UNIVERSITY OF CALIFORNIA
Los Angeles

Big data discovery of cancer immunotherapy targets arising from alternative splicing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics

by

Yang Pan

2020

ABSTRACT OF THE DISSERTATION

Big data discovery of cancer immunotherapy targets arising from alternative splicing

by

Yang Pan

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2020

Professor Yi Xing, Chair

Alternative pre-mRNA splicing (AS) is a prevalent mechanism and a main source of transcriptomic and proteomic complexity in cells. Dysregulation of AS is widespread in tumor transcriptomes. Cancer immunotherapies have transformed the treatment of aggressive tumors, but identification of novel tumor antigens remains challenging. Petabytes of sequencing data in public domains presents unprecedented opportunities to exploit AS-derived peptides as a new category in the tumor antigen repertoire. In this dissertation, novel computational methods were developed to detect AS variations in cancers with significant biological or therapeutic implications. Utilizing these new tools, we demonstrated that we can characterize the key AS changes responding to oncogenic signals alterations, and more importantly, systematically identify splicing events that are potential tumor antigens for targeted immunotherapies.

The first part of the dissertation describes Pathway Enrichment-Guided Activity Study of Alternative Splicing (PEGASAS), a novel computational framework identifying key splicing changes associated with oncogenic signals during disease progression from large-scale RNA-seq data. Although aberrant AS are widely detected in cancer, causes and consequences of AS dysregulations during cancer progression remain elusive. PEGASAS uses a pathway-guided approach for examining the effects of oncogenic signaling on splicing. Applying it to study a comprehensive prostate cancer dataset, we identified a conserved set of AS events regulated by oncogenic pathways and establish a role for Myc in regulating RNA processing. PEGASAS provides a generic framework to connect AS changes with a wide range of oncogenic alterations in cancers.

The second part of the dissertation presents Isoform peptides from RNA splicing for Immunotherapy target Screening (IRIS), a big data computational platform that integrates massive transcriptomic data along with proteomics data to characterize AS-derived tumor antigens for cancer immunotherapy. Exiting frameworks of tumor antigen discovery are predominantly somatic mutation-based, leaving AS-derived targets largely unexploited. IRIS employs a comprehensive reference panel that determines tumor AS events by leveraging splicing patterns from tens of thousands normal and tumor transcriptomes. Applying IRIS to analyze RNA-Seq data from 22 glioblastomas from patients, we identified candidate epitopes and validated their recognition by patient T cells. This work demonstrates IRIS's utility for expanding targeted cancer immunotherapy by enabling big data-informed discoveries of a variety of AS-derived tumor antigens.

The dissertation of Yang Pan is approved.

Douglas L. Black

Antoni Ribas

Owen N. Witte

Yi Xing, Committee Chair

University of California, Los Angeles

2020

To my family and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to express my profound gratitude to my advisor Prof. Yi Xing, for his great support of my PhD study and related research, and for his patience, dedication, and immense knowledge. I learned tremendously from him from all the aspects of doing research. He continually contributes valuable feedback, advice, and encouragement during my PhD training in his lab for the past of five years since I joined the lab in 2015. His cultivation to me has been extremely invaluable to my future research career.

Besides my advisor, I would like to thank my committee members, Dr. Douglas L. Black, Dr. Antoni Ribas, and Dr. Owen N. Witte, for their insightful advice and generous support. They have been instrumental throughout my PhD training.

I greatly appreciate the support received from all the collaborators and co-authors through various collaborations. Their help to my research is indispensable and I cannot be more thankful:

For the published prostate cancer study, it’s a joint work with Dr. Owen Witte’s group from UCLA. My thanks go to the lead collaborator Dr. John Phillips and all other authors contributing to the paper: Brandon L. Tsai, Zhijie Xie, Levon Demirdjian, Wen Xiao, Harry T. Yang, Yida Zhang, Chia Ho Lin, Donghui Cheng, Qiang Hu, Song Liu, Douglas L. Black, Owen Witte, Yi Xing.

For the IRIS study, the preprint of the work is currently online. This is a collaborative work with Dr. Robert Prins’s group from UCLA. My thanks go to the lead collaborators Alexander Lee and Harry T. Yang and all other authors contributing to the paper: Yuanyuan Wang, Yang Xu, Kathryn E. Kadash-Edmondson, John Phillips, Ameya Champhekar, Cristina Puig, Antoni Ribas, Owen N. Witte, Yi Xing. In addition, I want to thank Beatrice Zhang for her help with data processing during the manuscript revision.

For the ongoing writing of a review article in chapter 1, I would like to thank Kate Kadash-Edmondson for her tremendous help with the manuscript.

I am also very grateful to current and former lab members in Xing lab: Jinkai Wang,

Zhixiang Lu, Shihao Shen, Shaofang Li, Shayna Stein, Yida Zhang, Chengyang Wang, Zijun Zhang, Yang Guo, Zhicheng Pan, Yuanyuan Wang, Yongbo Wang, Samir Adhikari, Eddie Park, Ruijiao Xin, Yan Gao, Amal Katrib, Emad Bahrami-Samani, Xinyuan Chen, Harry Yang, Yang Xu, Levon Demirdjian, Wankun Deng, Yuan Gao, Feng Wang, Yungang Xu, Kate Kadash-Edmondson, Beatrice Zhang, Siwei Luo, Robert Wang and Emerson Hunter.

