). We hypothesized that DUX4 expression might be more common in the setting of metastatic disease, where immune evasion is particularly important. While our previous study implicated *DUX4* expression in resistance to checkpoint immunotherapy in advanced melanoma, it relied on small and thus statistically underpowered ICI cohorts. In our current study, we analyzed several larger cohorts to test the robustness and importance of *DUX4*'s association with ICI failure.

5.3 RESULTS

5.3.1 *DUX4 is reactivated broadly across primary and metastatic cancers*

To assess the prevalence of *DUX4*-expressing human malignancies, we performed a large-scale analysis of publicly available RNA-seq data across diverse cancer types (**Fig. 1A, Fig. S1A**). We

found that *DUX4* reactivation is a common feature observed across early and advanced stages, with cancers displaying physiological to supraphysiological *DUX4* expression levels. Interestingly, a higher proportion of metastatic cancers re-express *DUX4*— and tend to have higher gene expression levels— compared to their primary cancer counterparts (**Fig. 1B-C**).

We sought to determine if the *DUX4* transcripts in metastatic cancers harbor the entire coding sequence as *DUX4* truncations due to genomic rearrangements are frequent oncogenic drivers, such as in undifferentiated round cell sarcomas (CIC-DUX4 oncoprotein) and adolescent B-cell acute lymphoblastic leukemia (ALL). We aligned RNA-seq reads to the *DUX4* cDNA sequence and examined read coverage over the open reading frame. Resembling the cleavage stage embryo and *DUX4*-reactivated primary cancers, *DUX4*-positive metastatic tumors transcribe the full-length coding region. In contrast, B-cell ALL exhibited the expected C-terminal truncation due to *DUX4* fusion with the *IGH* locus (**Fig. 1D**).

Since *DUX4* is typically silent in most healthy contexts outside the cleavage stage embryo (Das & Chadwick, 2016; Snider et al., 2010), we investigated if artifacts related to sequencing and sample processing could account for the observed signal. We analyzed RNA-seq data from flash-frozen and formalin-fixed metastatic tumor samples wherein paired poly-A and hybrid probe capture sequencing library preparations were performed (Robinson et al., 2015b). *DUX4* expression is quantifiable in both processing types. Further, *DUX4* transcript levels in majority of the sequenced samples were higher in poly-A-enriched sequencing than the analogous measurements from hybrid capture (**Fig. S1B-C**). Of note, the primary cancer samples (The Cancer Genome Atlas) were sequenced via poly-A capture and prior work from our group determined that these tumors express the *DUX4* isoform containing the consensus polyadenylation signal (Chew

et al., 2019). Taken together, these observations are consistent with expression of a polyadenylated *DUX4* transcript in both primary and metastatic cancers.

5.3.2 *DUX4 expression is associated with immune cell exclusion*

To assess the possible downstream consequences of *DUX4* expression in metastatic cancer, we analyzed the pre-treatment transcriptome profiles of metastatic cancer patients enrolled in immune checkpoint inhibition (ICI) trials. As a representative example, we analyzed data from the patients of the IMvigor210 phase 2 trial, a large advanced urothelial carcinoma cohort treated with anti-*PD-L1* (atezolizumab) therapy (Balar et al., 2017; Mariathasan et al., 2018; Rosenberg et al., 2016).

We examined the gene expression consequences associated with *DUX4* reactivation in advanced bladder cancer. Specifically, we performed differential gene expression analyses on the individuals stratified according to tumor *DUX4* expression status. Gene Ontology (GO) network analyses on the upregulated genes in *DUX4*-positive cancers identified multiple clusters of development-associated terms, consistent with the known role of *DUX4* in early embryogenesis (**Fig. S2A;** de Iaco et al., 2017; Hendrickson et al., 2017; Sugie et al., 2020; Whiddon et al., 2017). By contrast, we found a single network associated with downregulated genes: GO terms corresponding to humoral or cell-mediated immunity (**Fig. 2A**). Using an IFN-γ gene signature predictive of response to blockade of the *PD-1*/*PD-L1* axis, we found that *DUX4*-expressing cancers have statistically lower levels of IFN-γ activity (**Fig. S2B**; Ayers et al., 2017). Consistent with IFN-γ suppression, we observed extensive downregulation of genes involved in anti-tumor immunity such as those involved in MHC-I-dependent antigen presentation and T cell activation, checkpoint proteins, and chemokines involved in effector T cell recruitment. *DUX4*-expression was also correlated with suppression of genes critical for MHC-II-mediated antigen presentation,

namely: MHC-II isotypes (*HLA–DP/DQ/DR)*, *HLA-DM* and *HLA-DO*, and the invariant chain (*CD74*) (Roche & Furuta, 2015). MHC-II gene expression is regulated by the transactivator *CIITA* via a conserved SXY-module present in the promoter regions of these genes. *CIITA* is induced by IFN-γ and is also conspicuously downregulated in *DUX4*-expressing tumors (**Fig. 2B**; (Glimcher & Kara, 1992; Masternak et al., 2000; Steimle et al., 1993, 1994). These analyses suggest that *DUX4* expression in the metastatic context induces an immunosuppressive gene expression program, concordant with its established function in inhibiting JAK-STAT signaling in primary cancers (Chew et al., 2019).

We hypothesized that *DUX4* expression in these cancers will generate related transcriptomic signals consistent with *CD8*+ T cell exclusion from the tumor. We assessed this using an effector *CD8*+ T cell transcriptomic signature developed from initial studies of the IMvigor210 phase 2 trial (Balar et al., 2017; Rosenberg et al., 2016). *DUX4*-expressing cancers had lower measures of the gene signature, consistent with decreased *CD8*+ T cell infiltration into the tumor (**Fig. 2C**). We also investigated the possible effects of *DUX4* expression on the exclusion of other immune cell types using gene signatures developed from The Cancer Genome Atlas (Danaher et al., 2017). In these analyses, we recapitulated the observation of lower *CD8*+ T cell signature associated with *DUX4* positivity (**Fig. S2C)**. In addition, we observed patterns consistent with widespread immune cell exclusion from the tumor microenvironment (**Fig. S2D**).

Defects in chemokine signaling could partially account for the observed *DUX4*-associated decrease in immune gene signature measurements. To test this hypothesis, we examined expression of chemokines involved in immune cell recruitment. In *DUX4*-expressing cancers, we observed lower mRNA levels of *CXCL9* and *CXCL10*, chemokines which recruit T cells to the tumor site (**Fig. 2D**, **Fig. 2E**; Nagarsheth et al., 2017). Utilizing a chemokine signature associated

with host immune response to solid tumors, we observed that *DUX4* expression was correlated with broad inhibition of chemokine signaling, beyond T cell-associated signals (**Fig. S2E**; Coppola et al., 2011).

We directly assessed the correlation of *DUX4* expression to immune cell exclusion by examining *CD8+* T cell abundance in the tumor microenvironment, measured by immunohistochemistry (IHC) on formalin fixed paraffin embedded (FFPE) patient tumor sections. We verified that *DUX4* expression in the advanced urothelial carcinoma tumors was associated with an immune exclusion phenotype: a higher proportion of *DUX4+* tumors exhibit either an immune-excluded or immune-desert phenotype compared to malignancies where *DUX4* is silent (**Fig. 2F**, **Fig. S2F**). We similarly examined the correlation of *DUX4* expression status with *PD-L1* levels in the tumor and immune compartments, quantified via IHC. We determined that *DUX4* expression was associated with a significant decrease in *PD-L1* levels on both tumor and host immune cells, consistent with *DUX4*-induced suppression of IFN-γ signaling (**Fig. 2G**, **Fig. S2G, Fig. 2H**, **Fig. S2H**). *PD-L1* expression on immune cells such as dendritic cells and macrophages modulate anti-tumor immune suppression and response to ICI in *in vivo* mouse models (Lau et al., 2017; Lin et al., 2018; Noguchi et al., 2017). Importantly, *PD-L1* levels on immune cells are correlated with response to ICI in clinical trials (Powles et al., 2014; Rosenberg et al., 2016).

5.3.3 *DUX4 expression is correlated with poor response in metastatic bladder cancer immune checkpoint inhibition*

Given the association between tumor *DUX4* expression and suppression of anti-tumor immune response, we next sought to understand if tumor *DUX4* expression conferred changes to patient overall survival during *PD-L1* inhibition. *DUX4* expression was associated with a significant decrease in objective response rates (RECIST) (**Fig. 3A**). As expected, higher tumor mutational

burden (TMB) was associated with improved survival outcomes in this cohort (**Fig. 3B**). *DUX4* expression on the other hand was correlated with a significant reduction in median overall survival (**Fig. S3A**). We attempted to control for the possible confounding effects of TMB on the *DUX4* signal by removing the bottom quartile of patients, those with the lowest number of missense mutations in their tumors. *DUX4* expression was associated with statistically lower survival rates in this cohort, even after controlling for TMB (**Fig. 3C**). Our results suggest that *DUX4* expression modulates clinical response to ICI in advanced cancer patients. *TGFB1* expression has been linked to diminished clinical response to ICI in advanced urothelial carcinoma patients (Mariathasan et al., 2018). However, differences in *TGFB1* expression did not correlate with augmented patient survival after controlling for TMB in our analyses (**Fig. S3B-C**).

5.3.4 *Risk assignments are improved with DUX4 expression*

We used Cox Proportional Hazards (PH) regression to quantify the effect of the various clinical, demographic, and molecular features on risk of death during ICI. In the multivariate Cox PH regression context, which controls for the confounding effects of all other covariates simultaneously, we observed that TMB was positively associated with survival [hazard ratio (HR) = 0.14] as expected. On the other hand, *DUX4* expression, Eastern Cooperative Oncology Group Performance Status (ECOG PS) > 0, and previous administration of platinum chemotherapy were correlated with increased risk, or shorter survival (**Fig. 4A**, **Table 1**). In particular, *DUX4*-positivity was associated with dramatically worse survival, an increase of more than 24-fold in risk compared to *DUX4*-negative status (HR = 3.19) (**Fig. S3B-C**). Of note, we found that differences in tumor *TGFB1* expression did not augment risk consistent with our TMB-adjusted KM analyses (**Table S1**).

We next investigated if *DUX4* expression status carried added value as a predictor over routinely collected clinical and molecular information. We focused on the variables with significant hazard ratios under both the univariate and multivariate regression settings: *DUX4* expression status, TMB, ECOG PS, and history of platinum chemotherapy. We employed goodness of fit measurements which compare the observed data to expectations from Cox PH models created using various combinations of the covariates. In these analyses, we observed a quantifiable improvement in data-model congruence with the addition of *DUX4* expression status (**Fig. 4B**, **Fig. S4A-B**). Additionally, we measured statistically significant differences in the likelihoods of the reduced models (without *DUX4* expression as a predictor) when compared to the full model (employs all covariates) (**Table 2**). Taken together, these analyses suggest that *DUX4* expression status is an informative predictor of risk under ICI treatment.

We evaluated the utility *DUX4* reactivation status for pre-treatment risk assignment in predicting patient response to ICI. We trained full and reduced Cox PH models on randomly sampled patients (training set, 70% of the cohort) and quantified their respective risk scores. A reference risk score per model was computed as the median score across the training set and was used to ascribe patients into low- vs. high-risk groups. Using these models, we quantified risk scores on the individuals excluded from model construction (test set, 30% of the patients), and similarly assigned patients into low- or high-risk groups based on the training set reference score. By empirically quantifying survival of the two risk groups using KM estimation, we found that the full model stratifies patients in an informative manner, appropriately discriminating patients with longer vs. shorter survival times (**Fig. 4C**, **Fig. S4C-D**). The time-dependent Brier score, which measures survival prediction accuracy at specific timepoints, of the full and reduced models also

demonstrates improvement in model performance associated with the addition of *DUX4* expression status **(Fig. S4E**).

5.3.5 *DUX4 expression impedes response to ICI after controlling for other clinical characteristics*

We used a Random Survival Forest (RSF) model to quantify the effect of *DUX4* expression on survival in ICI-treated advanced urothelial carcinoma patients (Ishwaran et al., 2008). The Random Survival Forest (RSF) is a machine learning ensemble, an extension of the Random Forest algorithm for right-censored data (Breiman, 2001). It can provide accurate estimates of risk and survival probability at definite times by aggregating predictions from a multitude of base learners (survival trees) (Ishwaran et al., 2008). RSFs have been successfully used to study time-to-event problems in medicine, including measurement of variable importance (S. Dietrich et al., 2016; Hsich et al., 2019; Ishwaran et al., 2009; O'Brien et al., 2021; Semeraro et al., 2011). We utilized the RSF model to address potential limitations of our Cox PH analyses. First, the RSF model is fully non-parametric and as such does not operate under the Cox PH assumptions: a constant relative hazard between strata over time (proportional hazards), a linear relationship between the predictors and the log hazard, and the unspecified baseline hazard function. Second, the RSF model can compute estimates of absolute risk and survival probability over time independent of a reference, unlike relative risk models such as Cox PH (Ishwaran et al., 2008).

We used all available molecular, clinical, and demographic covariates to grow an RSF. We randomly selected 70% of the patients to grow the forest, with the resulting model having an Out-of-Bag (OOB) error of 38.4%. The OOB error approaches the leave-one-out cross-validation error, once the error stabilizes with increasing number of trees, and is characterized as an unbiased estimate of the model's true prediction error (Breiman, 2001; Hastie et al., 2009). In some

instances, the OOB error provides overestimates and some reports have recommended treating it as an upper bound (Bylander, 2002; Janitza & Hornung, 2018; Mitchell, 2011). Thus, we measured the RSF model's test error using a holdout set (the remaining 30% of the cohort) excluded from training. The RSF model recorded a test error of 32.6% illustrating an appropriate fit (**Fig. S5A**). Our error measurements are comparable to Ishwaran, et al. (2008), suggesting our model can be used for inference purposes. Further, the time-dependent Brier score of the RSF model on the training and test sets confirms informative survival prediction (**Fig. S5B**).

The RSF model predicted worse survival outcomes in patients with *DUX4*-expressing cancers compared to their *DUX4*-silent counterparts. These predictions were mirrored in the test dataset, illustrating robustness of the model (**Fig. 5A**). Using time-dependent Receiver Operating Characteristic (ROC) curve analyses, we identified the time range for which the RSF predictive performance is statistically divergent from random guessing: approximately 6 to 20 months (**Fig. S5C**). In this window we measured significant survival differences between patients with *DUX4*+ and *DUX4*- tumors. We highlighted the model's performance at predicting 1-year and 1.5-year survival, typical timepoints of clinical interest. For these times, the RSF appropriately discriminates patient death and survival (**Fig. S5C)**. Examining the absolute effects of *DUX4* expression on survival, the RSF model predicted an approximately 20% decrease in 1-year and 1.5-year survival probabilities in patients with *DUX4*-reactivated cancers (**Fig. 5B).**

We sought to determine the importance of *DUX4* expression status relative to the other covariates in the RSF model. We measured feature importance using estimated Shapley values, which quantify the marginal contribution of each variable to the RSF prediction (Lundberg & Lee, 2017; Maksymiuk et al., 2020; Shapley, 1953; Štrumbelj & Kononenko, 2014). Specifically, Shapley values measure variable contributions to predictions at the level of each patient.

Contributions to the overall performance of the RSF model can be assessed by examining the aggregated summary: the average of the absolute Shapley values for a predictor across the patient cohort. We estimated Shapley values associated with predicting ensemble mortality, the RSF risk estimate. In these analyses, ECOG PS had the largest contribution, followed by TMB and *DUX4* expression (**Fig. S5E**). We validated these feature rankings through the use of two independent metrics. The first metric was permutation importance, which quantifies the change in prediction error associated with permutation of a variable's data; important covariates will record large deviations from the original predictions (Breiman, 2001; Ishwaran, 2007). The second measure employed was minimal depth, a measure of the variable-node-to-root-node distance within the survival trees of the RSF; important variables tend to have smaller minimal depth values as they are typically used for earlier decision splits (Ishwaran et al., 2010, 2011). Feature contributions measured using permutation importance and minimal depth were consistent with the Shapley-based assignments, importantly identifying *DUX4* expression as an important contributor to patient survival outcomes (**Fig. S5E**). We investigated time-dependent changes in variable importance by estimating Shapley values associated with predicting survival probability at distinct time points along the observation window. Interestingly, we observed the strong dependence on ECOG PS for predicting survival at early timepoints under this paradigm. Additionally, the importance of *DUX4* expression rises at later time points (**Fig. 5C**). Taken together, we found that diverse variable importance measures converge on identifying *DUX4* as a major contributor to patient survival prediction.

We sought to quantify the effect of *DUX4* expression on survival predictions after controlling for the effects of the other covariates. With Shapley dependence plots, which allows visualization of the marginal effects of a variable on the predicted outcome, we measured the

expected negative correlation between TMB and mortality (**Fig. S5F**; Lundberg et al., 2020). We performed a similar dependence analysis on *DUX4* expression and observed a clear separation of positive and negative Shapley values based on *DUX4*-positive and -negative status, respectively. These results signify an increase in predicted risk of death associated with *DUX4* expression (**Fig. S5E**). To quantify the effects of TMB and *DUX4* expression in the appropriate risk units (expected number of deaths), we utilized partial dependence as an alternative way to represent mortality predictions as a function of these variables, marginalized over the other predictors in the data (Friedman, 2001). Specifically, the average model predictions across the individuals in the cohort are calculated over the unique predictor values. The marginal effects of TMB and *DUX4* expression measured via partial dependence mirror the results of the Shapley dependence analyses. Patients with the lowest mutational burden exceed the individuals with the highest TMB by approximately 20 expected deaths on average. Further, we measured an increase in the number of predicted deaths associated with *DUX4*-positivity by approximately 16, over *DUX4*-negative status (**Fig. S5F-G**). We then extended the partial dependence analyses to survival probability predictions over time. In this paradigm, we similarly observed that higher TMB was correlated with increased survival probability, more pronounced at later times (**Fig. 5D**). *DUX4* expression was correlated with poorer survival outcomes, with a 1-year and 1.5-year survival difference of 20.7% and 19.2% between patients with *DUX4*+ and *DUX4*- tumors, respectively. Strikingly, our analyses measure a difference of at least 12.5 months in median survival between the *DUX4*+ and *DUX4*- strata (**Fig. 5E**). Our analyses demonstrate a significant decrease in survival attributable to reactivation of *DUX4* in advanced cancers.

5.4 DISCUSSION

Our study demonstrates that *DUX4* re-expression is a common feature of metastasis and important driver of immune evasion. While the mechanism governing *DUX4* de-repression in cancer remains to be elucidated, we show that *DUX4* expression in the metastatic context was associated with suppression of anti-tumor immunity, mirroring our observations in primary cancers (Chew et al., 2019). The immunomodulatory function of *DUX4* in metastatic cancer is consistent with biochemical studies from a recent preprint detailing the physical interaction between the *DUX4* and *STAT1* proteins. Specifically, *DUX4* sequesters phosphorylated *STAT1* from the promoters of IFN-γ-stimulated genes and further inhibits transcription by interfering with the recruitment of RNA polymerase II. This inhibitory function is dependent on the conserved (L)LxxL(L) motif in the C-terminal domain of *DUX4* (Spens et al., 2022). Accordingly, our analyses show that the *DUX4* transcript in the metastatic context contains the full-length coding region. *DUX4* may also act to suppress the immune response outside MHC-I-associated pathways. We found broad down-regulation of chemokines, transcriptomic signatures consistent with exclusion of other immune cells, and inhibition of MHC-II-mediated antigen presentation which is involved in natural killer cell-mediated cytotoxicity (Johnson et al., 2020). Our results demonstrate that *DUX4* re-expression is an underappreciated contributor ICI resistance through modulation of IFN-γ-mediated antigen presentation. The prognostic value of IFN-γ activity (Grasso et al., 2020; Newell et al., 2022) and its non-redundancy relative to TMB in terms of influencing ICI response is widely appreciated (Cristescu et al., 2018; Newell et al., 2022; Rozeman et al., 2021). Recent studies examining cutaneous melanoma have catalogued somatic mutations in IFN-γ pathway genes which modulate anti-tumor immunity. Their estimates of the frequency of cancers which employ IFN-γ-based

mechanisms for ICI escape, measurements performed with *DUX4* excluded, likely represent underestimations (Gao et al., 2016; Nguyen et al., 2021; Sucker et al., 2017).

Our analyses identify ECOG PS and history of treatment with platinum chemotherapy as factors that modulate response to ICI in advanced urothelial carcinoma patients. First, our data suggest that increasing ECOG PS is associated with reduced response to ICI, consistent with the original study. In these patients, the subgroup of individuals with ECOG PS = 2 (n = 24) had a median overall survival of 8.1 months, less than the patient subset with ECOG PS < 2 (n = 35) whose median survival was not reached during the observation period (Balar et al., 2017). Other studies have similarly reported poorer outcomes associated with ICI treatment in patients with high ECOG PS (Chalker et al., 2022; Krishnan et al., 2022; Petrillo et al., 2020; Sehgal et al., 2021). These results, in addition to our observation that ECOG PS drives predictions for survival at earlier times, possibly indicate that patients with higher degrees of disability are more likely to carry comorbidities that predispose them to adverse effects associated with ICI treatment. Second, our analyses suggest that pre-treatment with platinum chemotherapy may blunt patient response to PD-L1 blockade compared to their treatment-naïve counterparts. Results from trials involving advanced urothelial carcinoma patients with previous exposure to cisplatin or carboplatin regimens demonstrated sustained benefit from second-line atezolizumab (Bellmunt et al., 2017; Necchi et al., 2017; Rosenberg et al., 2016) resulting in its approval by the U.S. Food and Drug Administration (USFDA, 2016). However, a follow-up phase 3 trial showed that while atezolizumab was associated longer duration of response and led to fewer adverse events, it was not superior to chemotherapy (vinflunine, paclitaxel, or docetaxel) alone (Powles et al., 2018). As a result, the indication for atezolizumab in patients previously treated with platinum was recently withdrawn by the manufacturer (Roche, 2021). Interestingly, platinum chemotherapy affects

response to ICI in other treatment paradigms. As a first-line therapy, atezolizumab and pembrolizumab combined with platinum chemotherapy was associated with significantly higher overall survival metastatic non-small-cell lung cancer (Gandhi et al., 2018; West et al., 2019). In advanced urothelial carcinoma, combination of ICI to platinum chemotherapy was associated with improvement in progression-free survival (Galsky et al., 2020) but not overall survival (Galsky et al., 2020; Powles et al., 2021). As a maintenance therapy, avelumab after gemcitabine plus platinum chemotherapy significantly improved overall survival (Powles et al., 2020).

Our results may have profound consequences for ICI treatment. First *DUX4* expression may promote patient resistance in a wide array of ICI modalities. Our previous work showed *DUX4* expression is associated with resistance to anti-*CTLA4* and anti-*PD-1* therapies (Chew et al., 2019). In this study, we comprehensively demonstrate that *DUX4* modulates patient response to *PD-L1* blockade. We also report that DUX4 expression in metastasis results in downregulation of *TIGIT* (Zhang et al., 2018) and other immune checkpoints whose interception are currently under clinical investigation: *HAVCR2/TIM3* (NCT02608268; Dixon et al., 2021) and *LAG3* (NCT02658981; Amaria et al., 2022; Tawbi et al., 2022). Second, the pervasive reactivation of *DUX4* in all metastatic cohorts we examined exhibits its potential as a pan-cancer biomarker. We show that binary categorization of patients according to *DUX4* expression status was sufficient to stratify patients according to ICI response, which could translate to clinical screening whose results are binarized as well such as through IHC using anti-DUX4 antibodies. Our data motivate the investigation of the need for additional randomized trial data from diverse metastatic cancer cohorts to definitively evaluate DUX4's ICI-relevant prognostic value.

5.5 FIGURES

Figure 1

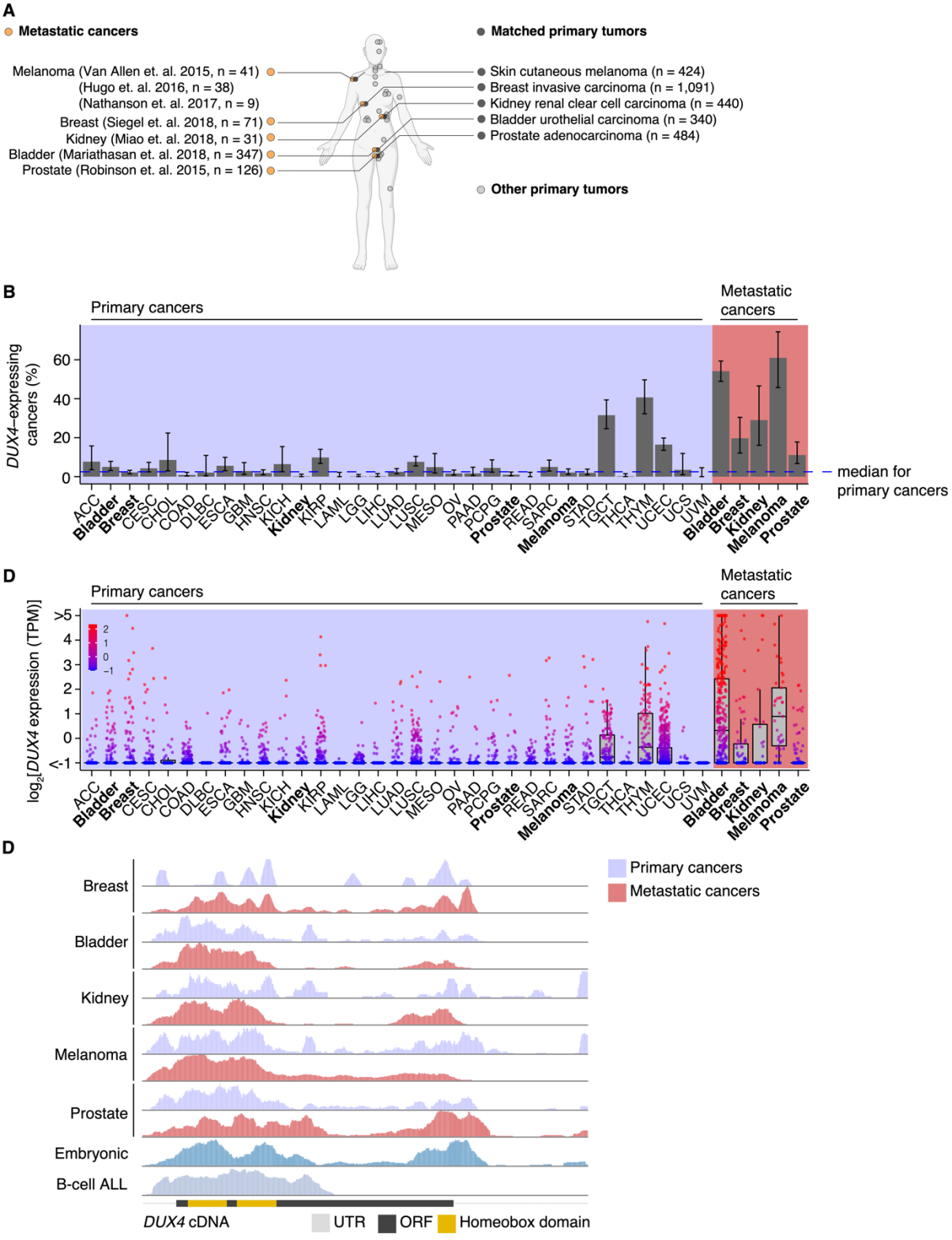


Figure 1. DUX4 is re-expressed as a full-length transcript in most primary and metastatic cancers.
(**A**) Matched primary (gray, The Cancer Genome Atlas) and metastatic (orange) cancer datasets analyzed in our study.
(**B**) The proportion of *DUX4*-expressing cancers in primary (purple shading) and metastatic (red shading) cancers. The blue line indicates the median over the primary cancer cohorts. The 95% confidence interval was estimated via a two-sided proportion test.
(**C**) DUX4 expression values (TPM, transcripts per million) in the primary (purple shading) and metastatic (red shading) cancer cohorts analyzed in our study.
(**D**) Representative RNA-seq coverage plots from primary and metastatic cancers for reads mapping to the *DUX4* cDNA. Open reading frame (ORF, black rectangle); UTR (untranslated region, gray line); Homeobox domains (yellow rectangles).

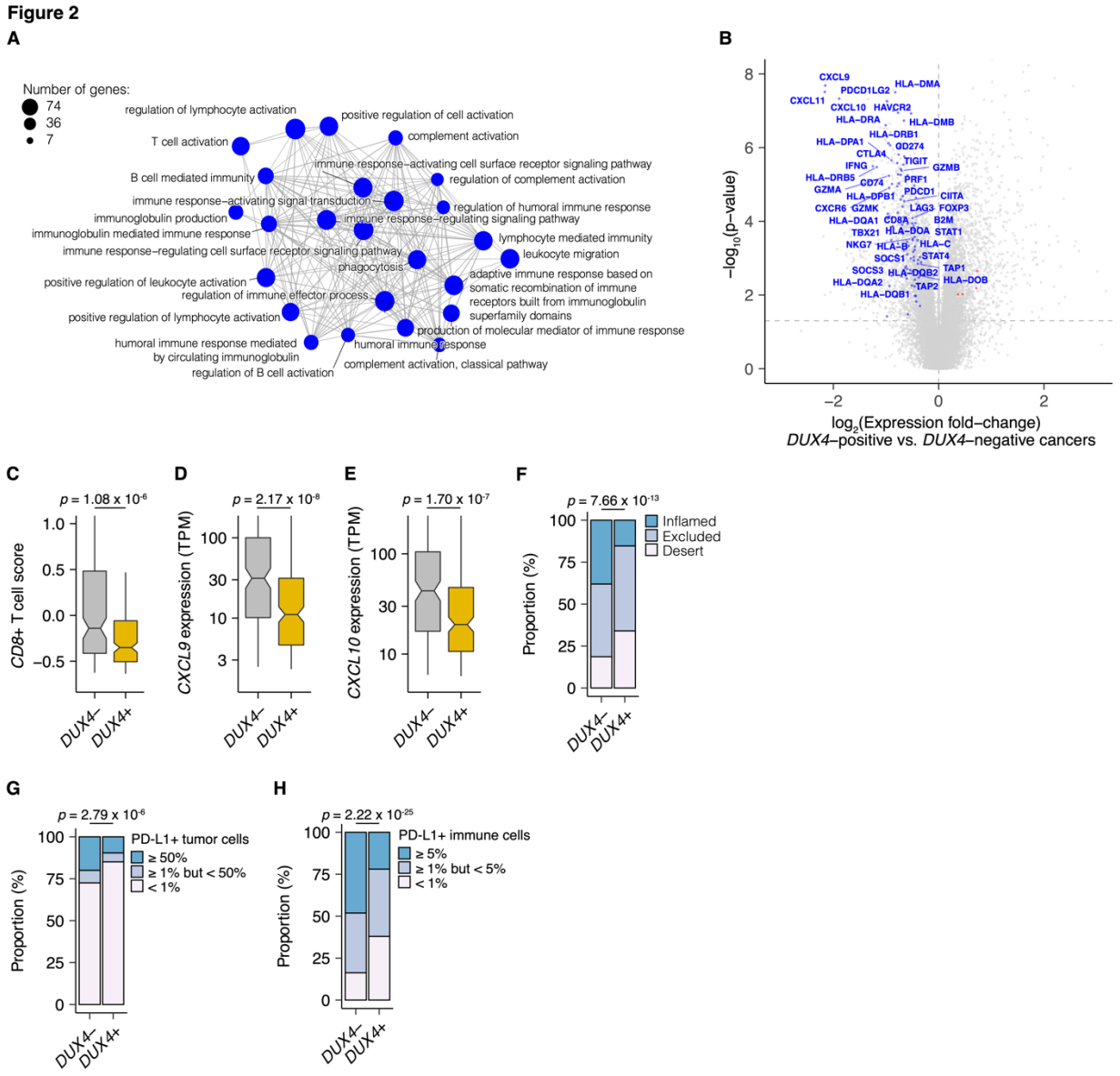


Figure 2. DUX4 expression in advanced cancers is associated with inhibition of host anti-tumor immunity

(**A**) Gene Ontology (GO) enrichment network analysis of *DUX4*-downregulated genes compared against the set of coding genes. Differentially expressed genes were identified from the comparison of advanced urothelial carcinoma tumors with high (> 1 TPM) vs. low (≤ 1 TPM) *DUX4* expression. The nodes and node sizes correspond to significantly enriched GO terms (Benjamini-Hochberg-adjusted *p*-value < 0.05) and the number of *DUX4*-downregulated genes in each, respectively. The edges connecting nodes correspond to terms with common genes.

(**B**) Downregulated (blue) and upregulated (red) anti-tumor immunity genes in tumors with *DUX4*-positive (> 1 TPM) vs. -negative (≤ 1 TPM) advanced urothelial carcinomas.

(**C**) Effector *CD8*+ T cell score, defined as the mean of the z-score normalized gene expression values in the signature (Mariathasan et al., 2018) for *DUX4*+/- tumors. The *p*-value was estimated via a Mann-Whitney *U* test.
(**D**) *CXCL9* expression for *DUX4*+/- tumors. The *p*-value was estimated via a Mann-Whitney *U* test.
(**E**) As in (D), but illustrating *CXCL10* expression.
(**F**) Proportion of immune phenotypes in *DUX4*+/- cancers. The phenotypes were based on the *CD8*+ T cell abundance and degree of tumor infiltration determined by anti-*CD8* staining of tumor FFPE sections in the original study (Mariathasan et al., 2018). The *p*-value was estimated via a multinomial proportion test.
(**G**) *PD-L1*-expression on tumor cells stratified by *DUX4* expression status measured by immunohistochemistry in the original study. The samples were categorized based on the percentage of *PD-L1*-positive tumor cells. The *p*-value was estimated via a multinomial proportion test.
(**H**) As in (G), but *PD-L1* staining on tumor-infiltrating immune cells (lymphocytes, macrophages, and dendritic cells) is represented.

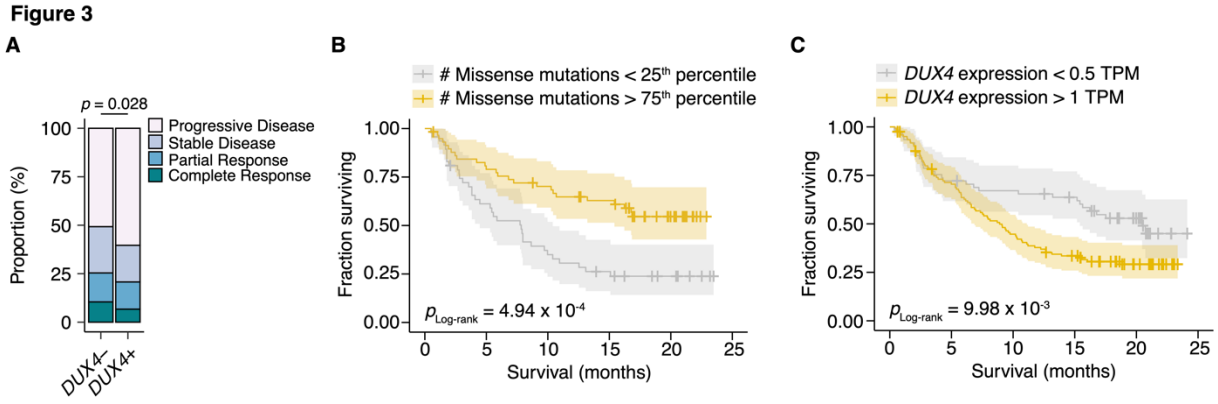


Figure 3. DUX4-positivity is associated with decreased response to immune checkpoint inhibition.

(**A**) The proportion of clinical response classifications (RECIST) in *DUX4*-positive (*DUX4*+, > 1 TPM) or -negative (*DUX4*-, ≤ 1 TPM) advanced urothelial carcinoma patients. RECIST categories were assigned in the original study (Mariathasan et al., 2018). The *p*-value was estimated via a multinomial proportion test.

(**B**) Kaplan-Meier (KM) estimates of overall survival for the patients in (A) stratified by tumor mutational burden (TMB, number of missense mutations). The estimated survival functions (solid lines) and 95% confidence intervals (transparent ribbons) for the patients in the top and bottom TMB quartiles are plotted. Censored events (crosses) The *p*-value was estimated via a log-rank test.

(**C**) As in (B), but patients are stratified by *DUX4* expression. To control for possible confounding by TMB, the quartile of patients with the lowest TMB was excluded.

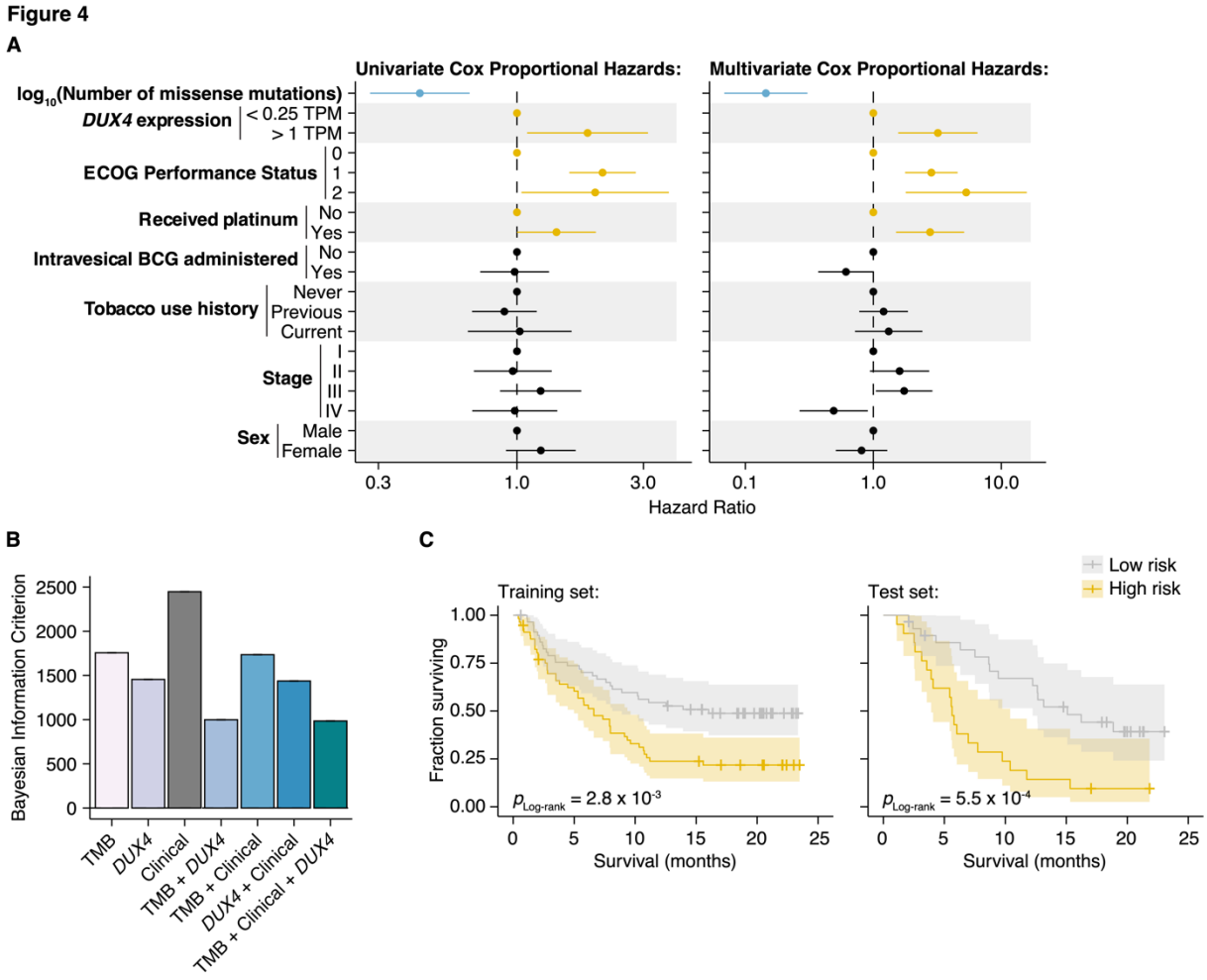


Figure 4. DUX4 expression status in patient tumors is associated with a higher risk of death
(**A**) Hazard ratios (HR) and 95% confidence intervals for the variables included in univariate (left) or multivariate (right) Cox Proportional Hazards (PH) regression. For categorical variables, the reference groups are indicated by points at HR = 1. Statistically significant predictors that are associated with increased (orange) or decreased (blue) risk in both the univariate and multivariate contexts are highlighted. ECOG (Eastern Cooperative Oncology Group); BCG (Bacillus Calmette-Guerin).
(**B**) Bayesian information criterion (BIC) measurements for goodness of fit for the full (TMB, Clinical, DUX4 expression) vs. reduced Cox PH models, where lower values indicate better fit. The bootstrapped BIC mean is illustrated. Clinical (ECOG Performance Status and Platinum treatment history).
(**C**) Kaplan-Meier (KM) estimates of overall survival and 95% confidence interval (transparent ribbon) for low-risk (solid gray line) and high-risk (solid orange line) patients in the training (left) and test (right) sets. Risk group assignments were based on risk scores estimated by the full Cox PH model. *p*-values were estimated via a log-rank test.