Last but not least, I would like to thank my grandparents, my parents and my wife Xiao Sun for their support and unconditional love. I love them deeply every moment.

VITA

EDUCATION

University of Pennsylvania / Children's Hospital of Philadelphia, Philadelphia, Pennsylvania.
Visiting Ph.D. student, Bioinformatics Aug 2018 - Jun 2020 (expected)

University of California, Los Angeles, Los Angeles, California.
Ph.D., Bioinformatics Jun 2020 (expected)

The George Washington University, Washington D.C.
M.S., Bioinformatics May 2014

Jilin University, Changchun, Jilin, China.
B.S., Biological Science Jul 2012

HONORS & AWARDS

Emerging Innovators Jan 2020
Emerging Innovators in Collaborative Research, Children's Hospital of Philadelphia

Bursary Aug 2018
Proteomics Bioinformatics, Wellcome Genome Campus Advanced Courses & Scientific Conferences

Graduate Assistantship & Tuition Award Apr 2013
Teaching Assistant Fellowship, the George Washington University

Oak Ridge Associated Institute Science & Education (ORISE) Fellowship May 2013
Research Participation Program, Center for Biologics Evaluation and Research (CBER / U.S. FDA)

SELECTED PUBLICATIONS

Yang Pan†, Alexander H. Lee†, Harry Yang†, Yuanyuan Wang, Yang Xu, Kathryn E. Kadash-Edmondson, John Phillips, Ameya Champhekar, Cristina Puig, Antoni Ribas, Owen N. Witte, Robert M. Prins*, Yi Xing*. *IRIS: Big data-informed discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing.* In preparation.
GitHub: github.com/Xinglab/IRIS; Preprint: www.biorxiv.org/content/10.1101/843268v1

John Phillips†, **Yang Pan†,** Brandon Tsai, Zhijie Xie, Levon Demirdjian, Wen Xiao, Harry Yang, Yida Zhang, Chia Ho Lin, Donghui Cheng, Qiang Hu, Song Liu, Douglas Black, Owen Witte*, Yi Xing*. *Pathway-guided analysis reveals Myc-dependent alternative pre-*

mRNA splicing in aggressive prostate cancers. PNAS. Feb 2020. PMID: 32086391 GitHub: github.com/Xinglab/PEGASAS; Web app: xingshiny.research.chop.edu:3838/PEGASASServer

Levon Demirdjian, Yungang Xu, **Yang Pan**, Shayna Stein, Zhijie Xie, Eddie Park, Ying Nian Wu, Yi Xing. *Detecting allele-specific alternative splicing from population-scale RNA-seq data*. American Journal of Human Genetics (AJHG). Jan 2020. In revision.

Yida Zhang, Harry Yang, Kate Kadash-Edmondson, **Yang Pan**, Zhicheng Pan, Davidson Black, Yi Xing*. Regional variation of splicing QTLs in human brain*. American Journal of Human Genetics (AJHG). Oct 2019. In revision.

Yu Fan, Yu Hu, Cheng Yan, Radoslav Goldman, **Yang Pan**, Raja Mazumder, Hayley M. Dingerdissen*. Loss and Gain of N-linked Glycosylation Sequons due to Single-nucleotide Variation in Cancer.* Scientific reports. Mar 2018. PMID: 29531238

Yang Pan†, Cheng Yan†, Yu Hu, Yu Fan, Qing Pan, Quan Wan, John Torcivia-Rodriguez, Raja Mazumder*. *Distribution bias analysis of germline and somatic single-nucleotide variations that impact protein functional site and neighboring amino acids.* Scientific Reports. Jan 2017. PMID: 28176830

Jinkai Wang, **Yang Pan**, Shihao Shen, Lan Lin, and Yi Xing. *rMATS-DVR: rMATS discovery of Differential Variants in RNA.* Bioinformatics (Oxford). Jan 2016. PMID: 28334241

Amirhossein Shamsaddini†, **Yang Pan†**, W. Evan Johnson, Konstantinos Krampis, Mariya Shcheglovitova, Vahan Simonyan, Amy Zanne, Raja Mazumder*. *Census-based rapid and accurate metagenome taxonomic profiling*. BMC Genomics. Oct 2014. PMID: 25336203

Yang Pan†, Konstantinos Karagiannis†, Haichen Zhang, Hayley Dingerdissen, Amirhossein Shamsaddini, Quan Wan, Vahan Simonyan, Raja Mazumder*. *Human germline and pan-cancer variomes and their distinct functional profiles.* Nucleic Acids Research. Sep 2014. PMID: 25232094