Figure 5

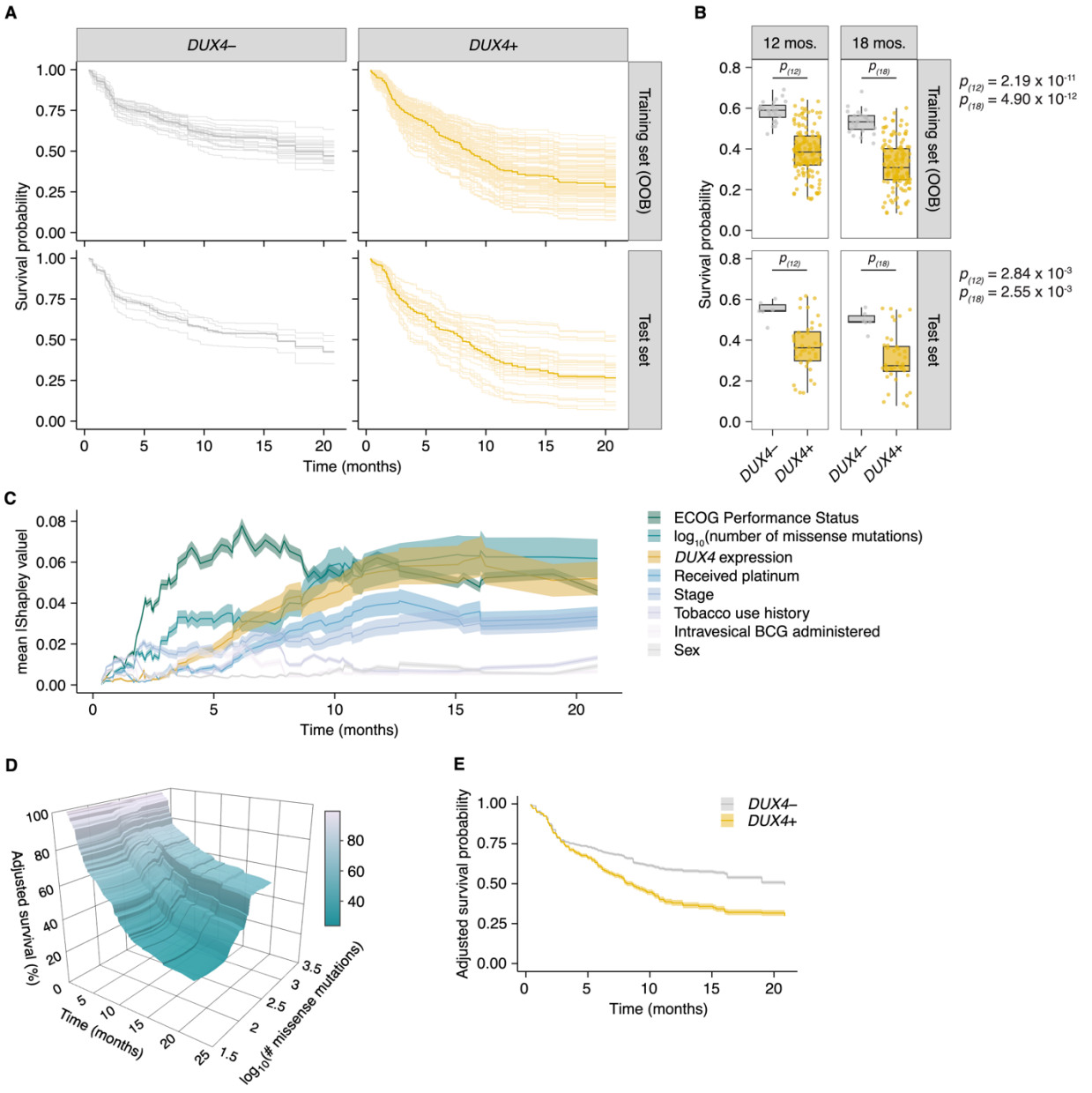


Figure 5. Overall survival during immune checkpoint inhibition therapy is decreased in patients with DUX4-expressing tumors

(**A**) Random Survival Forest (RSF) predicted overall survival for patients with either *DUX4*-positive or -negative tumors in the training and test sets. Out-of-bag (OOB) survival predictions are shown for the patients in the training set. Survival predictions for individual patients (thin lines) and the median survival function across the cohort (thick line) are represented. *DUX4*- (< 0.25 TPM); *DUX4*+ (> 1 TPM).

(**B**) Training (OOB) and test set survival probability predictions for patients with *DUX4*+/- tumors at 12 and 18 months. The *p*-values were estimated using a two-sided Mann-Whitney *U* test.

(**C**) Feature importance for variables used in the RSF model. The average absolute estimated Shapley values (solid lines) are shown, associated with predicting survival probability at particular times. The 95% confidence interval of the mean (transparent ribbon) is plotted.
(**D**) Surface plot showing adjusted (marginal) survival probability, measured via partial dependence, as a function of tumor mutational burden (TMB, number of missense mutations) and time. Each point on the surface corresponds to the mean survival prediction (at the respective timepoint) after TMB is fixed to the respective value for all patients.
(**E**) Partial plot showing adjusted survival probability as a function of *DUX4* expression status. The median survival probability (solid lines) and the 95% confidence interval (transparent ribbon) after *DUX4* expression status is fixed to the indicated value for all patients are plotted.

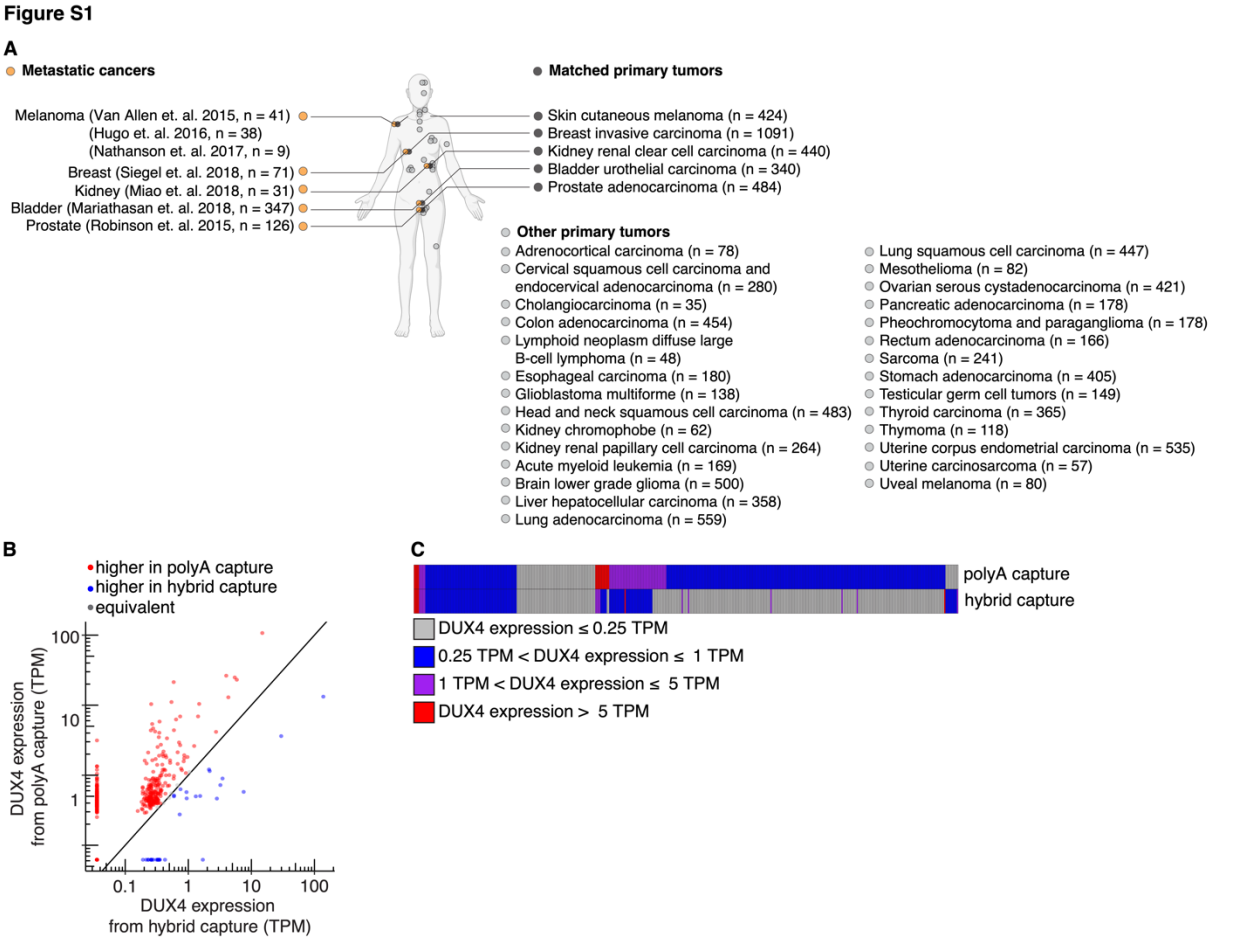


Figure S1. The DUX4 transcript is likely polyadenylated.
(**A**) As in (Figure 1A), but the primary cancer cohorts without matched normal sample analyzed in our study are shown.
(**B**) A comparison of *DUX4* expression values (TPM, transcripts per million) measured from sequencing libraries prepared via poly-A capture or hybrid capture.
(**C**) As in (B), but a heatmap where patient samples (columns) were stratified according to the indicated categories of *DUX4* expression.

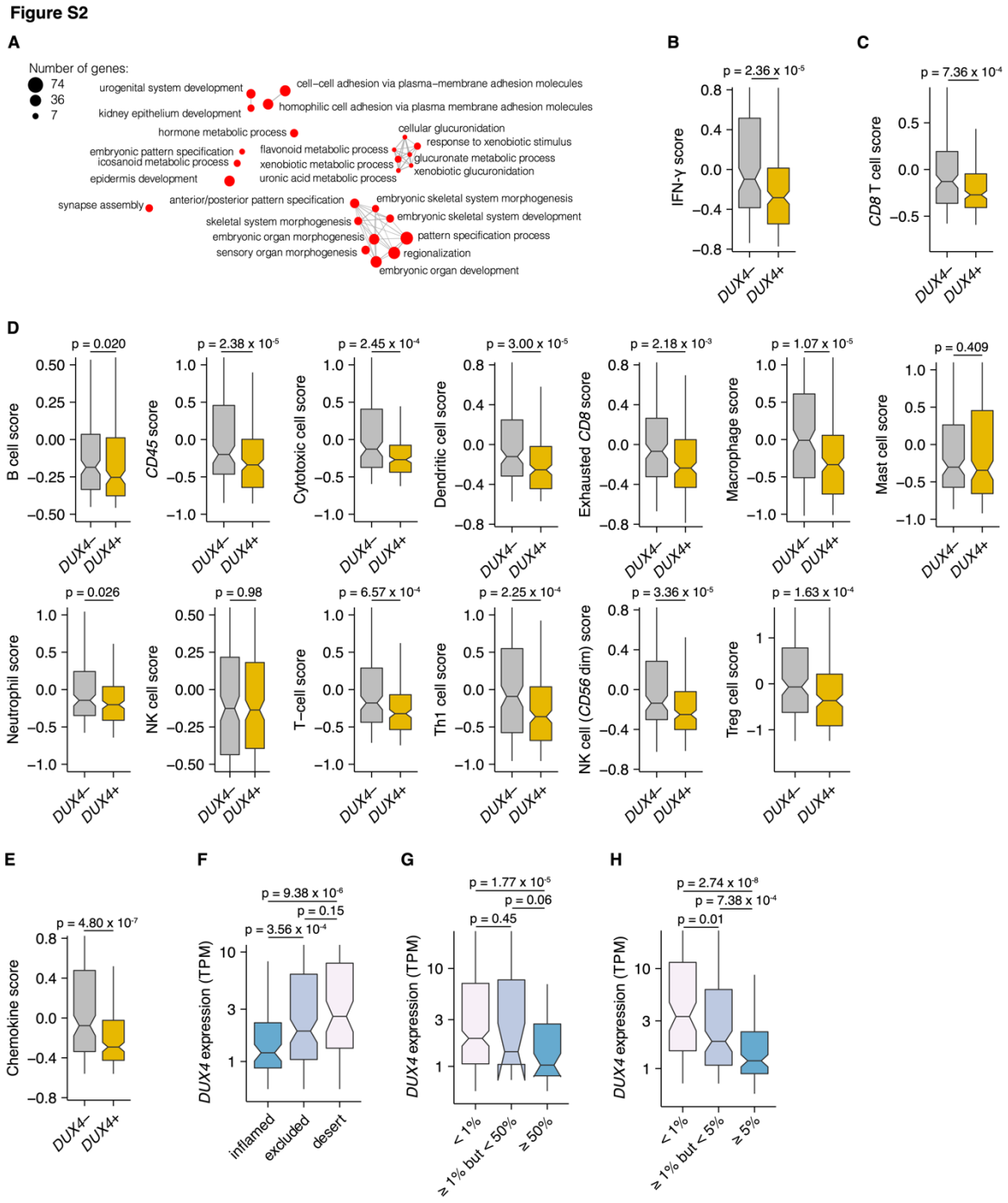


Figure S2. DUX4-positivity is correlated with an embryonic gene expression signature, downregulation of interferon-gamma signaling, and exclusion of diverse immune cell types
(**A**) As in (Figure 2A), but the Gene Ontology (GO) enrichment network analysis corresponding to *DUX4*-upregulated genes, compared against the set of coding genes, is shown.
(**B**) Interferon-gamma (IFN-γ) signature score (Ayers et al., 2017). The *p*-value was estimated via a Mann-Whitney *U* test.

(**C**) *CD8* T cell score from (Danaher et al., 2017). The *p*-value was estimated via a Mann-Whitney *U* test.
(**D**)As in (**C**), showing the other immune cell signatures available in (Danaher et al., 2017).
(**E**) Chemokine signature score (Coppola et al., 2011). The *p*-value was estimated via a Mann-Whitney *U* test.
(**F**) *DUX4* expression (TPM) in inflamed, immune excluded, and immune desert tumors. The phenotypes are based on *CD8*+ T cell abundance and degree of tumor infiltration determined by anti-*CD8* staining of tumor FFPE sections in the original study (Mariathasan et al., 2018). The *p*-values were estimated via the Mann-Whitney *U* test.
(**G**) *DUX4* expression (TPM) in advanced urothelial carcinoma tumors. The percentage of tumor cells with positive *PD-L1* staining are indicated on the x-axis. The *p*-values were estimated via the Mann-Whitney *U* test.
(**H**) As in (**G**), but showing the percentage of tumor-infiltrating immune cells (lymphocytes, macrophages, and dendritic cells) with positive *PD-L1* staining on the x-axis.

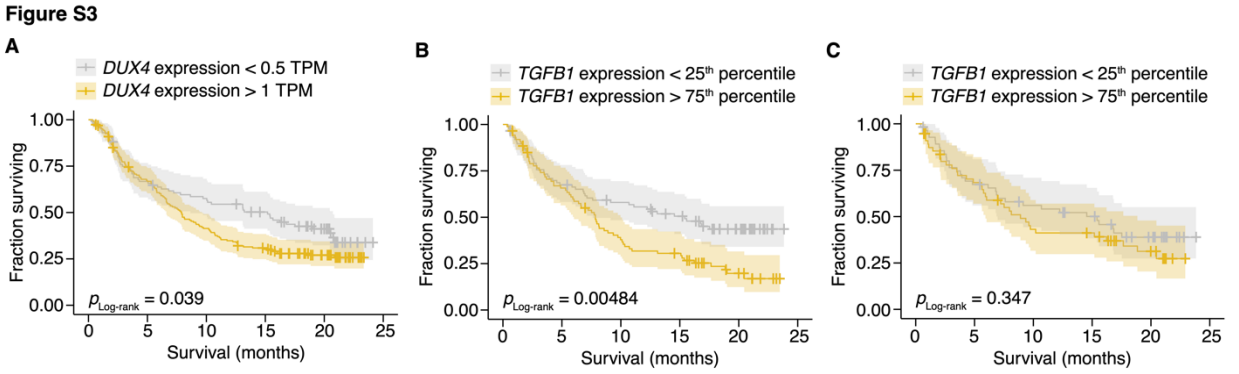


Figure S3. DUX4 expression status, but not TGFB1 expression, stratifies patients according to survival

(**A**) Kaplan-Meier (KM) estimates of overall survival (solid lines) and the corresponding 95% confidence intervals (transparent ribbons) for ICI-treated advanced urothelial carcinoma patients stratified by *DUX4* expression status. The *p*-value was estimated via a log-rank test. Censored events (crosses).

(**B**) As in (**B**), but patients falling in the top (orange) and bottom (gray) quartiles based on *TGFB1* expression are plotted.

(**C**) As in (**C**), but the quartile of patients with the lowest tumor mutational burden (TMB) was excluded to adjust for possible confounding by TMB.

Figure S4

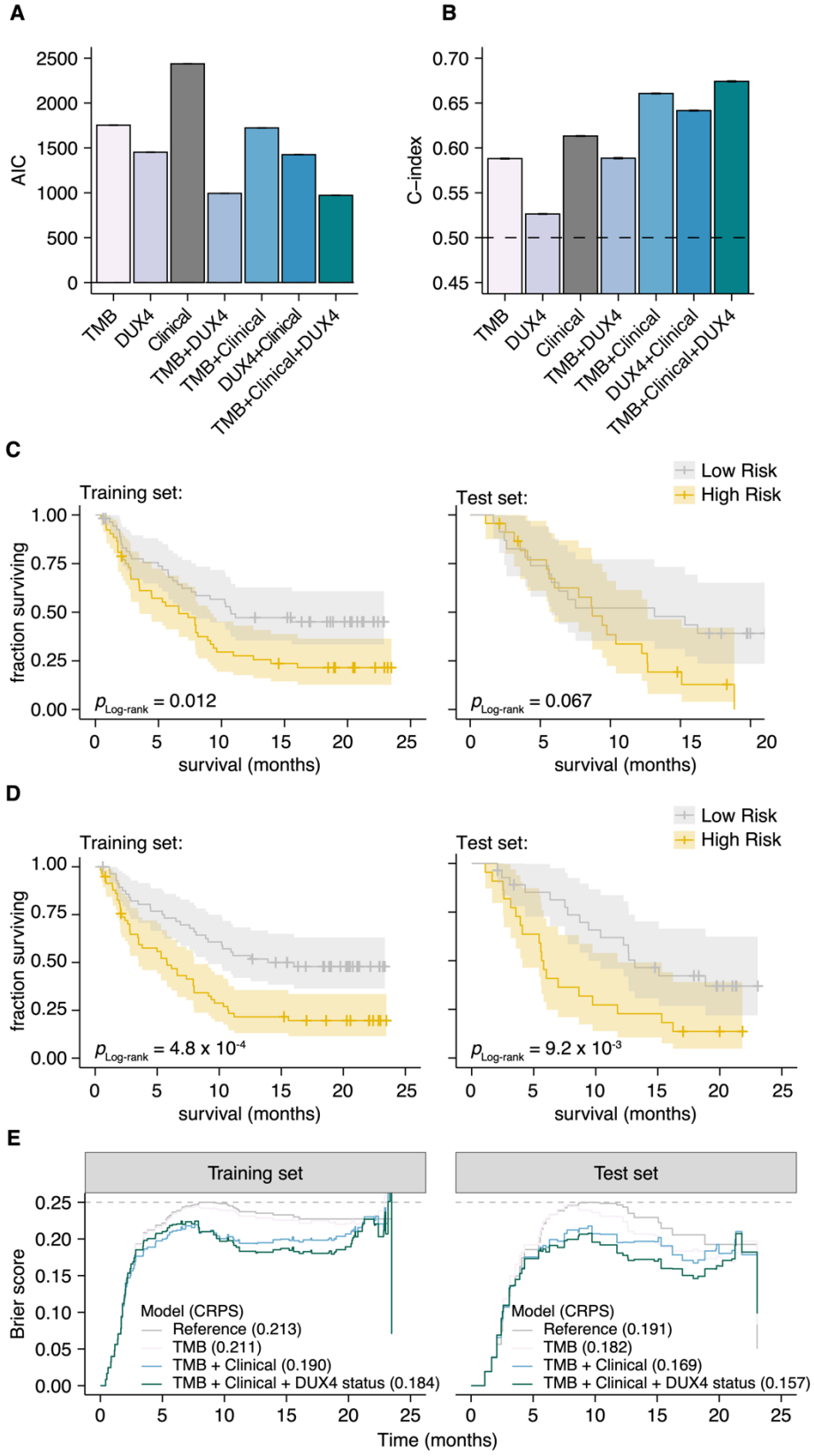

A
AIC
2500
2000
1500
1000
500
0
TMB
DUX4
Clinical
TMB+DUX4
TMB+Clinical
DUX4+Clinical
TMB+Clinical+DUX4
B
C–index
0.70
0.65
0.60
0.55
0.50
0.45
C
Training set:
Test set:
Low Risk
High Risk
fraction surviving
$p_{Log\text{-}rank}$ = 0.012
$p_{Log\text{-}rank}$ = 0.067
survival (months)
D
$p_{Log\text{-}rank}$ = 4.8 x 10^{-4}
$p_{Log\text{-}rank}$ = 9.2 x 10^{-3}
E
Training set
Test set
Brier score
Model (CRPS)
Reference (0.213)
TMB (0.211)
TMB + Clinical (0.190)
TMB + Clinical + DUX4 status (0.184)
Reference (0.191)
TMB (0.182)
TMB + Clinical (0.169)
TMB + Clinical + DUX4 status (0.157)
Time (months)

Figure S4. Cox Proportional Hazards regression models containing DUX4 expression status as a predictor have a better fit to the data

(**A**) Akaike information criterion (AIC) measurements for goodness of fit for the full (TMB, Clinical, DUX4 expression) vs. reduced Cox PH models, where lower values indicate better fit. The bootstrapped AIC mean is illustrated. Clinical (ECOG Performance Status and Platinum treatment history).

(**B**) Harrell's concordance indices (C-index) for the full (TMB, Clinical, DUX4 expression) vs. reduced Cox PH models, where high values indicate better model performance. The bootstrapped C-index mean is illustrated.

(**C**) Kaplan-Meier (KM) estimates of overall survival and 95% confidence interval (transparent ribbon) for low-risk (solid gray line) and high-risk (solid orange line) patients in the training (left) and test (right) sets. Risk group assignments were based on risk scores estimated by the Cox PH model with only TMB as a predictor. *p*-values were estimated via a log-rank test.

(**D**) As in (C), but the risk group assignments were based on risk scores estimated by the Cox PH model with TMB, ECOG Performance Status, and Platinum treatment history as predictors.

(**E)** Time-dependent Brier scores for the full and reduced Cox PH models applied on the training (left) and test (right) sets. The Continuous Ranked Probability Scores (CRPS), defined as the integrated Brier score divided by time, are shown in parentheses. Reference refers to the Kaplan-Meier prediction model. A Brier score = 0.25 indicates random guessing (gray dashed line).

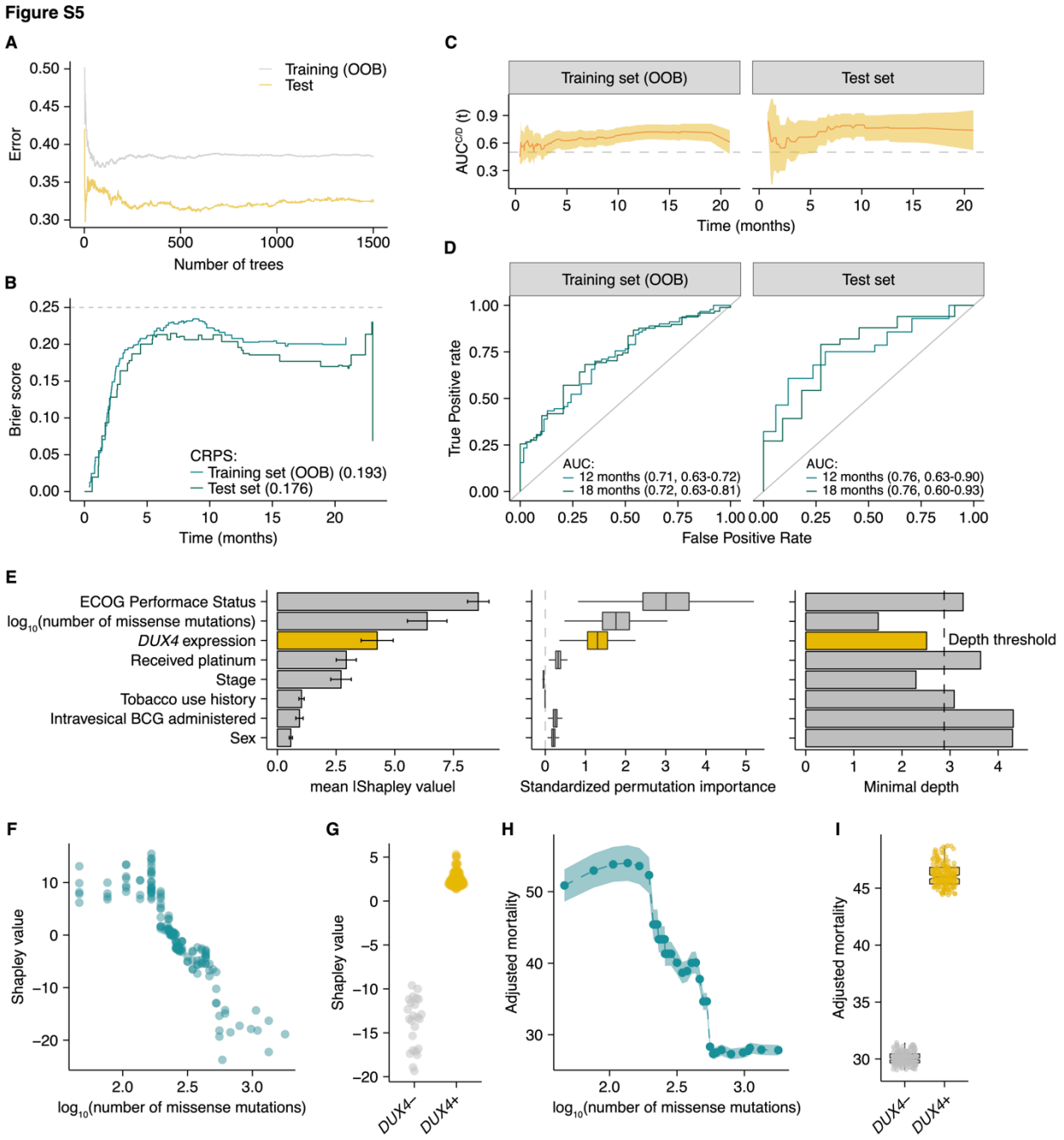


Figure S5. A Random Survival Forest model quantifies the effect of DUX4 status on overall survival probability in the context of immune checkpoint inhibition

(**A**) Error (1 – Harrell's concordance index) as a function of the number of trees in the Random Survival Forest (RSF) model. The training out-of-bag error (OOB error, solid gray line) and the test error (solid orange line) are shown. 1500 trees were used in the final model.

(**B**) Time-dependent Brier scores for the RSF model estimated from the training (solid turquoise line) or test (solid teal line) sets. OOB survival predictions were used to calculate the Brier score for the training set. The Continuous Ranked Probability Scores (CRPS), defined as the integrated

Brier score divided by time, for both sets are shown. A Brier score = 0.25 indicates random guessing (gray dashed line).
(**C**) Time-dependent ROC analyses. The $AUC^{C/D}$ (solid orange line) and the 95% confidence interval (transparent orange ribbon) are shown over the observation time for the training and test sets. The OOB mortality predictions were used to calculate the training set $AUC^{C/D}$.
(**D**) The Receiver Operating Characteristic (ROC) curves for the RSF model at 12 (solid turquoise line) and 18 (solid teal line) months. The Cumulative/Dynamic Area Under the ROC Curve ($AUC^{C/D}$) and 95% confidence interval are specified. The OOB mortality predictions were used to calculate the training set $AUC^{C/D}$.
(**E**) RSF feature importance. The mean absolute Shapley values and the 95% confidence intervals are shown (left). The standardized permutation importance and confidence regions estimated via delete-*d* jackknife subsampling are plotted (middle); noise variables are indicated by permutation importance measures ≤ 0. The minimal depth measures are shown (right); variables with values exceeding the depth threshold (gray dashed line) are designated as noise variables.
(**F**) Shapley dependence plot illustrating the relationship between tumor mutational burden (TMB, number of missense mutations) and mortality. Each point corresponds to a single patient.
(**G**) As in (**F**), but showing *DUX4* expression status.
(**H**) Partial plot illustrating the marginal effect of TMB on mortality. Each point corresponds to the average RSF mortality predictions when TMB is fixed to the indicated value for all patients. The transparent ribbon corresponds to the 95% confidence interval.
(**I**) Partial plot illustrating the marginal effect of *DUX4* expression status. The points correspond to the RSF prediction for mortality for each patient when *DUX4* expression status is fixed to the indicated value for the entire cohort.

5.6 TABLES

Table 1. Cox Proportional Hazards Regression for Overall Survival

	Univariate				Multivariate[a,b]			
Characteristic	HR	95% CI	p-value	q-value[c]	HR	95% CI	p-value	q-value[c]
Tumor mutational burden[d]	0.43	0.28 - 0.66	1.10×10^{-4}	6.62×10^{-4}	0.14	0.07 - 0.30	2.88×10^{-7}	3.46×10^{-6}
DUX4 expression:								
< 0.25 TPM (reference)	—	—	—	—	—	—	—	—
> 1 TPM	1.85	1.10 - 3.11	0.021	0.084	3.19	1.58 - 6.46	1.24×10^{-3}	3.71×10^{-3}
ECOG Performance Status[e]:								
0 (reference)	—	—	—	—	—	—	—	—
1	2.10	1.58 - 2.79	2.95×10^{-7}	3.54×10^{-6}	2.84	1.79 - 4.52	1.00×10^{-5}	6.02×10^{-5}
2	1.97	1.04 - 3.73	0.036	0.11	5.32	1.81 - 15.66	2.41×10^{-3}	5.79×10^{-3}
Received platinum:								
No (reference)	—	—	—	—	—	—	—	—
Yes	1.41	1.01 - 1.98	0.047	0.11	2.78	1.52 - 5.08	9.19×10^{-4}	3.67×10^{-3}
Intravesical BCG administered:								
No (reference)	—	—	—	—	—	—	—	—
Yes	0.98	0.73 - 1.32	0.89	0.92	0.61	0.37 - 1.01	0.054	0.080
Tobacco use history:								
Never (reference)	—	—	—	—	—	—	—	—
Previous	0.896	0.68 - 1.19	0.45	0.67	1.20	0.78 - 1.86	0.41	0.41
Current	1.02	0.65 - 1.61	0.92	0.92	1.32	0.72 - 2.42	0.37	0.41
Stage:								
I	—	—	—	—	—	—	—	—
II	0.96	0.69 - 1.35	0.83	0.92	1.60	0.62 - 0.94	0.082	0.11
III	1.23	0.86 - 1.75	0.25	0.44	1.74	1.05 - 2.90	0.033	0.056
IV	0.98	0.68 - 1.42	0.92	0.92	0.49	0.27 - 0.91	0.022	0.046
Sex:								
Male (reference)	—	—	—	—	—	—	—	—
Female	1.23	0.91 - 1.66	0.18	0.36	0.81	0.51 - 1.29	0.37	0.41

[a]Log-rank test: p-value = 1.40×10^{-8}
[b]Akaike Information Criterion = 970.78; Bayesian Information Criterion = 1003.08; Harrell's Concordance Index = 0.70
[c]Benjamini-Hochberg FDR correction
[d]$\log_{10}$(number of missense mutations)
[e]Eastern Oncology Cooperative Group Performance Status

Table 1. Cox Proportional Hazards Regression for Overall Survival

Table 2. Likelihood ratio test

	Reduced models	
	TMB only	TMB + clinical[a]
TMB + Clinical[a] + *DUX4* expression	$p = 1.08 \times 10^{-6}$	$p = 1.95 \times 10^{-3}$

[a]Eastern Cooperative Oncology Group Peformance Status + Platinum treatment history

Table 2. Likelihood ratio test

Table S1. Cox Proportional Hazards Regression for Overall Survival (TGFB1 expression included)

	Univariate				Multivariate[a,b]			
Characteristic	HR	95% CI	p-value	q-value[c]	HR	95% CI	p-value	q-value[c]
Tumor mutational burden[d]	0.43	0.28 - 0.66	1.10×10^{-4}	8.27×10^{-4}	0.14	0.06 - 0.30	8.41×10^{-7}	1.26×10^{-5}
DUX4 expression:								
< 0.25 TPM (reference)	—	—	—	—	—	—	—	—
> 1 TPM	1.85	1.10 - 3.11	0.021	0.079	3.12	1.52 - 6.38	1.87×10^{-3}	7.01×10^{-3}
ECOG Performace Statuts[e]:								
0 (reference)	—	—	—	—	—	—	—	—
1	2.10	1.58 - 2.79	2.95×10^{-7}	4.43×10^{-6}	2.86	1.79 - 4.57	1.18×10^{-5}	8.84×10^{-5}
2	1.97	1.04 - 3.73	0.036	0.091	5.31	1.79 - 15.7	2.58×10^{-3}	7.73×10^{-3}
Received platinum:								
No (reference)	—	—	—	—	—	—	—	—
Yes	1.41	1.01 - 1.98	0.047	0.10	2.69	1.47 - 4.93	1.32×10^{-3}	6.61×10^{-3}
Intravesical BCG administered:								
No (reference)	—	—	—	—	—	—	—	—
Yes	0.98	0.73 - 1.32	0.89	0.92	0.64	0.38 - 1.06	0.08	0.15
Tobacco use history:								
Never (reference)	—	—	—	—	—	—	—	—
Previous	0.90	0.68 - 1.19	0.45	0.61	1.21	0.78 - 1.87	0.39	0.45
Current	1.02	0.65 - 1.61	0.92	0.92	1.43	0.76 - 2.67	0.27	0.36
Stage:								
I	—	—	—	—	—	—	—	—
II	0.96	0.69 - 1.35	0.83	0.92	1.61	0.91 - 2.86	0.10	0.17
III	1.23	0.86 - 1.75	0.25	0.38	1.62	0.96 - 2.74	0.07	0.15
IV	0.98	0.68 - 1.42	0.92	0.92	0.46	0.24 - 0.88	0.020	0.050
Sex:								
Male (reference)	—	—	—	—	—	—	—	—
Female	1.23	0.91 - 1.66	0.18	0.31	0.76	0.47 - 1.23	0.26	0.36
TGFB1 expression:								
≤ 25th percentile (Q1, reference)	—	—	—	—	—	—	—	—
> 25th percentile and ≤ 50th percentile (Q2)	1.30	0.88 - 1.91	0.19	0.31	0.75	0.40 - 1.41	0.38	0.45
> 50th percentile and ≤ 75th percentile (Q3)	1.52	1.04 - 2.23	0.031	0.091	1.04	0.55 - 1.95	0.90	0.91
> 75th percentile (Q4)	1.68	1.16 - 2.45	6.46×10^{-3}	0.032	0.96	0.50 - 1.87	0.91	0.91

[a]Log-rank test: p-value = 1.27×10^{-7}
[b]Akaike Information Criterion = 975.30; Bayesian Information Criterion = 1015.67; Harrell's Concordance Index = 0.71
[c]Benjamini-Hochberg FDR correction
[d]$\log_{10}$(number of missense mutations)
[e]Eastern Oncology Cooperative Group Performance Status

Table S1. Cox Proportional Hazards Regression for Overall Survival (TGFB1 expression included)

5.7 MATERIALS AND METHODS

5.7.1 *Method Details*

5.7.1.1 Genome annotations, gene expression and Gene Ontology (GO) enrichment analyses

A genome annotation was created, a combination of the UCSC knownGene (Meyer et al., 2013), Ensembl 71 (Flicek et al., 2013), and MISO v2.0 (Katz et al., 2010) versions for the hg19/GRCh37 assembly. Further, this annotation was expanded by generating all possible combinations of annotated 5' and 3' splice sites within each gene. RNA-seq reads were mapped to the transcriptome using RSEM v1.2.4 (B. Li & Dewey, 2011) calling Bowtie v1.0.0 (Langmead et al., 2009), with the option "-v 2." TopHat v.2.0.8b (Trapnell et al., 2009) was used to map the unaligned reads to the genome and to the database of splice junctions obtained from the annotation merging described previously. Gene expression estimates (TPM, transcripts per million) obtained were normalized using the trimmed mean of M values (TMM) method (Robinson & Oshlack, 2010). In the differential gene expression analyses for the DUX4-positive vs. -negative comparison, gene expression values per sample group were compared using a two-sided Mann-Whitney *U* test. Differentially expressed genes illustrated in Figure 2B were identified as those with an absolute $\log_2$(fold-change) $\geq$ $\log_2$(1.25) and a *p*-value < 0.05. GO enrichment analyses, using the clusterProfiler package (Wu et al., 2021; Yu et al., 2012), were performed on *DUX4*-upregulated or -downregulated genes [absolute $\log_2$(fold-change) $\geq$ $\log_2$(1.5) and a *p*-value < 0.05] compared against the set of coding genes. Significant GO terms were defined as "Biological Process" terms with a Benjamini-Hochberg FDR-adjusted *p*-value < 0.05. The top 25 significant GO terms were illustrated (Fig. 2A and Fig. S2A). To investigate *DUX4* RNA-seq coverage patterns, a fasta file containing the *DUX4* cDNA sequence was assembled, indexed using samtools (H. Li et al., 2009),

and used as a reference for read pseudoalignment by kallisto v.0.46.1 (Bray et al., 2016). The following kallisto parameters were used: kmer size of 31, estimated fragment length of 200, and estimated fragment length standard deviation of 80. Usage of the single-end option ("--single") and bias correction ("--bias") were also specified. *DUX4* read coverage was visualized using the Integrative Genomics Viewer (IGV, Thorvaldsdóttir et al., 2013).

5.7.1.2 Gene signature analyses

For a given gene set, z-score normalization of the expression values per gene was performed across the patient cohort. The signature score was defined as the mean of the normalized values across the genes of the set.