Yu Wang†, **Yang Pan†**, Zeqiang Zhang, Ruoxi Sun, Dahai Yu*, Xuexun Fang*. *Combination use of ultrasound irradiation and ionic liquid in enzymatic isomerization of glucose to fructose*. Process Biochemistry. Jun 2012

Chapter 1 Introduction - RNA Dysregulation-Derived Antigen Targets for Cancer Immunotherapy

1.1 Exploiting RNA Processing Dysfunctions as Targets in Cancer Immunotherapy

Variations in the human transcriptome provide enormous diversity at the transcript and protein levels. RNA-level variations are generated from a wide range of tightly regulated processes, such as alternative pre-mRNA splicing and RNA editing, among others. Aberrant RNA processing is a major cause or contributor to many human diseases. Indeed, several large-scale cancer studies have reported elevated levels of recurrent RNA-processing dysregulations, many of which are key or driver alterations for cancer.

Cancer immunotherapy has gained momentum in the clinic because of its success in treating various types of aggressive malignancy. A critical aspect in developing immunotherapies with long-lasting antitumor immunity is discovering targetable tumor antigens (TAs). Although shown to be effective in many clinical studies, current TA discovery approaches search a very limited fraction of tumor variations by only focusing on genome-level alterations, such as single-nucleotide variations (SNVs), insertion-deletion mutations (indels), and, occasionally, genomic fusion events. This approach cannot be feasibly applied for designing targeted immunotherapy for patients with moderate or low numbers of somatic mutations.

RNA variations in the tumor transcriptome can give rise to TAs, generating a potential new repertoire of emerging targets for cancer immunotherapy. Here, we review the major types of RNA dysregulation that shape the tumor transcriptome landscape and, in turn, transform the tumor-cell immunopeptidome and surfaceome. We explore how these RNA dysregulation-derived alterations can be targeted for immunotherapy, highlighting the tools and resources required for their discovery. Finally, we discuss exciting works and new strategies that can be applied to facilitate the discovery of novel therapeutics.

1.2 Cancer Immunotherapy and Tumor Antigens

Cancer immunotherapy revolutionized the cancer treatment paradigm

By harnessing and augmenting the patient's antitumor immunity, cancer immunotherapy shifted the paradigm of treating human malignancies. Various cancer types, including aggressive cancers and those previously considered untreatable, have been effectively treated with immunotherapy, leading to improved survival, durable responses, and other benefits (1-3).

One major type of immunotherapy works by augmenting the suppressed immune response in the tumor environment. This strategy, immune checkpoint blockade (ICB), includes a class of therapies that use immune checkpoint inhibitors, such as neutralizing antibodies against PD-1 or CTLA-4, to reactivate tumor-specific T cells (4). Another class of therapies, including therapeutic antibodies and adoptive cell therapies (ACTs) (5), work by directing or engineering immune cells to improve the anticancer response. For therapeutic antibodies (e.g., anti-CD20 rituximab), the therapy binds directly to the TA to direct the immune response (6). By contrast, in ACTs, the patient's own immune cells are removed and modified ex-vivo to enhance their antitumor ability. Successful cases of ACTs include using tumor-infiltrating lymphocyte (TIL) therapy to treat metastatic melanoma (7,8) and therapeutic cancer vaccines for metastatic prostate cancer (9). With engineered ACTs, which include T-cell receptor (TCR) (10) and chimeric antigen receptor T-cell (CAR-T) (11) therapies, the patient's T cells are engineered ex-vivo for improved reactivity to known TAs. Use of engineered ACTs has led to promising clinical outcomes, including landmark studies in progressive metastatic melanoma (12,13) and an FDA-approved CD19 CAR-T cell therapy for B-cell malignancies, such as large B-cell lymphoma (14) and leukemia (15).

TAs and anticancer immunity are keys to immunotherapy success

The choice of TA is an essential consideration for generating strong anticancer immunity for successful immunotherapy (16,17). TAs are formed from peptides generated from genetic, transcriptional, or translational alterations in the tumor, and ideally should be rarely or not expressed on normal cells. Both targeted immunotherapies and ICB therapies require that TAs be recognizable as foreign antigens to T cells (17-19). Specifically, for CAR-T and TCR therapies using engineered ACTs, TAs are recognizable through distinct pathways – either by extracellular domain expression or by presentation via major histocompatibility complex (MHC) molecules on the cell surface.

A key feature of TAs for immunotherapy is its selectivity of expression or association with tumors. Based on this feature, TAs can be grouped into three major classes (17). Tumor-specific antigens (TSAs) are exclusively expressed by tumor cells. Among TAs, TSAs have the highest association with tumors. TSAs are often referred to as neoantigens (2,19,20). Although difficult to identify, TSAs are ideal targets for immunotherapy because they offer the potential for effective, low-toxic targeting due to their highly specific tumor expression. Tumor-associated antigens (TAAs) are a common class of TAs that are overexpressed in malignant cells but also expressed (to a limited extent) in normal cells. TAAs can be useful targets because their presence can be generic across patients and even tumor types. One important TAA is ERBB2 (HER2/NEU). Expressed in normal adult tissues, ERBB2 is overexpressed in many epithelial tumors, including breast tumors (21). However, immunogenicity (22,23) and potential toxicity are major concerns for targeting TAAs (16,24). Lastly, cancer/testis antigens (CTAs) comprise a special class of TAs that are present at elevated levels in tumors and reproductive tissues, while showing limited expression in normal adult tissues. CTAs have been developed as therapeutic targets for tumors, by taking advantage of the fact that normal reproductive cells do not express MHC class I molecules (25).

A key step for developing targeted immunotherapies is efficiently identifying targetable TAs. Advancements in genomic sequencing have allowed emergence of a consensus TA discovery framework, which is largely based on genome (mostly exome) sequencing (17,18,26,27). Multiple successful applications of this consensus discovery framework have been described (28-31). Genomic sequencing-based strategies primarily focus on TAs derived from SNVs, although TAs from indels (32), CNVs, and gene fusions (33) are also common.

These strategies identify somatic variation-based antigens that are often highly tumor-specific and technically easy and robust to detect. However, one major problem is that somatic variation changes are specific to a very small subset of patients; therefore, therapy needs to be determined on an individual (or “personalized”) basis (17). Furthermore, only a small proportion of somatic changes can alter the protein-coding sequence, preventing identification of TAs from tumors with moderate or low numbers of genetic mutations (18). In addition, many SNV-based antigens suffer from poor immunogenicity because the sequence is changed by only a few amino acids (34).

Given the limitations of targeting TAs derived from somatic variations alone, new TA sources are required (35). Tumor alterations are not limited to genomic mutations. Thus, large numbers of transcriptomic variations in tumor are overlooked by using genomic sequencing-based methods, presenting huge opportunities for discovering novel targets.

1.3 RNA Dysregulations in the Tumor Transcriptome

RNA processing diversifies the human transcriptome

Under tight cis- and trans-level regulatory control, RNA processing events generate diverse transcript and protein isoforms that are required for essential biological functions in humans (36). The transcriptome is governed by cis regulation (e.g., DNA mutations in a nearby genomic region) and trans regulation (e.g., binding of proteins with regulatory roles). Dysfunction of these RNA-related biological processes will result in abnormal cell functions, which can be a major source of and contributor to human diseases. In cancer, many RNA-related processes are often dysregulated. These dysregulations can reshape the landscape of the tumor transcriptome by changing the abundance and diversity of RNAs or transcripts.

In this section, we summarize current knowledge on the mechanisms and consequences of dysregulation of RNA-related processes, noting that there is frequently overlap among processes in their mechanisms/consequences of dysfunction. In particular, we review the effects of dysfunction on the tumor transcriptome, including how dysregulation can generate novel transcripts or RNA species (aberrantly expressed RNA) or alter the relative abundance of RNAs compared to the normal condition (differentially expressed RNA).

Alterations in RNA expression

Alterations in RNA abundance between two biological conditions are widely studied and commonly seen in human transcriptomes under both normal and disease conditions (37-41). RNAs that are overexpressed in cancer but expressed at a baseline level in normal cells may play essential oncogenic roles and are a major source of variation in the tumor transcriptome (42). A common example in cancer is the overexpression of RNA for transcription factor-encoding oncogenes in tumors (43).

RNA expression changes in tumors derive from cis regulatory effects caused by genetic differences (e.g., genetic polymorphisms between different individuals) or trans regulatory effects caused by differentially expressed upstream regulatory proteins (e.g., tissue-specific RNA expression) (37-39). Additionally, shifts in RNA expression can be caused by other types of transcriptional regulations. For instance, changes in RNA abundance can result from alterations in regulatory RNAs (44), such as microRNAs (miRNA) (45), piwi-interacting RNAs (piRNAs) (46,47), and long noncoding RNAs (lncRNAs) (48). Changes in RNA abundance might also be caused by alterations in RNA splicing or other RNA-processing events (49). Thus, the dysregulation of RNA expression in cancer is the combined result of many alterations (36). Due to the relevance to protein product, RNA expression changes have important implications for the tumor phenotype. In detailing the types of RNA dysregulation that are found in cancer, we note that many of these dysfunctions can alter RNA expression in the tumor.

Pre-mRNA alternative splicing

As the major source of RNA and protein product diversity in human and other eukaryotic cells (36,50), alternative pre-mRNA splicing has fundamental physiological functions in tissue identity and development (50). There are five common forms of alternative splicing events: intron retention, exon skipping, mutually exclusive exons, and alternative 3′ and alternative 5′ splice site changes (51). More sophisticated forms are also detected in human cells (52). As with other RNA processing events, alternative splicing is tightly regulated via genetic (cis) and RNA/protein (trans) effects (53).

Alternative splicing dysregulation contributes to every “hallmark of cancer” (54,55). Initially identified through genomic studies, the hallmarks of cancer are the key phenotypic and energetic characteristics of tumors compared to normal cells (56,57). Subsequent RNA sequencing (RNA-seq) reports found that every hallmark of cancer involves splicing dysfunctions (55). Aberrant alternative splicing can contribute to cancer development in various ways, such as by generating isoforms that promote tumorigenesis or metastasis, by inhibiting apoptosis, promoting growth signaling, or enabling epithelial-to-mesenchymal transition (58).