5.7.1.3 Survival analyses, goodness of fit measures, and risk modeling

Kaplan-Meier (KM) estimation, *p*-value estimates from the log-rank test, and Cox Proportional Hazards (PH) regression in the univariate and multivariate contexts were performed using the survival package (T. Therneau, 2022; T. M. Therneau & Grambsch, 2000). Goodness of fit evaluations of the Cox PH models were done by measuring the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC metrics balance model complexity with maximized likelihood, penalizing feature number increases without a concomitant improvement in performance. The likelihood ratio test was also used to compare goodness of fit of full (all variables) vs. reduced (subset of variables) Cox PH models. Specifically, the null hypothesis that the simple model provides as good as a fit as the more complex model was evaluated. The AIC, BIC, and likelihood ratio test *p*-values were computed using R's stats package (R Core Team, 2022). For the Cox PH risk modeling, the patients were randomly assigned into training (70%) and test (30%) datasets. The createDataPartition() function from the caret package (Kuhn, 2022) was used to preserve the *DUX4* status class distribution after splitting. Full and

reduced Cox PH models were created using the training data and the risk scores for each respective model were calculated using caret's predict.coxph() function. For a given patient, the calculated risk score is equal to the hazard ratio relative to a "reference patient" (an individual whose covariate values are set to the respective means, from the training set). Specifically, the risk score is the quotient of the patient's and the reference's exponentiated linear predictors (the sum of the covariates in the model, weighted by the model's regression coefficients). A "reference risk score" for each model was defined as the median risk score in the training data. Patients were assigned into low- or high-risk groups if their risk scores were lower or higher than the reference, respectively. The trained models were used to calculate risk scores and assign risk labels (based on the training set risk score reference) in the test set. The survival difference between low- and high-risk patients were empirically assessed via KM estimation and the log-rank test. Visualizations were created using the ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2022), and survminer (Kassambara et al., 2021) packages.

5.7.1.4 Random Survival Forest, feature importance, and partial dependence

We implemented a Random Survival Forest (RSF) model, an ensemble of multiple base learners (survival trees), using the randomForestSRC package (Ishwaran et al., 2008). The RSF algorithm is an extension of the Random Forest Algorithm (Breiman, 2001) for usage with right-censored data. Here, B bootstrap datasets are created from the original data, used to grow B concomitant survival trees (usually constrained by membership size in the terminal nodes) constructed using a randomly selected subset of the variables. Terminal node statistics are obtained for each tree: the survival function (via the Kaplan-Meier estimator), the cumulative hazard function (CHF, via the Nelson-Aalen estimator), and mortality (expected number of deaths; sum of the CHF over time).

The RSF prediction is the average across the forest. Of note, each bootstrap dataset excludes 36.8% of the original data on average, the out-of-bag (OOB) samples. Thus, predictions for a particular sample can be made using the subset of the trees for which it was excluded from training (OOB predictions). Similarly, the associated OOB error for the RSF model can calculated, representing an unbiased estimate of the test error. We randomly assigned patients into training (70%) and test (30%) datasets. Since the *DUX4*-positive status was a minority class, we utilized the createDataPartition() function from the caret package (Kuhn, 2022) to preserve the class distribution within the splits. To determine optimal hyperparameters, we evaluated 5,616 RSF models representing different combinations of ntree (number of trees), nodesize (minimum terminal node size), mtry (number of randomly selected splitting variables), na.action (handling of missing data), splitrule (splitting rule), and samptype (type of bootstrap). We selected the model with hyperparameters which minimized both the OOB training and the test errors (defined as 1 – concordance index), namely: ntree = 1500, nodesize = 15, mtry = 3, na.action = "na.impute", splitrule = "bs.gradient", and samptype = "swr." We specified the use of an nsplit (number of random splits) value of 0 to indicate evaluation of all possible split points and usage of the optimum. For test set predictions, patients with missing data were omitted (na.action = "na.omit").

Feature importance in the final RSF model was evaluated using 3 metrics. First, permutation importance was measured using randomForestSRC's subsample() function. RSF permutation importance utilizes OOB values: a variable's OOB data is permuted and the change in the new vs. original OOB prediction error is quantified. The RSF permutation importance values were standardized by dividing by the variance and multiplying by 100, and the variance and confidence regions were obtained via the delete-*d* jackknife estimator. Second, the tree-based feature importance metric minimal depth was calculated using randomForestSRC's var.select()

function. The minimal depth threshold (mean minimal depth) is the tree-averaged threshold (conservative = "medium"). Last, Shapley values were estimated using the fastshap package (Greenwell, 2021), using 1000 Monte Carlo repetitions. For each prediction, the sum of the estimated Shapley values was corrected (adjust = TRUE) to satisfy the efficiency (or local accuracy) property: for an individual *i*, the sum of *i*'s feature contributions equal the difference between the prediction for *i* and the average prediction across the entire cohort. For the overall measure of importance, the Shapley values were estimated from the mortality predictions from the RSF model (Fig S5E). Mortality is defined as the number of expected deaths. That is, if all patients in the cohort shared the same covariate values as patient *i* who has mortality m_i, then an average of *m* deaths is expected (Ishwaran et al., 2008). For the time-dependent implementation, we estimated Shapley values associated with the per timepoint RSF survival probability predictions along the observation window (Fig 5C).

The relationships of *DUX4* expression and TMB to mortality or survival probability (marginal contributions) were assessed via Shapley dependence plots and partial dependence plots. Partial dependence values were obtained using randomForestSRC's partial() function and OOB predictions for mortality and survival probability were used as input. Visualizations were created in the R programming environment using the dplyr (Wickham et al., 2022), ggplot2 (Wickham, 2016), pammtools (Bender & Scheipl, 2018), and plotly (Sievert, 2020) packages.

5.7.1.5 Measuring survival model predictive accuracy

The time-dependent Receiver Operating Characteristic (ROC) curve analyses were done to evaluate the RSF model's accuracy in differentiating patients who die before a particular time *t*, from those who survive past *t* (Heagerty & Zheng, 2005). Specifically, for each timepoint, the cumulative/dynamic Area Under the ROC curve ($AUC^{C/D}$) was computed by computing the

sensitivity (true positive rate) and specificity (1 – false positive rate) associated with using RSF-predicted mortality as the prognostic marker. The time-dependent $AUC^{C/D}$ and 95% confidence interval per time point were estimated using the timeROC package, which adds the inverse-probability-of-censoring weights (IPCW) to the sensitivity calculation to correct for selection bias due to right-censoring (Blanche et al., 2013). The out-of-bag (training) or the test mortality predictions were used as input. The time-dependent Brier score and the Continuous Ranked Probability Score (CRPS, integrated Brier score divided by time) for the Cox PH models were computed using the pec package (Mogensen et al., 2012). The time-dependent Brier score and the CRPS for the RSF model was calculated using the randomForestSRC package (Ishwaran et al., 2008). The Kaplan-Meier estimator for the censoring times was used to estimate the IPCW (cens.model = "marginal"). Harrell's concordance index for the Cox PH and RSF models were calculated using the survival (T. Therneau, 2022; T. M. Therneau & Grambsch, 2000) and randomForestSRC packages, respectively. Visualizations were created in the R programming environment using the dplyr (Wickham et al., 2022) and ggplot2 (Wickham, 2016) packages.

5.8 ACKNOWLEDGMENTS

R.K.B. was supported in part by the NIH/NCI (R01 CA251138), NIH/NHLBI (R01 HL128239 and R01 HL151651) and the Blood Cancer Discoveries Grant program through the Leukemia & Lymphoma Society, Mark Foundation for Cancer Research, and Paul G. Allen Frontiers Group (8023-20). R.K.B is a Scholar of The Leukemia & Lymphoma Society (1344-18) and holds the McIlwain Family Endowed Chair in Data Science. The results published here are based in part upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

5.9 Author Contributions

J.M.B.P. and R.K.B. designed the study, analyzed the data, and wrote the paper.

5.10 Competing Interests

The authors declare that they have no competing interests.

5.11 References

Alspach, E., Lussier, D. M., & Schreiber, R. D. (2019). Interferon γ and Its Important Roles in Promoting and Inhibiting Spontaneous and Therapeutic Cancer Immunity. *Cold Spring Harbor Perspectives in Biology*, *11*(3). https://doi.org/10.1101/cshperspect.a028480

Amaria, R. N., Postow, M., Burton, E. M., Tezlaff, M. T., Ross, M. I., Torres-Cabala, C., Glitza, I. C., Duan, F., Milton, D. R., Busam, K., Simpson, L., McQuade, J. L., Wong, M. K., Gershenwald, J. E., Lee, J. E., Goepfert, R. P., Keung, E. Z., Fisher, S. B., Betof-Warner, A., … Tawbi, H. A. (2022). Neoadjuvant relatlimab and nivolumab in resectable melanoma. *Nature*, *611*(7934), 155–160. https://doi.org/10.1038/s41586-022-05368-8

Ascierto, P. A., Ferrucci, P. F., Fisher, R., del Vecchio, M., Atkinson, V., Schmidt, H., Schachter, J., Queirolo, P., Long, G. v, di Giacomo, A. M., Svane, I. M., Lotem, M., Bar-Sela, G., Couture, F., Mookerjee, B., Ghori, R., Ibrahim, N., Moreno, B. H., & Ribas, A. (2019). Dabrafenib, trametinib and pembrolizumab or placebo in BRAF-mutant melanoma. *Nature Medicine*, *25*(6), 941–946. https://doi.org/10.1038/s41591-019-0448-9

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., Albright, A., Cheng, J. D., Kang, S. P., Shankaran, V., Piha-Paul, S. A., Yearley, J., Seiwert, T. Y., Ribas, A., & McClanahan, T. K. (2017). IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *The Journal of Clinical Investigation*, *127*(8), 2930–2940. https://doi.org/10.1172/JCI91190

Balar, A. v, Galsky, M. D., Rosenberg, J. E., Powles, T., Petrylak, D. P., Bellmunt, J., Loriot, Y., Necchi, A., Hoffman-Censits, J., Perez-Gracia, J. L., Dawson, N. A., van der Heijden, M. S., Dreicer, R., Srinivas, S., Retz, M. M., Joseph, R. W., Drakaki, A., Vaishampayan, U. N., Sridhar, S. S., … IMvigor210 Study Group. (2017). Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet (London, England)*, *389*(10064), 67–76. https://doi.org/10.1016/S0140-6736(16)32455-2

Bellmunt, J., de Wit, R., Vaughn, D. J., Fradet, Y., Lee, J.-L., Fong, L., Vogelzang, N. J., Climent, M. A., Petrylak, D. P., Choueiri, T. K., Necchi, A., Gerritsen, W., Gurney, H., Quinn, D. I., Culine, S., Sternberg, C. N., Mai, Y., Poehlein, C. H., Perini, R. F., … KEYNOTE-045 Investigators. (2017). Pembrolizumab as Second-Line Therapy for Advanced Urothelial

Carcinoma. *The New England Journal of Medicine*, *376*(11), 1015–1026. https://doi.org/10.1056/NEJMoa1613683

Bender, A., & Scheipl, F. (2018). *pammtools: Piece-wise exponential Additive Mixed Modeling tools*.

Blanche, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, *32*(30), 5381–5397. https://doi.org/10.1002/sim.5958

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bylander, T. (2002). Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates. *Machine Learning*, *48*(1/3), 287–297. https://doi.org/10.1023/A:1013964023376

Chalker, C., Voutsinas, J. M., Wu, Q. V., Santana-Davila, R., Hwang, V., Baik, C. S., Lee, S., Barber, B., Futran, N. D., Houlton, J. J., Laramore, G. E., Liao, J. J., Parvathaneni, U., Martins, R. G., Eaton, K. D., & Rodriguez, C. P. (2022). Performance status (PS) as a predictor of poor response to immune checkpoint inhibitors (ICI) in recurrent/metastatic head and neck cancer (RMHNSCC) patients. *Cancer Medicine*. https://doi.org/10.1002/cam4.4722

Chew, G.-L., Campbell, A. E., de Neef, E., Sutliff, N. A., Shadle, S. C., Tapscott, S. J., & Bradley, R. K. (2019). DUX4 Suppresses MHC Class I to Promote Cancer Immune Evasion and Resistance to Checkpoint Blockade. *Developmental Cell*, *50*(5), 658-671.e7. https://doi.org/10.1016/j.devcel.2019.06.011

Coppola, D., Nebozhyn, M., Khalil, F., Dai, H., Yeatman, T., Loboda, A., & Mulé, J. J. (2011). Unique ectopic lymph node-like structures present in human primary colorectal carcinoma are identified by immune gene array profiling. *The American Journal of Pathology*, *179*(1), 37–45. https://doi.org/10.1016/j.ajpath.2011.03.007

Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X. Q., Lu, H., Nebozhyn, M., Zhang, C., Lunceford, J. K., Joe, A., Cheng, J., Webber, A. L., Ibrahim, N., Plimack, E. R., Ott, P. A., Seiwert, T. Y., … Kaufman, D. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science (New York, N.Y.)*, *362*(6411). https://doi.org/10.1126/science.aar3593

Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., Geller, M. A., Odunsi, K., Beechem, J., & Fling, S. P. (2017). Gene expression markers of Tumor Infiltrating Leukocytes. *Journal for Immunotherapy of Cancer*, *5*, 18. https://doi.org/10.1186/s40425-017-0215-8

Das, S., & Chadwick, B. P. (2016). Influence of Repressive Histone and DNA Methylation upon D4Z4 Transcription in Non-Myogenic Cells. *PloS One*, *11*(7), e0160022. https://doi.org/10.1371/journal.pone.0160022

de Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., & Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nature Genetics*, *49*(6), 941–945. https://doi.org/10.1038/ng.3858

Dietrich, S., Floegel, A., Troll, M., Kühn, T., Rathmann, W., Peters, A., Sookthai, D., von Bergen, M., Kaaks, R., Adamski, J., Prehn, C., Boeing, H., Schulze, M. B., Illig, T., Pischon, T., Knüppel, S., Wang-Sattler, R., & Drogan, D. (2016). Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. *International Journal of Epidemiology*, *45*(5), 1406–1420. https://doi.org/10.1093/ije/dyw145

Dixon, K. O., Tabaka, M., Schramm, M. A., Xiao, S., Tang, R., Dionne, D., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., & Kuchroo, V. K. (2021). TIM-3 restrains anti-tumour immunity by regulating inflammasome activation. *Nature*, *595*(7865), 101–106. https://doi.org/10.1038/s41586-021-03626-9

Doki, Y., Ajani, J. A., Kato, K., Xu, J., Wyrwicz, L., Motoyama, S., Ogata, T., Kawakami, H., Hsu, C.-H., Adenis, A., el Hajbi, F., di Bartolomeo, M., Braghiroli, M. I., Holtved, E., Ostoich, S. A., Kim, H. R., Ueno, M., Mansoor, W., Yang, W.-C., … CheckMate 648 Trial Investigators. (2022). Nivolumab Combination Therapy in Advanced Esophageal Squamous-Cell Carcinoma. *The New England Journal of Medicine*, *386*(5), 449–462. https://doi.org/10.1056/NEJMoa2111380

Ebert, P. J. R., Cheung, J., Yang, Y., McNamara, E., Hong, R., Moskalenko, M., Gould, S. E., Maecker, H., Irving, B. A., Kim, J. M., Belvin, M., & Mellman, I. (2016). MAP Kinase Inhibition Promotes T Cell and Anti-tumor Activity in Combination with PD-L1 Checkpoint Blockade. *Immunity*, *44*(3), 609–621. https://doi.org/10.1016/j.immuni.2016.01.024

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., … Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Research*, *41*(D1), 48–55. https://doi.org/10.1093/nar/gks1236

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Gajic, Z. Z., Deshpande, A., Legut, M., Imieliński, M., & Sanjana, N. E. (2022). Recurrent somatic mutations as predictors of immunotherapy response. *Nature Communications*, *13*(1), 3938. https://doi.org/10.1038/s41467-022-31055-3

Galsky, M. D., Arija, J. Á. A., Bamias, A., Davis, I. D., de Santis, M., Kikuchi, E., Garcia-Del-Muro, X., de Giorgi, U., Mencinger, M., Izumi, K., Panni, S., Gumus, M., Özgüroğlu, M., Kalebasty, A. R., Park, S. H., Alekseev, B., Schutz, F. A., Li, J.-R., Ye, D., … IMvigor130

Study Group. (2020). Atezolizumab with or without chemotherapy in metastatic urothelial cancer (IMvigor130): a multicentre, randomised, placebo-controlled phase 3 trial. *Lancet (London, England)*, *395*(10236), 1547–1557. https://doi.org/10.1016/S0140-6736(20)30230-0

Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., de Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., Powell, S. F., Cheng, S. Y.-S., Bischoff, H. G., Peled, N., Grossi, F., Jennens, R. R., Reck, M., Hui, R., Garon, E. B., Boyer, M., … KEYNOTE-189 Investigators. (2018). Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *The New England Journal of Medicine*, *378*(22), 2078–2092. https://doi.org/10.1056/NEJMoa1801005

Gao, J., Shi, L. Z., Zhao, H., Chen, J., Xiong, L., He, Q., Chen, T., Roszik, J., Bernatchez, C., Woodman, S. E., Chen, P.-L., Hwu, P., Allison, J. P., Futreal, A., Wargo, J. A., & Sharma, P. (2016). Loss of IFN-γ Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. *Cell*, *167*(2), 397-404.e9. https://doi.org/10.1016/j.cell.2016.08.069

Gide, T. N., Quek, C., Menzies, A. M., Tasker, A. T., Shang, P., Holst, J., Madore, J., Lim, S. Y., Velickovic, R., Wongchenko, M., Yan, Y., Lo, S., Carlino, M. S., Guminski, A., Saw, R. P. M., Pang, A., McGuire, H. M., Palendira, U., Thompson, J. F., … Wilmott, J. S. (2019). Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell*, *35*(2), 238-255.e6. https://doi.org/10.1016/j.ccell.2019.01.003

Glimcher, L. H., & Kara, C. J. (1992). Sequences and factors: a guide to MHC class-II transcription. *Annual Review of Immunology*, *10*, 13–49. https://doi.org/10.1146/annurev.iy.10.040192.000305

Goel, S., DeCristo, M. J., Watt, A. C., BrinJones, H., Sceneay, J., Li, B. B., Khan, N., Ubellacker, J. M., Xie, S., Metzger-Filho, O., Hoog, J., Ellis, M. J., Ma, C. X., Ramm, S., Krop, I. E., Winer, E. P., Roberts, T. M., Kim, H.-J., McAllister, S. S., & Zhao, J. J. (2017). CDK4/6 inhibition triggers anti-tumour immunity. *Nature*, *548*(7668), 471–475. https://doi.org/10.1038/nature23465

Grasso, C. S., Giannakis, M., Wells, D. K., Hamada, T., Mu, X. J., Quist, M., Nowak, J. A., Nishihara, R., Qian, Z. R., Inamura, K., Morikawa, T., Nosho, K., Abril-Rodriguez, G., Connolly, C., Escuin-Ordinas, H., Geybels, M. S., Grady, W. M., Hsu, L., Hu-Lieskovan, S., … Peters, U. (2018). Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer Discovery*, *8*(6), 730–749. https://doi.org/10.1158/2159-8290.CD-17-1327

Grasso, C. S., Tsoi, J., Onyshchenko, M., Abril-Rodriguez, G., Ross-Macdonald, P., Wind-Rotolo, M., Champhekar, A., Medina, E., Torrejon, D. Y., Shin, D. S., Tran, P., Kim, Y. J., Puig-Saus, C., Campbell, K., Vega-Crespo, A., Quist, M., Martignier, C., Luke, J. J., Wolchok, J. D., … Ribas, A. (2020). Conserved Interferon-γ Signaling Drives Clinical Response to Immune Checkpoint Blockade Therapy in Melanoma. *Cancer Cell*, *38*(4), 500-515.e3. https://doi.org/10.1016/j.ccell.2020.08.005

Greenwell, B. (2021). *fastshap: Fast Approximate Shapley Values* (R package version 0.0.7). https://CRAN.R-project.org/package=fastshap

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Hellmann, M. D., Paz-Ares, L., Bernabe Caro, R., Zurawski, B., Kim, S.-W., Carcereny Costa, E., Park, K., Alexandru, A., Lupinacci, L., de la Mora Jimenez, E., Sakai, H., Albert, I., Vergnenegre, A., Peters, S., Syrigos, K., Barlesi, F., Reck, M., Borghaei, H., Brahmer, J. R., … Ramalingam, S. S. (2019). Nivolumab plus Ipilimumab in Advanced Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, *381*(21), 2020–2031. https://doi.org/10.1056/NEJMoa1910231