Large-scale RNA-seq analyses have revealed shifts in the relative abundances of alternatively spliced isoforms across tumor types, with functional implications for cancer (59). Comparing tumor samples to tumor-matched normal tissues, researchers found increased numbers of novel splicing isoforms in tumor, with enrichment in unannotated skipped exon and alternative 3' splice site events (60). Several studies reported dysregulation of intron retention as a common pathway in cancer (61,62). Indeed, certain splicing dysregulations have been shown to promote oncogenesis or cancer development, and may be used to predict cancer prognosis (55,63).

Noncanonical splicing

Dysregulated noncanonical splicing can form aberrant RNAs (64). Trans-splicing and cis-splicing between adjacent genes (cis-SAGe) are rare RNA-processing events that can produce novel transcripts in the absence of DNA-level alterations (65-67). Generated transcripts are often called "chimeric RNAs" or "fusion transcripts" because they are comprised of exons from different genes (68). Although many chimeric RNAs have been identified as biomarkers or targets for cancers (65,69), trans-splicing and cis-SAGes have been found in embryonic stem cells (70) and other noncancer tissues and cells (71).

Circular RNAs (circRNAs) are another cancer-implicated RNA product resulting from noncanonical splicing. These closed-loop RNAs result from the back-splicing of exons within pre-mRNA (72). Although circRNAs exhibit certain biological functions in normal cells (72), they are expressed at higher levels and show different patterns of isoform expression in various tumor types (73). At present, much remains unknown about the pathomechanisms underlying the increased expression of specific circRNA isoforms in cancer(74). Researchers have begun to create databases of cancer-associated circRNAs (e.g., MiOncoCirc) with the hope that these isoforms can be used as diagnostic or therapeutic targets (73).

Despite the intriguing implications for cancer identification and treatment, the current understanding of noncanonical splicing dysregulation is still limited. Thus, systematic evaluations of the patterns and roles of noncanonical splicing in the tumor transcriptome are needed.

RNA editing

RNA editing, an important mechanism whereby mRNA undergoes site-specific nucleotide modifications (75,76), has important regulatory roles in protein recoding, RNA activity (77), and RNA secondary structure (78). The most common type of RNA editing is adenosine (A)-to-inosine (I) editing, in which adenosine deaminase acting on RNA (ADAR) enzymes (79) catalyze conversion of A to I at discrete sites on mRNA. A-to-I RNA editing is functionally important in many human tissues, especially brain (77,80). Its dysregulation contributes to many human diseases, including cancer (81-83). For example, large-scale RNA-seq analyses comparing tumor to normal tissues revealed distinct RNA-editing patterns. Other studies have associated dysregulated RNA-editing activities with patient survival (83) and cancer development (84,85).

Retrotransposons

Retrotransposons, or class I transposable elements (TEs), are portions of the genome that undergo transposition by converting RNA back into DNA via an RNA transposition intermediate. Examples of retrotransposons in humans include long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons (e.g., Alu elements), and endogenous retroviruses (ERVs). In addition to their physiological functions and their importance in increasing transcript diversity, retrotransposons also play roles in human diseases such as cancer (86-88).

Alu elements are among the most active retrotransposons in the human genome (89). As a family of short, primate-specific elements, Alu sequences are expressed in mature mRNA through a splicing-mediated process called exonization (90,91). Specifically, a genetically inserted Alu element can introduce novel splice sites for splicing machinery to recognize, which could lead to creation of a new exon. According to reports, Alu exonization is tissue-specific and can play regulatory roles (92,93). Recent studies have characterized thousands of retrotransposon-generated novel splicing sites in cancer genes (94). Genomic regions within Alu elements are enriched with splice site-creating somatic mutations across cancers (95), suggesting that cancer takes advantage of this mechanism during cancer genome evolution.

Other posttranscriptional RNA-level events

Dysfunctions in other steps of RNA posttranscriptional regulation have been studied in cancer, including dysfunctions in alternative polyadenylation (96) and RNA modifications such as m6a RNA methylation (97). These events typically result in changes to the RNA translational efficiency (96), RNA localization (98), or RNA stability (99) without creating novel transcripts or increasing the isoform diversity. Therefore, most likely these dysfunctions are not observed to alter the observed RNA expression profile.

1.4 Targeting RNA Dysregulation-Generated TAs for Immunotherapy

Many dysregulated RNA processes have essential functions in cancer development. For example, aberrant products resulting from dysfunctional RNA processes may be translated into proteins carrying unique tumor-related signatures. Peptides derived from these proteins could be potential TAs if they are located on the protein extracellular domain or are able to be presented by MHC molecules on the cell surface.

In this section, we discuss possible sources of TAs derived from the dysregulation of various RNA-level processes. Some sources for TA candidates have been studied computationally or by preliminary experimental works. Other promising TA sources derived from dysfunctional RNA-level processes have not yet been described in published works.

RNA overexpression-derived TAAs

In cancer, genes or transcript isoforms that are overexpressed can generate proteins that are enriched in tumor cells. If the protein is stably overexpressed across patients with the tumor of interest, or even across tumor types, and if the protein expression is sparse or limited in normal tissues, then this protein may be a candidate for a TAA. Clinical studies have been performed with several TAAs that show overexpression in tumors or specific expression in certain cell lineages (e.g. CD19-specific CAR, HER2/neu-specific CAR) (100-102). RNA overexpression-derived TAAs are commonly used as generic (nonspecific) targets in immunotherapy. However, issues of limited immunogenicity and potential for toxicity remain to be solved (102,103).

Alternative splicing-derived TAs

Owing to its prevalence and protein diversification function, alternative splicing is a key source of TAs through shifting splicing patterns across different cancer types. TAs can result from any pattern of alternative splicing, whether basic (e.g., skipped exon, intron retention, etc.) or complex. Depending on whether the skipping or inclusion form is identified as the isoform in tumor, the corresponding skipping or inclusion splice junction(s) is used to generate the splice-junction peptides as TA candidates. For example, if inclusion of an exon (or intron, in the case of intron retention) is enriched in tumors, then the entire exon/intron body can be translated to peptides as TA candidates.

There are two scenarios of splicing in cancer, leading to TAs with different tumor specificity. Differentially spliced isoforms in tumors, if translated, can be potential TAAs due to their expression in normal cells. Novel spliced isoforms that are specifically expressed in tumors can be TSA candidates. Possible sources of novel isoforms could be splice-junction region-derived peptides coming from a novel exon-exon combination or from completely novel splice sites.

Early attempts at utilizing alternative splicing-derived TAs have been effective in treating lymphoma and ovarian cancers (104,105), but focused on very specific targets. A recent comprehensive re-analysis of RNA-seq data from The Cancer Genome Atlas (TCGA) evaluated putative splice junction-derived TAs from cancer-specific splicing events by comparing tumor samples to normal tissues from the GTEx database (60). Researchers only included possible TSAs from the five simple alternative splicing types, and only considered splice-junction peptides as TA candidates. They cross-validated a fraction of these peptides using mass-spectrometry (MS) data for protein expression. Around the same time, Smart et al. reported a more focused analysis to identify cancer-specific intron retention events in patent samples. They performed additional experimental validation for MHC presentation of intron retention-derived epitopes (not tumor-specific) (106), indicating potential immunogenicity. They used the sequence of the entire intron body to construct peptides. Tumor-specific expression of peptides was determined by comparing to a small cohort of manually selected normal samples using a database of known proteins. However, these studies have some limitations. For example, both use heuristic approaches in defining the tumor specificity of antigens due to the lack of a standardized reference of normal splicing pattern (107). They lack experimental evidence for the immunogenicity of alternative splicing-derived TSAs, and limit their scope to only TCR targets. Nevertheless, despite limitations, these pioneering studies provide evidence that alternative splicing is a promising source for TAs that can be used as targets for immunotherapy.

Chimeric RNA- or circRNA-derived TAs

Rare splicing events are found at increased levels in cancer compared to normal tissues (69). Theoretically, if a novel chimeric RNA generated by trans-splicing can be translated in the tumor cell, then the trans-spliced junction peptide could be used as a candidate TSA. If a new open-reading frame (ORF) is introduced to the downstream exon in the tumor-specific chimeric RNA, then the entire translated peptide sequence from the downstream exon before the stop codon could be considered as a TSA source. However, one major concern is the quality of chimeric RNAs that are detected from RNA-seq, as a considerable proportion might result from sequencing artifacts. The tumor specificity of chimeric RNAs also must be thoroughly evaluated due to their expression in normal tissues (71,108). Although a few reports have targeted chimeric RNAs (109), to date, no study has performed a focused analysis of TA-derived chimeric RNAs. Thus, there is a need for additional systematic research, building on existing fusion transcript and fusion gene detection tools (33,110,111), to explore the expression landscape of chimeric RNAs in cancer.

Similarly, circRNAs have been shown to be translated in cancer and normal tissues (112,113). With minor adaption from the canonical splicing antigen framework, peptides formed from these back-splice junctions could be a possible source of TAs to explore. Although some regulatory functions of circRNAs in cancer have been uncovered, there have been very limited studies looking at translated peptides from circRNAs or the potential of circRNAs to serve as TAs (114). To evaluate the level of tumor association, comprehensive analysis is needed, given the functional roles of circRNAs in normal cells.