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., Weaver, B. D., Pflueger, C., Emery, B. R., Wilcox, A. L., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T., & Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, *49*(6), 925–934. https://doi.org/10.1038/ng.3844

Himeda, C. L., & Jones, P. L. (2019). The Genetics and Epigenetics of Facioscapulohumeral Muscular Dystrophy. *Annual Review of Genomics and Human Genetics*, *20*, 265–291. https://doi.org/10.1146/annurev-genom-083118-014933

Hsich, E. M., Thuita, L., McNamara, D. M., Rogers, J. G., Valapour, M., Goldberg, L. R., Yancy, C. W., Blackstone, E. H., Ishwaran, H., & Transplantation of HEarts to MaxImize Survival (THEMIS) Investigators. (2019). Variables of importance in the Scientific Registry of Transplant Recipients database predictive of heart transplant waitlist mortality. *American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, *19*(7), 2067–2076. https://doi.org/10.1111/ajt.15265

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*(none). https://doi.org/10.1214/07-EJS039

Ishwaran, H., Blackstone, E. H., Apperson-Hansen, C., & Rice, T. W. (2009). A novel approach to cancer staging: application to esophageal cancer. *Biostatistics (Oxford, England)*, *10*(4), 603–620. https://doi.org/10.1093/biostatistics/kxp016

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3). https://doi.org/10.1214/08-AOAS169

Ishwaran, H., Kogalur, U. B., Chen, X., & Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *4*(1), 115–132. https://doi.org/10.1002/sam.10103

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, *105*(489), 205–217. https://doi.org/10.1198/jasa.2009.tm08622

Ishwaran, H., & Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, *38*(4), 558–582. https://doi.org/10.1002/sim.7803

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PloS One*, *13*(8), e0201904. https://doi.org/10.1371/journal.pone.0201904

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J.-R., Wang, S., Rabasha, B., Liu, D., Zhang, G., Margolais, C., Ashenberg, O., Ott, P. A., Buchbinder, E. I., Haq, R., … Regev, A. (2018). A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell*, *175*(4), 984-997.e24. https://doi.org/10.1016/j.cell.2018.09.006

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., Li, Z., Traugh, N., Bu, X., Li, B., Liu, J., Freeman, G. J., Brown, M. A., Wucherpfennig, K. W., & Liu, X. S. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature Medicine*, *24*(10), 1550–1558. https://doi.org/10.1038/s41591-018-0136-1

Johnson, A. M., Bullock, B. L., Neuwelt, A. J., Poczobutt, J. M., Kaspar, R. E., Li, H. Y., Kwak, J. W., Hopp, K., Weiser-Evans, M. C. M., Heasley, L. E., Schenk, E. L., Clambey, E. T., & Nemenoff, R. A. (2020). Cancer Cell-Intrinsic Expression of MHC Class II Regulates the Immune Microenvironment and Response to Anti-PD-1 Therapy in Lung Adenocarcinoma. *Journal of Immunology (Baltimore, Md. : 1950)*, *204*(8), 2295–2307. https://doi.org/10.4049/jimmunol.1900778

Kalbasi, A., & Ribas, A. (2020). Tumour-intrinsic resistance to immune checkpoint blockade. *Nature Reviews. Immunology*, *20*(1), 25–39. https://doi.org/10.1038/s41577-019-0218-4

Kassambara, A., Kosinski, M., & Biecek, P. (2021). *survminer: Drawing Survival Curves using "ggplot2"* (R package version 0.4.9). https://CRAN.R-project.org/package=survminer

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

Klein, O., Kee, D., Nagrial, A., Markman, B., Underhill, C., Michael, M., Jackett, L., Lum, C., Behren, A., Palmer, J., Tebbutt, N. C., Carlino, M. S., & Cebon, J. (2020). Evaluation of Combination Nivolumab and Ipilimumab Immunotherapy in Patients With Advanced Biliary Tract Cancers: Subgroup Analysis of a Phase 2 Nonrandomized Clinical Trial. *JAMA Oncology*, *6*(9), 1405–1409. https://doi.org/10.1001/jamaoncol.2020.2814

Krishnan, M., Kasinath, P., High, R., Yu, F., & Teply, B. A. (2022). Impact of Performance Status on Response and Survival Among Patients Receiving Checkpoint Inhibitors for Advanced

Solid Tumors. *JCO Oncology Practice*, *18*(1), e175–e182. https://doi.org/10.1200/OP.20.01055

Kuhn, M. (2022). *caret: Classification and Regression Training* (R package version 6.0-93). https://CRAN.R-project.org/package=caret

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Lao, C. D., Cowey, C. L., Schadendorf, D., Wagstaff, J., Dummer, R., Ferrucci, P. F., Smylie, M., Hogg, D., Hill, A., Márquez-Rodas, I., Haanen, J., Guidoboni, M., Maio, M., Schöffski, P., … Wolchok, J. D. (2019). Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *The New England Journal of Medicine*, *381*(16), 1535–1546. https://doi.org/10.1056/NEJMoa1910836

Lau, J., Cheung, J., Navarro, A., Lianoglou, S., Haley, B., Totpal, K., Sanders, L., Koeppen, H., Caplazi, P., McBride, J., Chiu, H., Hong, R., Grogan, J., Javinal, V., Yauch, R., Irving, B., Belvin, M., Mellman, I., Kim, J. M., & Schmidt, M. (2017). Tumour and host cell PD-L1 is required to mediate suppression of anti-tumour immunity in mice. *Nature Communications*, *8*(1), 14572. https://doi.org/10.1038/ncomms14572

Lee, J. H., Shklovskaya, E., Lim, S. Y., Carlino, M. S., Menzies, A. M., Stewart, A., Pedersen, B., Irvine, M., Alavi, S., Yang, J. Y. H., Strbenac, D., Saw, R. P. M., Thompson, J. F., Wilmott, J. S., Scolyer, R. A., Long, G. v, Kefford, R. F., & Rizos, H. (2020). Transcriptional downregulation of MHC class I and melanoma de- differentiation in resistance to PD-1 inhibition. *Nature Communications*, *11*(1), 1897. https://doi.org/10.1038/s41467-020-15726-7

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Zhu, M., Pan, R., Fang, T., Cao, Y.-Y., Chen, S., Zhao, X., Lei, C.-Q., Guo, L., Chen, Y., Li, C.-M., Jokitalo, E., Yin, Y., Shu, H.-B., & Guo, D. (2016). The tumor suppressor PTEN has a critical role in antiviral innate immunity. *Nature Immunology*, *17*(3), 241–249. https://doi.org/10.1038/ni.3311

Lin, H., Wei, S., Hurt, E. M., Green, M. D., Zhao, L., Vatan, L., Szeliga, W., Herbst, R., Harms, P. W., Fecher, L. A., Vats, P., Chinnaiyan, A. M., Lao, C. D., Lawrence, T. S., Wicha, M., Hamanishi, J., Mandai, M., Kryczek, I., & Zou, W. (2018). Host expression of PD-L1

determines efficacy of PD-L1 pathway blockade-mediated tumor regression. *The Journal of Clinical Investigation*, *128*(2), 805–815. https://doi.org/10.1172/JCI96113

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Maksymiuk, S., Gosiewska, A., & Biecek, P. (2020). *Landscape of R packages for eXplainable Artificial Intelligence*.

Mariathasan, S., Turley, S. J., Nickles, D., Castiglioni, A., Yuen, K., Wang, Y., Kadel, E. E., Koeppen, H., Astarita, J. L., Cubas, R., Jhunjhunwala, S., Banchereau, R., Yang, Y., Guan, Y., Chalouni, C., Ziai, J., Şenbabaoğlu, Y., Santoro, S., Sheinson, D., … Powles, T. (2018). TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature*, *554*(7693), 544–548. https://doi.org/10.1038/nature25501

Masternak, K., Muhlethaler-Mottet, A., Villard, J., Zufferey, M., Steimle, V., & Reith, W. (2000). CIITA is a transcriptional coactivator that is recruited to MHC class II promoters by multiple synergistic interactions with an enhanceosome complex. *Genes & Development*, *14*(9), 1156–1166.

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B. K., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., van Allen, E. M., … Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science (New York, N.Y.)*, *351*(6280), 1463–1469. https://doi.org/10.1126/science.aaf1490

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., … Kent, W. J. (2013). The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Research*, *41*(D1), 64–69. https://doi.org/10.1093/nar/gks1048

Mitchell, M. W. (2011). Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. *Open Journal of Statistics*, *01*(03), 205–211. https://doi.org/10.4236/ojs.2011.13024

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of Statistical Software*, *50*(11), 1–23. https://doi.org/10.18637/jss.v050.i11

Motzer, R. J., Escudier, B., George, S., Hammers, H. J., Srinivas, S., Tykodi, S. S., Sosman, J. A., Plimack, E. R., Procopio, G., McDermott, D. F., Castellano, D., Choueiri, T. K., Donskov,

F., Gurney, H., Oudard, S., Richardet, M., Peltola, K., Alva, A. S., Carducci, M., … Tannir, N. M. (2020). Nivolumab versus everolimus in patients with advanced renal cell carcinoma: Updated results with long-term follow-up of the randomized, open-label, phase 3 CheckMate 025 trial. *Cancer*, *126*(18), 4156–4167. https://doi.org/10.1002/cncr.33033

Nagarsheth, N., Wicha, M. S., & Zou, W. (2017). Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. *Nature Reviews. Immunology*, *17*(9), 559–572. https://doi.org/10.1038/nri.2017.49

Nassar, A. H., Adib, E., Abou Alaiwi, S., el Zarif, T., Groha, S., Akl, E. W., Nuzzo, P. V., Mouhieddine, T. H., Perea-Chamblee, T., Taraszka, K., El-Khoury, H., Labban, M., Fong, C., Arora, K. S., Labaki, C., Xu, W., Sonpavde, G., Haddad, R. I., Mouw, K. W., … Gusev, A. (2022). Ancestry-driven recalibration of tumor mutational burden and disparate clinical outcomes in response to immune checkpoint inhibitors. *Cancer Cell*, *40*(10), 1161-1172.e5. https://doi.org/10.1016/j.ccell.2022.08.022

Necchi, A., Joseph, R. W., Loriot, Y., Hoffman-Censits, J., Perez-Gracia, J. L., Petrylak, D. P., Derleth, C. L., Tayama, D., Zhu, Q., Ding, B., Kaiser, C., & Rosenberg, J. E. (2017). Atezolizumab in platinum-treated locally advanced or metastatic urothelial carcinoma: post-progression outcomes from the phase II IMvigor210 study. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *28*(12), 3044–3050. https://doi.org/10.1093/annonc/mdx518

Newell, F., Pires da Silva, I., Johansson, P. A., Menzies, A. M., Wilmott, J. S., Addala, V., Carlino, M. S., Rizos, H., Nones, K., Edwards, J. J., Lakis, V., Kazakoff, S. H., Mukhopadhyay, P., Ferguson, P. M., Leonard, C., Koufariotis, L. T., Wood, S., Blank, C. U., Thompson, J. F., … Long, G. v. (2022). Multiomic profiling of checkpoint inhibitor-treated melanoma: Identifying predictors of response and resistance, and markers of biological discordance. *Cancer Cell*, *40*(1), 88-102.e7. https://doi.org/10.1016/j.ccell.2021.11.012

Nguyen, T.-T., Ramsay, L., Ahanfeshar-Adams, M., Lajoie, M., Schadendorf, D., Alain, T., & Watson, I. R. (2021). Mutations in the IFNγ-JAK-STAT Pathway Causing Resistance to Immune Checkpoint Inhibitors in Melanoma Increase Sensitivity to Oncolytic Virus Treatment. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *27*(12), 3432–3442. https://doi.org/10.1158/1078-0432.CCR-20-3365

Noguchi, T., Ward, J. P., Gubin, M. M., Arthur, C. D., Lee, S. H., Hundal, J., Selby, M. J., Graziano, R. F., Mardis, E. R., Korman, A. J., & Schreiber, R. D. (2017). Temporally Distinct PD-L1 Expression by Tumor and Host Cells Contributes to Immune Escape. *Cancer Immunology Research*, *5*(2), 106–117. https://doi.org/10.1158/2326-6066.CIR-16-0391

O'Brien, R. C., Ishwaran, H., Szczotka-Flynn, L. B., Lass, J. H., & Cornea Preservation Time Study (CPTS) Group. (2021). Random Survival Forests Analysis of Intraoperative Complications as Predictors of Descemet Stripping Automated Endothelial Keratoplasty

Graft Failure in the Cornea Preservation Time Study. *JAMA Ophthalmology*, *139*(2), 191–197. https://doi.org/10.1001/jamaophthalmol.2020.5743

Patterson, A., & Auslander, N. (2022). Mutated processes predict immune checkpoint inhibitor therapy benefit in metastatic melanoma. *Nature Communications*, *13*(1), 5151. https://doi.org/10.1038/s41467-022-32838-4

Peng, W., Chen, J. Q., Liu, C., Malu, S., Creasy, C., Tetzlaff, M. T., Xu, C., McKenzie, J. A., Zhang, C., Liang, X., Williams, L. J., Deng, W., Chen, G., Mbofung, R., Lazar, A. J., Torres-Cabala, C. A., Cooper, Z. A., Chen, P.-L., Tieu, T. N., … Hwu, P. (2016). Loss of PTEN Promotes Resistance to T Cell-Mediated Immunotherapy. *Cancer Discovery*, *6*(2), 202–216. https://doi.org/10.1158/2159-8290.CD-15-0283

Powles, T., Csőszi, T., Özgüroğlu, M., Matsubara, N., Géczi, L., Cheng, S. Y.-S., Fradet, Y., Oudard, S., Vulsteke, C., Morales Barrera, R., Fléchon, A., Gunduz, S., Loriot, Y., Rodriguez-Vida, A., Mamtani, R., Yu, E. Y., Nam, K., Imai, K., Homet Moreno, B., … KEYNOTE-361 Investigators. (2021). Pembrolizumab alone or combined with chemotherapy versus chemotherapy as first-line therapy for advanced urothelial carcinoma (KEYNOTE-361): a randomised, open-label, phase 3 trial. *The Lancet. Oncology*, *22*(7), 931–945. https://doi.org/10.1016/S1470-2045(21)00152-2

Powles, T., Durán, I., van der Heijden, M. S., Loriot, Y., Vogelzang, N. J., de Giorgi, U., Oudard, S., Retz, M. M., Castellano, D., Bamias, A., Fléchon, A., Gravis, G., Hussain, S., Takano, T., Leng, N., Kadel, E. E., Banchereau, R., Hegde, P. S., Mariathasan, S., … Ravaud, A. (2018). Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *Lancet (London, England)*, *391*(10122), 748–757. https://doi.org/10.1016/S0140-6736(17)33297-X

Powles, T., Eder, J. P., Fine, G. D., Braiteh, F. S., Loriot, Y., Cruz, C., Bellmunt, J., Burris, H. A., Petrylak, D. P., Teng, S., Shen, X., Boyd, Z., Hegde, P. S., Chen, D. S., & Vogelzang, N. J. (2014). MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature*, *515*(7528), 558–562. https://doi.org/10.1038/nature13904

Powles, T., Park, S. H., Voog, E., Caserta, C., Valderrama, B. P., Gurney, H., Kalofonos, H., Radulović, S., Demey, W., Ullén, A., Loriot, Y., Sridhar, S. S., Tsuchiya, N., Kopyltsov, E., Sternberg, C. N., Bellmunt, J., Aragon-Ching, J. B., Petrylak, D. P., Laliberte, R., … Grivas, P. (2020). Avelumab Maintenance Therapy for Advanced or Metastatic Urothelial Carcinoma. *The New England Journal of Medicine*, *383*(13), 1218–1230. https://doi.org/10.1056/NEJMoa2002788

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., Bhatia, S., Hwu, W.-J., Gajewski, T. F., Slingluff, C. L., Chowell, D., Kendall, S. M., Chang, H., Shah, R., Kuo, F., … Chan, T.

A. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, *171*(4), 934-949.e16. https://doi.org/10.1016/j.cell.2017.09.028

Ribas, A., Lawrence, D., Atkinson, V., Agarwal, S., Miller, W. H., Carlino, M. S., Fisher, R., Long, G. v, Hodi, F. S., Tsoi, J., Grasso, C. S., Mookerjee, B., Zhao, Q., Ghori, R., Moreno, B. H., Ibrahim, N., & Hamid, O. (2019). Combined BRAF and MEK inhibition with PD-1 blockade immunotherapy in BRAF-mutant melanoma. *Nature Medicine*, *25*(6), 936–940. https://doi.org/10.1038/s41591-019-0476-5

Robinson, D., van Allen, E. M., Wu, Y.-M., Schultz, N., Lonigro, R. J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C. C., Attard, G., Beltran, H., Abida, W., Bradley, R. K., Vinson, J., Cao, X., Vats, P., Kunju, L. P., Hussain, M., Feng, F. Y., … Chinnaiyan, A. M. (2015). Integrative clinical genomics of advanced prostate cancer. *Cell*, *161*(5), 1215–1228. https://doi.org/10.1016/j.cell.2015.05.001

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3). https://doi.org/10.1186/gb-2010-11-3-r25

Roche. (2021, March 7). *Roche provides update on Tecentriq US indication in prior-platinum treated metastatic bladder cancer*. https://www.roche.com/media/releases/med-cor-2021-03-08

Roche, P. A., & Furuta, K. (2015). The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature Reviews. Immunology*, *15*(4), 203–216. https://doi.org/10.1038/nri3818

Rosenberg, J. E., Hoffman-Censits, J., Powles, T., van der Heijden, M. S., Balar, A. v, Necchi, A., Dawson, N., O'Donnell, P. H., Balmanoukian, A., Loriot, Y., Srinivas, S., Retz, M. M., Grivas, P., Joseph, R. W., Galsky, M. D., Fleming, M. T., Petrylak, D. P., Perez-Gracia, J. L., Burris, H. A., … Dreicer, R. (2016). Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet (London, England)*, *387*(10031), 1909–1920. https://doi.org/10.1016/S0140-6736(16)00561-4

Rozeman, E. A., Hoefsmit, E. P., Reijers, I. L. M., Saw, R. P. M., Versluis, J. M., Krijgsman, O., Dimitriadis, P., Sikorska, K., van de Wiel, B. A., Eriksson, H., Gonzalez, M., Torres Acosta, A., Grijpink-Ongering, L. G., Shannon, K., Haanen, J. B. A. G., Stretch, J., Ch'ng, S., Nieweg, O. E., Mallo, H. A., … Blank, C. U. (2021). Survival and biomarker analyses from the OpACIN-neo and OpACIN neoadjuvant immunotherapy trials in stage III melanoma. *Nature Medicine*, *27*(2), 256–263. https://doi.org/10.1038/s41591-020-01211-7

Sade-Feldman, M., Jiao, Y. J., Chen, J. H., Rooney, M. S., Barzily-Rokni, M., Eliane, J.-P., Bjorgaard, S. L., Hammond, M. R., Vitzthum, H., Blackmon, S. M., Frederick, D. T., Hazar-Rethinam, M., Nadres, B. A., van Seventer, E. E., Shukla, S. A., Yizhak, K., Ray, J. P., Rosebrock, D., Livitz, D., … Hacohen, N. (2017). Resistance to checkpoint blockade

therapy through inactivation of antigen presentation. *Nature Communications*, *8*(1), 1136. https://doi.org/10.1038/s41467-017-01062-w

Sehgal, K., Gill, R. R., Widick, P., Bindal, P., McDonald, D. C., Shea, M., Rangachari, D., & Costa, D. B. (2021). Association of Performance Status With Survival in Patients With Advanced Non-Small Cell Lung Cancer Treated With Pembrolizumab Monotherapy. *JAMA Network Open*, *4*(2), e2037120. https://doi.org/10.1001/jamanetworkopen.2020.37120

Semeraro, F., Parrinello, G., Cancarini, A., Pasquini, L., Zarra, E., Cimino, A., Cancarini, G., Valentini, U., & Costagliola, C. (2011). Predicting the risk of diabetic retinopathy in type 2 diabetic patients. *Journal of Diabetes and Its Complications*, *25*(5), 292–297. https://doi.org/10.1016/j.jdiacomp.2010.12.002

Shapley, L. S. (1953). A Value for n-person Games. In *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307--318). Princeton University Press.