Edited RNA transcript-derived TAs

The RNA-editing process can introduce protein variations (115) and, hypothetically, peptides derived from cancer-associated RNA-edited transcripts can be a source of TAs. Although many studies have shown that transcripts are differentially edited in tumor cells, edited transcripts and proteins are also expressed in normal cells (85). Therefore, edited peptides can be either TAAs or TSAs. Owing to the site-specific nature of RNA editing, the resulting sequence change may be insufficient for desired immunogenicity. Using immunopeptidomic data and a T-cell–mediated cell-killing assay, a recent study demonstrated the first experimental evidence that RNA-edited peptides can be presented by MHC molecules and can elicit tumor responses (116). Notably, the study showed that RNA-derived TAAs are recognized by T cells that are physiologically present in cancer tissue, alleviating concerns regarding safety and toxicity of targeting TAAs. Despite the promising therapeutic implications of their findings, the authors acknowledged the need for a careful evaluation of toxicity. These reports highlight RNA editing as an emerging alternative source of TAs in cancer immunotherapy. Future studies should be aimed at the large-scale characterization and rigorous immunological validation of TAs derived from RNA-edited transcripts.

Retrotransposon-derived TAs

There is now extensive evidence that retrotransposons, especially exonized Alu elements, are translated into proteins in normal and cancer cells in a tissue-specific fashion (117). Thus, this process may represent a promising source of TAs (35). Splice site-creating somatic mutations are significantly enriched in Alu regions in tumors, suggesting the functional importance of these novel transcripts (95). After translation, newly exonized sequences can generate peptides that have never been exposed to the host's immune system, offering the potential for strong immunogenicity. As very few studies have focused on identifying retrotransposon-derived exons (117), the scale and landscape of retrotransposon-derived TAs are unknown. Using RNA-seq data and immunopeptidomic MS analysis, a recent study of TSAs found evidence of aberrantly expressed but nonmutated transcripts from endogenous retroelements (118). The authors hypothesized that these TSAs could be shared across multiple tumor types.

Although retrotransposon-derived TAs offer the advantage of immunogenicity due to their foreignness, validation of their translation by tumors remains a difficult task. Full-length sequences of retrotransposon-derived antigens are difficult to obtain by traditional short-read sequencing, and no tools are currently available to identify retroelement TSAs (35). Further technological advances, such as high-throughput third-generation sequencing, are needed to improve detection of TAs derived from retrotransposons in tumor cells.

1.5 Multi-omic and Big-data Strategies to Discover RNA-derived TAs

Any approach for discovering RNA dysregulation-derived TAs should be more detailed than existing frameworks for DNA-derived TAs and should include several key components. At minimum, an effective discovery strategy for RNA-derived TAs should include following features: 1) accurately characterize the sequences and abundances of tumor transcripts, 2) efficiently identify translated protein products, 3) robustly determine the tumor association and tumor-selective expression levels of targets with the tumor, and 4) evaluate the likelihood that identified tumor peptides can be targeted by TCR, CAR-T, or other targeted immunotherapies.

Nevertheless, several challenges exist in creating such a discovery strategy. For tumor transcripts that are unannotated or derived from complicated RNA-processing events, identifying the exact or complete RNA sequence, splice site, and ORF used for translation can be difficult. Furthermore, it can be difficult to ascertain the tumor selectivity and association of the antigen (i.e., degrees to which the TA is selectively expressed by and associated with the tumor), which have critical immunological and clinical implications for immunotherapy success. Thus, due to the complexity of tumor transcriptomes, discovering RNA dysfunction-derived TAs that can be targeted for immunotherapy will require an integrated strategy incorporating multi-omics experimental and computational solutions. Based on existing DNA-derived TA identification frameworks, in this section, we review and preview the tools in our toolbox that can be used to discover RNA-derived TAs. Using these integrated tools, we discuss how possible solutions may be achieved to the outstanding issues mentioned above.

Accurate RNA sequence characterization

Dysregulations in RNA processes can result in tumor-specific RNAs having new, previously undescribed sequences. Therefore, a key component of any RNA dysregulation-derived TA discovery strategy is the capability to accurately detect various types of RNA-derived transcripts.

RNA-seq (119). A powerful next-generation sequencing (NGS) technique, RNA-seq is used to sequence the entire set of RNA molecules, or transcriptome, which is isolated from cells. With over a decade of development, RNA-seq and its variations have become the most commonly used tool to profile human transcriptomes. RNA-seq has been used extensively in large-scale studies to investigate various biological conditions (60,120). While capable of capturing expression-level changes at the genome level, RNA-seq uniquely detects alterations that are only seen at the transcriptome level. For this reason, RNA-seq is the primary approach used to study the diverse tumor-associated transcripts that arise from dysregulated RNA processes (121). When settings such as proper read length and sequencing depth are chosen(121,122), RNA-seq can capture important transcriptome events, including RNA splicing, RNA editing, and other RNA-processing events.

Sophisticated computational and statistical algorithms have been developed to accurately handle RNA-seq data. Algorithms have been developed to align and quantify RNA-seq reads, as well as to identify and characterize various transcriptome-level features, including alternatively spliced isoforms(122,123), RNA-editing sites(124), chimeric RNAs(110), and circRNAs(125). Combining RNA-seq with existing analytical tools, researchers can readily detect many putative RNA-derived TAs, as demonstrated in pioneering works (60,61,116). Construction of dedicated and standardized tools would enable the systematic utilization of RNA dysregulation-derived TAs (32).