Sheng, W., LaFleur, M. W., Nguyen, T. H., Chen, S., Chakravarthy, A., Conway, J. R., Li, Y., Chen, H., Yang, H., Hsu, P.-H., van Allen, E. M., Freeman, G. J., de Carvalho, D. D., He, H. H., Sharpe, A. H., & Shi, Y. (2018). LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade. *Cell*, *174*(3), 549-563.e19. https://doi.org/10.1016/j.cell.2018.05.052

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Snider, L., Geng, L. N., Lemmers, R. J. L. F., Kyba, M., Ware, C. B., Nelson, A. M., Tawil, R., Filippova, G. N., van der Maarel, S. M., Tapscott, S. J., & Miller, D. G. (2010). Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genetics*, *6*(10), e1001181. https://doi.org/10.1371/journal.pgen.1001181

Spens, A. E., Sutliff, N. A., Bennett, S. R., Campbell, A. E., & Tapscott, S. J. (2022). Human DUX4 and mouse Dux interact with STAT1 and broadly inhibit interferon-stimulated gene induction. *BioRxiv*. https://doi.org/10.1101/2022.08.09.503314

Spranger, S., Bao, R., & Gajewski, T. F. (2015). Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity. *Nature*, *523*(7559), 231–235. https://doi.org/10.1038/nature14404

Steimle, V., Otten, L. A., Zufferey, M., & Mach, B. (1993). Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). *Cell*, *75*(1), 135–146.

Steimle, V., Siegrist, C. A., Mottet, A., Lisowska-Grospierre, B., & Mach, B. (1994). Regulation of MHC class II expression by interferon-gamma mediated by the transactivator gene CIITA. *Science (New York, N.Y.)*, *265*(5168), 106–109. https://doi.org/10.1126/science.8016643

Stein, A., Paschold, L., Tintelnot, J., Goekkurt, E., Henkes, S.-S., Simnica, D., Schultheiss, C., Willscher, E., Bauer, M., Wickenhauser, C., Thuss-Patience, P., Lorenzen, S., Ettrich, T., Riera-Knorrenschild, J., Jacobasch, L., Kretzschmar, A., Kubicka, S., Al-Batran, S.-E., Reinacher-Schick, A., … Binder, M. (2022). Efficacy of Ipilimumab vs FOLFOX in Combination With Nivolumab and Trastuzumab in Patients With Previously Untreated ERBB2-Positive Esophagogastric Adenocarcinoma: The AIO INTEGA Randomized Clinical Trial. *JAMA Oncology*, *8*(8), 1150–1158. https://doi.org/10.1001/jamaoncol.2022.2228

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sucker, A., Zhao, F., Pieper, N., Heeke, C., Maltaner, R., Stadtler, N., Real, B., Bielefeld, N., Howe, S., Weide, B., Gutzmer, R., Utikal, J., Loquai, C., Gogas, H., Klein-Hitpass, L., Zeschnigk, M., Westendorf, A. M., Trilling, M., Horn, S., … Paschen, A. (2017). Acquired IFNγ resistance impairs anti-tumor immunity and gives rise to T-cell-resistant melanoma lesions. *Nature Communications*, *8*, 15440. https://doi.org/10.1038/ncomms15440

Sucker, A., Zhao, F., Real, B., Heeke, C., Bielefeld, N., Maβen, S., Horn, S., Moll, I., Maltaner, R., Horn, P. A., Schilling, B., Sabbatino, F., Lennerz, V., Kloor, M., Ferrone, S., Schadendorf, D., Falk, C. S., Griewank, K., & Paschen, A. (2014). Genetic evolution of T-cell resistance in the course of melanoma progression. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *20*(24), 6593–6604. https://doi.org/10.1158/1078-0432.CCR-14-0567

Sugie, K., Funaya, S., Kawamura, M., Nakamura, T., Suzuki, M. G., & Aoki, F. (2020). Expression of Dux family genes in early preimplantation embryos. *Scientific Reports*, *10*(1), 19396. https://doi.org/10.1038/s41598-020-76538-9

Sullivan, R. J., Hamid, O., Gonzalez, R., Infante, J. R., Patel, M. R., Hodi, F. S., Lewis, K. D., Tawbi, H. A., Hernandez, G., Wongchenko, M. J., Chang, Y., Roberts, L., Ballinger, M., Yan, Y., Cha, E., & Hwu, P. (2019). Atezolizumab plus cobimetinib and vemurafenib in BRAF-mutated melanoma patients. *Nature Medicine*, *25*(6), 929–935. https://doi.org/10.1038/s41591-019-0474-7

Tawbi, H. A., Schadendorf, D., Lipson, E. J., Ascierto, P. A., Matamala, L., Castillo Gutiérrez, E., Rutkowski, P., Gogas, H. J., Lao, C. D., de Menezes, J. J., Dalle, S., Arance, A., Grob, J.-J., Srivastava, S., Abaskharoun, M., Hamilton, M., Keidel, S., Simonsen, K. L., Sobiesk, A. M., … RELATIVITY-047 Investigators. (2022). Relatlimab and Nivolumab versus Nivolumab in Untreated Advanced Melanoma. *The New England Journal of Medicine*, *386*(1), 24–34. https://doi.org/10.1056/NEJMoa2109970

Therneau, T. (2022). *A Package for Survival Analysis in R* (R package version 3.4-0). https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer New York. https://doi.org/10.1007/978-1-4757-3294-8

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

USFDA. (2016, May 18). *FDA approves new, targeted treatment for bladder cancer*. https://www.fda.gov/news-events/press-announcements/fda-approves-new-targeted-treatment-bladder-cancer

Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W., & Tapscott, S. J. (2017). Conservation and innovation in the DUX4-family gene network. *Nature Genetics*, *49*(6), 935–940. https://doi.org/10.1038/ng.3846

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation* (R package version 1.0.10). https://CRAN.R-project.org/package=dplyr

Wolf, Y., Bartok, O., Patkar, S., Eli, G. B., Cohen, S., Litchfield, K., Levy, R., Jiménez-Sánchez, A., Trabish, S., Lee, J. S., Karathia, H., Barnea, E., Day, C.-P., Cinnamon, E., Stein, I., Solomon, A., Bitton, L., Pérez-Guijarro, E., Dubovik, T., … Samuels, Y. (2019). UVB-Induced Tumor Heterogeneity Diminishes Immune Response in Melanoma. *Cell*, *179*(1), 219-235.e21. https://doi.org/10.1016/j.cell.2019.08.032

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Cambridge (Mass.))*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, Q., Bi, J., Zheng, X., Chen, Y., Wang, H., Wu, W., Wang, Z., Wu, Q., Peng, H., Wei, H., Sun, R., & Tian, Z. (2018). Blockade of the checkpoint receptor TIGIT prevents NK cell exhaustion and elicits potent anti-tumor immunity. *Nature Immunology*, *19*(7), 723–732. https://doi.org/10.1038/s41590-018-0132-0

Chapter 6. DISCUSSION

The body of work outlined in this dissertation is comprised of distinct cases demonstrating the utility of large-scale analytical strategies to uncover and extract novel insights from existing publicly available genomics data. Specifically, we provide an increased mechanistic understanding of splicing and its regulation, identification of splicing-based strategies to improve drug response, and discernment of factors contributing to cancer therapeutic response. The potential implications of these works for cancer biology and treatment, and future directions, are discussed below.

6.1.1 *Utilizing branchpoint nucleotides to study dysregulated alternative splicing in cancer*

Chapter 2 describes a method to identify the positions of branchpoint nucleotides— obligate signals in splicing catalysis— in the human genome, uncovering the unexpected contribution of branchpoints to splicing regulation and providing an alternative means for examining the consequences of splicing-augmenting variants. I anticipate our annotation to be of broad utility to scientists investigating the significance of intronic disease-associated variants, the functional consequences of oncogenic mutations in splicing factors, and the characterization of non-canonical splicing modalities.

Intronic variants are generally linked with splicing perturbation due to interference with 5′ and 3′ splice site recognition (resulting in either ablation of the signal or introduction of an alternative site), or disruption of cis-acting regulatory element (intronic splicing enhancers or silencers) or RNA-binding protein motif sequences (Scotti & Swanson, 2016; Sveen et al., 2016). Notably, few variants have been linked to branchpoint sequences, possibly due to the previous deficiency in knowledge of branchpoint locations or the buffering effect of branchpoint multiplicity, which is characteristic of human introns (Pineda & Bradley, 2018). Mercer et al.

(2015) examined retinoblastoma and lung adenocarcinoma patients and identified variants in the branchpoint nucleotides within the introns of *RB1* and *MET*, respectively. These mutations induced aberrant exon exclusion in their corresponding parental transcripts (Mercer et al., 2015; Onozato et al., 2009; K. Zhang et al., 2008). Using our annotation, which represents a significantly expanded 'branchpoint map,' we resolved the pathogenicity of deep intronic variants in the gene *LZTR1* (Inoue et al., 2021), linked to Schwannomatosis (Piotrowski et al., 2014) and familial autosomal recessive Noonan syndrome (Johnston et al., 2018), both cancer pre-disposition syndromes. Here, we showed that mutations occurring within the branchpoint sequence of a minor intron in *LZTR1* resulted in intron retention and subsequent transcript down-regulation via nonsense-mediated decay. Recently, Zhang, et al. (2022) used our annotation to interpret somatic mutations from the Catalogue of Somatic Mutations in Cancer database which disrupt the branchpoint sequence within intronic regions. Here, they identified an intronic 59-nucleotide deletion in a lymphoma patient which simultaneously ablated 3 branchpoint locations in intron 3 of *ITPKB* (Tate et al., 2019; P. Zhang et al., 2022). These studies display examples of cancer-relevant mutations in branchpoint sequences and highlight the importance of a 'branchpoint map' to interpret somatic and germline variants occurring within introns.

Splicing fidelity is tightly controlled in human cells and its dysregulation, via mutations in genes encoding components of the spliceosome, drives the genesis and development hematological and solid cancers (Bonnal et al., 2020; Dvinge et al., 2016; Lee & Abdel-Wahab, 2016). For instance, *ZRSR2* is a component of the minor spliceosome commonly mutated in myelodysplastic syndromes and acute myeloid leukemia patients (Madan et al., 2015). Interestingly, our recent characterization of *ZRSR2*-mutant-associated minor intron retention in myeloid malignancies implicated branchpoint sequence features as potential determinants of *ZRSR2* mutant-induced

missplicing (Inoue et al., 2021). Our branchpoint annotation may similarly aid the investigation into the effects of mutations in *SF3B1*, a splicing factor involved in branchpoint selection. *SF3B1* is the most recurrently mutated splicing factor in cancer (Seiler et al., 2018) and is observed in hematopoietic malignancies (Gentien et al., 2014; Haferlach et al., 2014; Makishima et al., 2012; Malcovati et al., 2015; Papaemmanuil et al., 2011, 2013; Shiozawa et al., 2018; Wang et al., 2011; Yoshida et al., 2011) and a variety of solid tumors such as uveal melanoma and papillary and mucinous carcinomas of the breast (Furney et al., 2013; Harbour et al., 2013; Maguire et al., 2015; Martin et al., 2013; Seiler et al., 2018). These gain or alteration of function mutations occur in a heterozygous context, and cluster tightly within sequences corresponding to the HEAT Repeat (HR) domains of *SF3B1* (Dvinge et al., 2016). These HR domains are poorly characterized, contributing to difficulty in determining the precise molecular basis for splicing dysregulation in *SF3B1* mutants. Changes in 3′ splice site preference is appreciated to be a hallmark feature of *SF3B1* mutant transcriptomes. Several studies, which can be categorized according to branchpoint usage, have posited mechanistic hypotheses to explain this observation. The first group proposes the use of mutant-specific branchpoints leading to a corresponding selection of upstream cryptic 3′ splice sites (Alsafadi et al., 2016; Darman et al., 2015). The second set of reports proposes no shift in branchpoint preference, but a change in accessibility to alternative 3′ splice sites due to secondary RNA structure (Kesarwani et al., 2017), steric hindrance from *SF3B1* mutant-associated conformational changes (DeBoever et al., 2015), or changes in the interactions between mutant *SF3B1* and essential RNA-helicases (Z. Liu et al., 2020; J. Zhang et al., 2019). The models put forth lack consensus. Further, these models rely on computationally predicted branchpoint positions (DeBoever et al., 2015) or branchpoints identified in single-intron contexts (Darman et al., 2015; B. Liu et al., 2021)— carrying problems related to extrapolating from a small number of

select observations. Thus, determining *SF3B1* mutant branchpoint nucleotide preference is central to resolving the basis for splicing dysregulation in these cancers— an effort that is currently underway with the aid of our 'branchpoint map.' Importantly, I expect inquiry into the influence of branchpoint selection in *SF3B1* mutant splicing to shed light on the mechanisms behind mutant splicing changes outside alternative 3′ splice site selection, such as changes in alternative 5′ splice site selection, intron retention, and exon skipping (Tang et al., 2020). Our recent efforts have revealed usage of an alternative branchpoint associated with poison exon inclusion and subsequent repression of *BRD9*, a member of the non-canonical chromatin-remodeling complex in *SF3B1*-mutated cancers (Inoue et al., 2019). Our computational and experimental methods to detect branchpoints have also been used to help optimize the design of synthetic introns, which display *SF3B1* mutant-specific splicing, for directed tumor repression while sparing healthy host tissue (North et al., 2022).

Lariat sequencing (branchpoint profiling) can aid in the investigation non-canonical splicing mechanisms. We recently studied recursive splicing (RS) in humans: intron excision accomplished in a piece-wise manner (Blazquez et al., 2018; Gehring & Roignant, 2021). Specifically, we investigated the exon definition model of RS which is characterized by the use of short cryptic exons (RS exons) containing a splice donor site immediately adjacent to the 3′ splice site upstream of the exon (Joseph et al., 2018; Sibley et al., 2015). Here, we computationally identified putative lariats associated with RS exon splicing from bulk RNA-seq data: 287 high-confidence pairs which link branchpoint positions to the start (5′ splice site) of RS exons. We experimentally confirmed RS lariats associated with the gene *AP1G2* using direct lariat sequencing on cerebellum total RNA to indicate that the exon definition model of recursive splicing is employed in human transcriptomes. RS deregulation could occur outside normal physiological

conditions such as in cancer; our study provides a framework to investigate such processes through an annotation of RS-exon-associated branchpoints, the first of its kind.

6.1.2 *Cintron-based cancer vaccines to enhance the efficacy of immunotherapy*

Chapter 3 details the first systematic characterization of circular RNAs derived from introns. We demonstrate that these largely unrecognized RNA products, "cintrons," are produced by most human genes. To date, almost all research into circular RNAs has focused on circular RNAs derived from exons, a well-studied class whose mechanistic origins via backsplicing and functional roles in regulating gene expression and splicing have been elucidated in many studies (Kristensen et al., 2019). While the origin and function of cintrons remain to be elucidated, I hypothesize that these molecules could be engineered as a vehicle for cancer vaccines.

Splicing-based strategies for RNA circularization exist. First, there is the 'Tornado' expression system based on tRNA splicing. Here, an autocatalytic transcript releases an intron containing a hydroxyl group and a 2′,3′-cyclic phosphate at the 5′ and 3′ ends, respectively. These functional groups are recognized and joined by the RNA ligase *RTCB* resulting in circularization (Litke & Jaffrey, 2019). Second, synthesis of circular transcripts could be achieved via the permuted intron-exon (PIE) splicing strategy based on group I intron self-splicing (Wesselhoeft et al., 2018). In this system, the two exons upstream and downstream of a self-splicing intron are inverted and joined— resulting in fused partial exons and flanking partial introns. This structure retains the ability to undergo the two-step transesterification reactions. However, the resultant circularization of the fused exons is observed in this case due to the inverted nature of the sequence (reversed 5′ and 3′ splice sites). The first reports of circular RNA synthesis using the PIE strategy was based on a T4 phage intron (Ford & Ares, 1994; Puttaraju & Been, 1992). These studies also showed that circularization can proceed even with the introduction of foreign sequences ('cargo')

within the fused exonic sequence. The process has been further optimized (increased efficiency and yield) by using an engineered *Anabaena* pre-tRNA sequence (Wesselhoeft et al., 2018).

Circular RNAs have several advantageous properties for vaccination purposes. First, circular transcripts can be translated via an internal ribosome entry site (IRES) which eliminates the need for capping or polyadenylation. Second, circularization significantly increases the half-life of the RNA compared to its linear counterpart due to their inherent resistance to exonucleolytic degradation. This feature also increases the amount of protein that could be produced from translation of a single molecule. Lastly, while RNA itself is naturally immunogenic, circularization has been shown to escape Toll-like receptor activation and *RIG-I* signaling (Wesselhoeft et al., 2019).

Cintrons could be a promising alternative to the aforementioned circularization platforms for RNA-based cancer vaccines (Lorentzen et al., 2022). I envision a cintron-based IRES-driven system for tumor-specific (or tumor associated) antigen expression administered in combination with immunotherapy (immune checkpoint inhibition or chimeric antigen receptor therapy) to boost the efficacy of these regimens. We are currently investigating strategies to increase the efficiency of cintron production. Our results suggest that cintrons are derived from intron lariats, and the rate of conversion is primarily determined by potential for intramolecular binding and length of the lariat-tail. We plan to investigate the following intronic sequence modifications: modulating lariat tail length; modifying branchpoint nucleotide location, nucleotide identity, and number; and introducing complementary sequences (e.g., RNA hairpins) to promote BP-3′ splice site proximity in the lariat conformation. We also aim to investigate factors that could further reduce the immunogenicity of circular RNAs, such as introduction of chemically-modified nucleosides in the cintronic sequence (Martin & Lowery, 2020). Finally, our study found that cintron-yielding human

introns can be very long, presenting the possibility of circularizing large cargos and expression of large or multiple proteins form a single circular transcript.

6.1.3 *Improving patient response to cancer therapeutics*

Chapters 4 and 5 discussed strategies to improve cancer therapeutic response. In Chapter 4, we unified *in vitro* and *in vivo* experiments with analyses of acute-myeloid leukemia (AML) cell line RNA-seq data. We also linked our findings to patient transcriptomic and drug response profiles from the BeatAML study, a large reference cohort of AML patients (Tyner et al., 2018). Overall, we found that AML harbors a unique susceptibility to splicing interference, and our integrated analyses identified specific genetic (*RBM10* deletion) and pharmacologic (the novel splicing-dependent kinase inhibitor SM09419) splicing-modulating combinatorial strategies which controlled sensitivity and resistance to *BCL2* inhibition (venetoclax). In Chapter 5, we processed RNA-seq data from over 20,000 patients with early-stage and advanced cancers to determine that *DUX4* expression was a common strategy for immune evasion and resistance to immune checkpoint inhibition. In both cases, large-scale integrative analyses of numerous RNA-seq datasets were fundamental. Recently, we undertook a similar approach to identify gene expression signatures associated with response to a novel *CBP*/*P300* peptidomimetic termed CRYBMIM in *MYB*-driven AML: interference of *MYB*-associated oncogenic signaling (Takao et al., 2021).

I hypothesize that the complementation of large-scale genomic analyses with machine-learning methods will be an effective method to identify genetic factors associated with cancer development, progression, and response to therapy. The approaches we employed provided important insights insofar as being limited by the information included in the analyses. For instance, in Chapter 5, the initial survey was performed by solely examining *DUX4*-expressing vs. -silent comparisons across various cancers, before quantifying the effect on patient survival post-

hoc. Machine learning algorithms can simultaneously integrate multimodal and multidimensional data (e.g., patient genetic and phenotypic variables). This approach could be advantageous in terms of selecting specific clinically contextualized candidates to validate before the required, and often expensive, experimental approaches are employed (*in vitro* and *in vivo* models of cancer). One recent preprint has exhibited this potential by aggregating the predictions of diverse machine learning models to identify gene expression signatures associated with response to combinatorial drug treatment strategies in AML (Janizek et al., 2021).

6.1.4 *Closing remarks*

While the exploration framework for cancer genomic data remains challenging to implement at present, and often inaccessible, its democratization is imminent. Analysis of sequencing data requires the integration of data processing and its concomitant analysis— through the choreography of a multitude of software dependencies (e.g., reference genomes and annotations, functionalities from external libraries) and the associated hardware requisites— with specific domain expertise. Numerous computational tools have become readily accessible to streamline this workflow. For instance, the Partek Genomics Suite (https://www.partek.com/partek-genomics-suite/), Galaxy (Afgan et al., 2018; Giardine et al., 2005; Jalili et al., 2020), and the Empowering the Development of Genomics Expertise integrated platform (P.-E. Li et al., 2017) permit efficient and reproducible analyses with intuitive usage: the required bioinformatic tools are unified and the data processing is hosted on servers, allowing easy interactive web-based use. Large communities of scientists have organized around these tools (for example, the Galaxy Community: https://galaxyproject.org/) for continued development and growth. Similarly, programs which streamline and automate the construction and application of machine learning models are common and available (Bischl et al., 2016; Chollet, 2015; H2O.ai, 2020; Kuhn, 2022;

Kuhn & Wickham, 2020; Lang et al., 2019; Paszke et al., 2019; Pedregosa et al., 2011). These tools will likely catalyze scientists' and clinicians' increased capability to analyze cancer genomics data in parallel and at a massive scale. As the magnitude of available human sequencing data from research and clinical use exponentially grows, forecasting of the value of such approaches will undoubtedly yield underestimates.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Alsafadi, S., Houy, A., Battistella, A., Popova, T., Wassef, M., Henry, E., Tirode, F., Constantinou, A., Piperno-Neumann, S., Roman-Roman, S., Dutertre, M., & Stern, M. H. (2016). Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms10615

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Blazquez, L., Emmett, W., Faraway, R., Pineda, J. M. B., Bajew, S., Gohr, A., Haberman, N., Sibley, C. R., Bradley, R. K., Irimia, M., & Ule, J. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Molecular Cell*, *72*(3), 496-509.e9. https://doi.org/10.1016/j.molcel.2018.09.033

Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/s41571-020-0350-x

Chollet, F. et al. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Darman, R. B., Seiler, M., Agrawal, A. A., Lim, K. H., Peng, S., Aird, D., Bailey, S. L., Bhavsar, E. B., Chan, B., Colla, S., Corson, L., Feala, J., Fekkes, P., Ichikawa, K., Keaney, G. F., Lee, L., Kumar, P., Kunii, K., MacKenzie, C., … Buonamici, S. (2015). Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Reports*, *13*(5), 1033–1045. https://doi.org/10.1016/j.celrep.2015.09.053

DeBoever, C., Ghia, E. M., Shepard, P. J., Rassenti, L., Barrett, C. L., Jepsen, K., Jamieson, C. H. M., Carson, D., Kipps, T. J., & Frazer, K. A. (2015). Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3' Splice Site Selection in SF3B1-mutated Cancers. *PLoS Computational Biology*, *11*(3), 1–19. https://doi.org/10.1371/journal.pcbi.1004105

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews. Cancer*, *16*(7), 413–430. https://doi.org/10.1038/nrc.2016.51

Ford, E., & Ares, M. (1994). Synthesis of circular RNA in bacteria and yeast using RNA cyclase ribozymes derived from a group I intron of phage T4. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(8), 3117–3121. https://doi.org/10.1073/pnas.91.8.3117

Furney, S. J., Pedersen, M., Gentien, D., Dumont, A. G., Rapinat, A., Desjardins, L., Turajlic, S., Piperno-Neumann, S., de la Grange, P., Roman-Roman, S., Stern, M.-H., & Marais, R. (2013). SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discovery*, *3*(10), 1122–1129. https://doi.org/10.1158/2159-8290.CD-13-0330

Gehring, N. H., & Roignant, J.-Y. (2021). Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends in Genetics : TIG*, *37*(4), 355–372. https://doi.org/10.1016/j.tig.2020.10.008

Gentien, D., Kosmider, O., Nguyen-Khac, F., Albaud, B., Rapinat, A., Dumont, A. G., Damm, F., Popova, T., Marais, R., Fontenay, M., Roman-Roman, S., Bernard, O. A., & Stern, M.-H. (2014). A common alternative splicing signature is associated with SF3B1 mutations in malignancies from different cell lineages. *Leukemia*, *28*(6), 1355–1357. https://doi.org/10.1038/leu.2014.28

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., & Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, *15*(10), 1451–1455. https://doi.org/10.1101/gr.4086505

H2O.ai. (2020). *H2O: Scalable Machine Learning Platform* (version 3.30.0.6). https://github.com/h2oai/h2o-3

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., Yoshida, K., Roller, A., Nadarajah, N., Shiraishi, Y., Shiozawa, Y., Chiba, K., Tanaka, H., Koeffler, H. P., Klein, H.-U., … Ogawa, S. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, *28*(2), 241–247. https://doi.org/10.1038/leu.2013.336

Harbour, J. W., Roberson, E. D. O., Anbunathan, H., Onken, M. D., Worley, L. A., & Bowcock, A. M. (2013). Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nature Genetics*, *45*(2), 133–135. https://doi.org/10.1038/ng.2523

Inoue, D., Chew, G. L., Liu, B., Michel, B. C., Pangallo, J., D'Avino, A. R., Hitchman, T., North, K., Lee, S. C. W., Bitner, L., Block, A., Moore, A. R., Yoshimi, A., Escobar-Hoyos, L., Cho, H., Penson, A., Lu, S. X., Taylor, J., Chen, Y., … Bradley, R. K. (2019). Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature*, *574*(7778), 432–436. https://doi.org/10.1038/s41586-019-1646-9

Inoue, D., Polaski, J. T., Taylor, J., Castel, P., Chen, S., Kobayashi, S., Hogg, S. J., Hayashi, Y., Pineda, J. M. B., el Marabti, E., Erickson, C., Knorr, K., Fukumoto, M., Yamazaki, H., Tanaka, A., Fukui, C., Lu, S. X., Durham, B. H., Liu, B., … Abdel-Wahab, O. (2021). Minor intron retention drives clonal hematopoietic disorders and diverse cancer predisposition. *Nature Genetics*, *53*(5), 707–718. https://doi.org/10.1038/s41588-021-00828-9

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., Taylor, J., & Nekrutenko, A. (2020). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*, *48*(W1), W395–W402. https://doi.org/10.1093/nar/gkaa434

Janizek, J. D., Dincer, A. B., Celik, S., Chen, H., Chen, W., Naxerova, K., & Lee, S.-I. (2021). Uncovering expression signatures of synergistic drug response using an ensemble of explainable AI models. *BioRxiv*.

Johnston, J. J., van der Smagt, J. J., Rosenfeld, J. A., Pagnamenta, A. T., Alswaid, A., Baker, E. H., Blair, E., Borck, G., Brinkmann, J., Craigen, W., Dung, V. C., Emrick, L., Everman, D. B., van Gassen, K. L., Gulsuner, S., Harr, M. H., Jain, M., Kuechler, A., Leppig, K. A., … Biesecker, L. G. (2018). Autosomal recessive Noonan syndrome associated with biallelic LZTR1 variants. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *20*(10), 1175–1185. https://doi.org/10.1038/gim.2017.249

Joseph, B., Kondo, S., & Lai, E. C. (2018). Short cryptic exons mediate recursive splicing in Drosophila. *Nature Structural & Molecular Biology*, *25*(5), 365–371. https://doi.org/10.1038/s41594-018-0052-6

Kesarwani, A. K., Ramirez, O., Gupta, A. K., Yang, X., Murthy, T., Minella, A. C., & Pillai, M. M. (2017). Cancer-associated SF3B1 mutants recognize otherwise inaccessible cryptic 3' splice sites within RNA secondary structures. *Oncogene*, *36*(8), 1123–1133. https://doi.org/10.1038/onc.2016.279

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., & Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. *Nature Reviews Genetics*, *20*(11), 675–691. https://doi.org/10.1038/s41576-019-0158-7

Kuhn, M. (2022). *caret: Classification and Regression Training* (R package version 6.0-93). https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse*. https://www.tidymodels.org

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Lee, S. C. W., & Abdel-Wahab, O. (2016). Therapeutic targeting of splicing in cancer. *Nature Medicine*, *22*(9), 976–986. https://doi.org/10.1038/nm.4165

Li, P.-E., Lo, C.-C., Anderson, J. J., Davenport, K. W., Bishop-Lilly, K. A., Xu, Y., Ahmed, S., Feng, S., Mokashi, V. P., & Chain, P. S. G. (2017). Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Research*, *45*(1), 67–80. https://doi.org/10.1093/nar/gkw1027

Litke, J. L., & Jaffrey, S. R. (2019). Highly efficient expression of circular RNA aptamers in cells using autocatalytic transcripts. *Nature Biotechnology*, *37*(6), 667–675. https://doi.org/10.1038/s41587-019-0090-6

Liu, B., Liu, Z., Chen, S., Ki, M., Erickson, C., Reis-Filho, J. S., Durham, B. H., Chang, Q., de Stanchina, E., Sun, Y., Rabadan, R., Abdel-Wahab, O., & Chandarlapaty, S. (2021). Mutant SF3B1 promotes AKT- and NF-κB-driven mammary tumorigenesis. *The Journal of Clinical Investigation*, *131*(1). https://doi.org/10.1172/JCI138315

Liu, Z., Zhang, J., Sun, Y., Perea-Chamblee, T. E., Manley, J. L., & Rabadan, R. (2020). Pan-cancer analysis identifies mutations in SUGP1 that recapitulate mutant SF3B1 splicing dysregulation. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10305–10312. https://doi.org/10.1073/pnas.1922622117

Lorentzen, C. L., Haanen, J. B., Met, Ö., & Svane, I. M. (2022). Clinical advances and ongoing trials on mRNA vaccines for cancer treatment. *The Lancet. Oncology*, *23*(10), e450–e458. https://doi.org/10.1016/S1470-2045(22)00372-2

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., Sanada, M., Grossmann, V., Sundaresan, J., Shiraishi, Y., Miyano, S., Thol, F., Ganser, A., Yang, H., Haferlach, T., Ogawa, S., & Koeffler, H. P. (2015). Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nature Communications*, *6*, 6042. https://doi.org/10.1038/ncomms7042

Maguire, S. L., Leonidou, A., Wai, P., Marchiò, C., Ng, C. K., Sapino, A., Salomon, A.-V., Reis-Filho, J. S., Weigelt, B., & Natrajan, R. C. (2015). SF3B1 mutations constitute a novel therapeutic target in breast cancer. *The Journal of Pathology*, *235*(4), 571–580. https://doi.org/10.1002/path.4483

Makishima, H., Visconte, V., Sakaguchi, H., Jankowska, A. M., Abu Kar, S., Jerez, A., Przychodzen, B., Bupathi, M., Guinta, K., Afable, M. G., Sekeres, M. A., Padgett, R. A., Tiu, R. v, & Maciejewski, J. P. (2012). Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. *Blood*, *119*(14), 3203–3210. https://doi.org/10.1182/blood-2011-12-399774

Malcovati, L., Karimi, M., Papaemmanuil, E., Ambaglio, I., Jädersten, M., Jansson, M., Elena, C., Gallì, A., Walldin, G., della Porta, M. G., Raaschou-Jensen, K., Travaglino, E., Kallenbach, K., Pietra, D., Ljungström, V., Conte, S., Boveri, E., Invernizzi, R., Rosenquist, R., … Hellström Lindberg, E. (2015). SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts. *Blood*, *126*(2), 233–241. https://doi.org/10.1182/blood-2015-03-633537

Martin, C., & Lowery, D. (2020). mRNA vaccines: intellectual property landscape. *Nature Reviews. Drug Discovery*, *19*(9), 578. https://doi.org/10.1038/d41573-020-00119-8

Martin, M., Maßhöfer, L., Temming, P., Rahmann, S., Metz, C., Bornfeld, N., van de Nes, J., Klein-Hitpass, L., Hinnebusch, A. G., Horsthemke, B., Lohmann, D. R., & Zeschnigk, M. (2013). Exome sequencing identifies recurrent somatic mutations in EIF1AX and SF3B1 in uveal melanoma with disomy 3. *Nature Genetics*, *45*(8), 933–936. https://doi.org/10.1038/ng.2674

Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. *Genome Research*, *25*(2), 290–303. https://doi.org/10.1101/gr.182899.114

North, K., Benbarche, S., Liu, B., Pangallo, J., Chen, S., Stahl, M., Bewersdorf, J. P., Stanley, R. F., Erickson, C., Cho, H., Pineda, J. M. B., Thomas, J. D., Polaski, J. T., Belleville, A. E., Gabel, A. M., Udy, D. B., Humbert, O., Kiem, H.-P., Abdel-Wahab, O., & Bradley, R. K. (2022). Synthetic introns enable splicing factor mutation-dependent targeting of cancer cells. *Nature Biotechnology*, *40*(7), 1103–1113. https://doi.org/10.1038/s41587-022-01224-2

Onozato, R., Kosaka, T., Kuwano, H., Sekido, Y., Yatabe, Y., & Mitsudomi, T. (2009). Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer*, *4*(1), 5–11. https://doi.org/10.1097/JTO.0b013e3181913e0e

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., Godfrey, A. L., Rapado, I., Cvejic, A., Rance, R., McGee, C., Ellis, P., Mudie, L. J., Stephens, P. J., McLaren, S., … Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *The New England Journal of Medicine*, *365*(15), 1384–1395. https://doi.org/10.1056/NEJMoa1103283

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., van Loo, P., Yoon, C. J., Ellis, P., Wedge, D. C., Pellagatti, A., Shlien, A., Groves, M. J., Forbes, S. A., Raine, K., Hinton, J., Mudie, L. J., McLaren, S., Hardy, C., Latimer, C., … Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium. (2013).

Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, *122*(22), 3616–3627; quiz 3699. https://doi.org/10.1182/blood-2013-08-518886

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Jake Vanderplas, Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pineda, J. M. B., & Bradley, R. K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes and Development*, *32*(7–8), 577–591. https://doi.org/10.1101/gad.312058.118

Piotrowski, A., Xie, J., Liu, Y. F., Poplawski, A. B., Gomes, A. R., Madanecki, P., Fu, C., Crowley, M. R., Crossman, D. K., Armstrong, L., Babovic-Vuksanovic, D., Bergner, A., Blakeley, J. O., Blumenthal, A. L., Daniels, M. S., Feit, H., Gardner, K., Hurst, S., Kobelka, C., … Messiaen, L. M. (2014). Germline loss-of-function mutations in LZTR1 predispose to an inherited disorder of multiple schwannomas. *Nature Genetics*, *46*(2), 182–187. https://doi.org/10.1038/ng.2855

Puttaraju, M., & Been, M. D. (1992). Group I permuted intron-exon (PIE) sequences self-splice to produce circular exons. *Nucleic Acids Research*, *20*(20), 5357–5364. https://doi.org/10.1093/nar/20.20.5357

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews. Genetics*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3

Seiler, M., Peng, S., Agrawal, A. A., Palacino, J., Teng, T., Zhu, P., Smith, P. G., Cancer Genome Atlas Research Network, Buonamici, S., & Yu, L. (2018). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Reports*, *23*(1), 282-296.e4. https://doi.org/10.1016/j.celrep.2018.01.088

Shiozawa, Y., Malcovati, L., Gallì, A., Sato-Otsubo, A., Kataoka, K., Sato, Y., Watatani, Y., Suzuki, H., Yoshizato, T., Yoshida, K., Sanada, M., Makishima, H., Shiraishi, Y., Chiba, K., Hellström-Lindberg, E., Miyano, S., Ogawa, S., & Cazzola, M. (2018). Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nature Communications*, *9*(1), 3649. https://doi.org/10.1038/s41467-018-06063-x

Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V., & Ule, J. (2015). Recursive splicing in long vertebrate genes. *Nature*, *521*(7552), 371–375. https://doi.org/10.1038/nature14466

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., & Skotheim, R. I. (2016). Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*, *35*(19), 2413–2427. https://doi.org/10.1038/onc.2015.318

Takao, S., Forbes, L., Uni, M., Cheng, S., Pineda, J. M. B., Tarumoto, Y., Cifani, P., Minuesa, G., Chen, C., Kharas, M. G., Bradley, R. K., Vakoc, C. R., Koche, R. P., & Kentsis, A. (2021). Convergent organization of aberrant MYB complex controls oncogenic gene expression in acute myeloid leukemia. *ELife*, *10*. https://doi.org/10.7554/eLife.65905

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-15171-6

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Tyner, J. W., Tognon, C. E., Bottomly, D., Wilmot, B., Kurtz, S. E., Savage, S. L., Long, N., Schultz, A. R., Traer, E., Abel, M., Agarwal, A., Blucher, A., Borate, U., Bryant, J., Burke, R., Carlos, A., Carpenter, R., Carroll, J., Chang, B. H., … Druker, B. J. (2018). Functional genomic landscape of acute myeloid leukaemia. *Nature*, *562*(7728), 526–531. https://doi.org/10.1038/s41586-018-0623-z

Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., Zhang, L., Zhang, W., Vartanov, A. R., Fernandes, S. M., Goldstein, N. R., Folco, E. G., Cibulskis, K., Tesar, B., Sievers, Q. L., Shefler, E., … Wu, C. J. (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England Journal of Medicine*, *365*(26), 2497–2506. https://doi.org/10.1056/NEJMoa1109016

Wesselhoeft, R. A., Kowalski, P. S., & Anderson, D. G. (2018). Engineering circular RNA for potent and stable translation in eukaryotic cells. *Nature Communications*, *9*(1), 2629. https://doi.org/10.1038/s41467-018-05096-6

Wesselhoeft, R. A., Kowalski, P. S., Parker-Hale, F. C., Huang, Y., Bisaria, N., & Anderson, D. G. (2019). RNA Circularization Diminishes Immunogenicity and Can Extend Translation Duration In Vivo. *Molecular Cell*, *74*(3), 508-520.e4. https://doi.org/10.1016/j.molcel.2019.02.015

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., Chalkidis, G., Suzuki, Y., Shiosaka, M., Kawahata, R., Yamaguchi, T., Otsu, M., Obara, N., Sakata-Yanagimoto, M., Ishiyama, K., … Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, *478*(7367), 64–69. https://doi.org/10.1038/nature10496

Zhang, J., Ali, A. M., Lieu, Y. K., Liu, Z., Gao, J., Rabadan, R., Raza, A., Mukherjee, S., & Manley, J. L. (2019). Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Molecular Cell*, *76*(1), 82-95.e7. https://doi.org/10.1016/j.molcel.2019.07.017

Zhang, K., Nowak, I., Rushlow, D., Gallie, B. L., & Lohmann, D. R. (2008). Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression. *Human Mutation*, *29*(4), 475–484. https://doi.org/10.1002/humu.20664

Zhang, P., Philippot, Q., Ren, W., Lei, W.-T., Li, J., Stenson, P. D., Palacín, P. S., Colobran, R., Boisson, B., Zhang, S.-Y., Puel, A., Pan-Hammarström, Q., Zhang, Q., Cooper, D. N., Abel, L., & Casanova, J.-L. (2022). Genome-wide detection of human variants that disrupt intronic branchpoints. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), e2211194119. https://doi.org/10.1073/pnas.2211194119

VITA

Jose Mario Bello Pineda was born and raised in Baguio City in the Cordillera Administrative Region of the Republic of the Philippines. He attended the Saint Louis School Center for elementary school and the Baguio City National High School. From 2010 to 2015, he worked as a research assistant and technician in the lab of Dr. Wenying Shou at the Fred Hutchinson Cancer Center. Here, he studied the evolution of a synthetic, cross-feeding *Saccharomyces cerevisiae* mutualism using mathematical modeling and simulations, and experimental methods in molecular biology and genetics. He received his bachelor's degrees in Mathematics and Neurobiology at the University of Washington in 2015.