Despite the success of conventional RNA-seq technologies in cancer research, this NGS approach falls short in the inference of sophisticated or rare RNA transcripts (126), many of which have the potential to be strong TA candidates. Moreover, because RNA-seq is a short read-based sequencing technology, it only recovers a portion of the whole transcript. As a result, the ORF can remain unknown for unannotated or aberrantly expressed transcripts. This issue can lead to generation of an increased number of false targets, because all three ORFs must be used to generate putative peptide sequences for downstream antigen predictions.

Long-read sequencing (127,128). Third-generation, or long-read, sequencing technologies were designed to overcome inherent limitations of short-read sequencing. Whereas short-read technologies typically provide read lengths of up to 600 bases, long-read sequencers provide read lengths exceeding 10 kb (129). Thus, long-read sequencing has become an emerging tool for sequencing genomes and transcriptomes (130-133), resolving sequences for very long or complicated events that could not be determined by short-read methods. For example, full-length transcripts can be used to infer ORF information for in silico translation of peptide sequences, information that is critical for controlling the number of false targets in a TA candidate list.

Despite promising applications and rapid experimental (126) and computational(134) advances, long-read sequencing is still considered an underdeveloped technology. Limitations of current technologies including poor read accuracy, sometimes incomplete reads, and high cost with lower throughput than short-read sequencing (126). Further developments to decrease the error rate, increase throughput, and improve mapping accuracy would transform this emerging technology into a powerful tool for comprehensively identifying aberrant and unique transcripts in tumors.

Integrative detection of translation and protein expression

In cancer, dysregulations of different RNA processes can result in many novel transcripts. Evidence of translation and protein expression can greatly boost the validity that an RNA-level tumor event is an antigen. Transcriptome- or proteome-wide approaches may be applied to confirm translation of a transcript.

Ribo-seq. In this RNA-seq-based ribosome-profiling strategy (135,136), ribosome-protected mRNA fragments are captured and sequenced to infer whether active ribosomes during translation are present in cells. This strategy has been widely used to profile association of ribosomes with different RNA species and to quantify translated RNA isoforms, providing insights into many important translational processes (137,138). In particular, Ribo-seq has proven to be a powerful tool for identifying novel translated ORFs (139,140). Evidence of translation or translated ORFs can greatly reduce the need for computation and limit uncertainty during the *in-silico* search for TA candidates. For these reasons, applying Ribo-seq to characterize various aberrantly expressed RNA species in tumors may greatly improve detection of RNA-derived TAs.

MS-based proteomics. The primary approach to characterizing changes in translation at the proteome level is MS-based proteomics (141,142). Originating from MS techniques measuring the mass-to-charge ratio of ions, MS-based proteomics are used to characterize protein expression and modifications at a global scale in cells. Various sample preparation and labeling methods, combined with different MS instruments, have been described for analyzing protein expression, modifications, protein-protein interactions, and other features. More specifically, large-scale cancer studies have generated MS proteomics data to investigate direct protein changes and their roles in apoptosis and oncogenesis (143,144). These tumor-derived MS data with matched RNA-seq data are invaluable resources for exploring the TAs expressed in tumor cells (60), especially for putative candidates derived from novel tumor RNAs. The major limitation of this method is low sensitivity, and some major technological bottlenecks have not yet been addressed (145). Furthermore, the current capability of MS-based proteomics for detecting proteins and their variants is much lower than that of sequencing-based technologies at both the genome and transcriptome levels.

Spectrum-searching approaches can help improve the sensitivity and accuracy of peptide identification.

Proteogenomics is an emerging approach to improve the peptide-identification performance of MS-based proteomics (146). In proteogenomics, genomic and/or transcriptomic data from the same sample are used to augment the MS-based proteomic search (147). In contrast to NGS-based sequencing techniques, MS-based proteomics is an indirect approach that heavily relies on the completeness and accuracy of the MS search library (148). Customizing a standard library with sample-specific variations could dramatically increase detection. This data-driven approach can help avoid the inability to detect unannotated events, such as novel exons or mutated peptides, due to their absence from the library. RNA-seq data can be used to remove non expressed proteins from the MS library, avoiding the decreased detection power due to the inflated size of the MS library. Proteogenomics has been widely adopted by many large-scale studies to understand relationship between transcriptome and proteome (149), and has led to detection of numerous novel peptides in cancers (60,143,144). However, prospective works are needed to characterize RNA dysregulation-derived tumor peptides using this integrative approach. Explicitly integrating peptide sequences formed by dysregulated RNAs in cancer to guide the proteomic search may lead to novel discoveries.

Confirmation of antigen presentation for immunotherapy

Not all peptides from expressed proteins in cells can be accessible to T cells via TCRs or CARs. To be recognized, a peptide must be presented by MHC (HLA in human) or located on the cell surface. To confirm peptide presentation, specialized proteomic data are required.

Immunopeptidomic data. The immunopeptidome, or MHC/HLA peptidome/ligandome, refers to a collection of short peptides presented by MHC/HLA molecules on the cell surface (150). Immunopeptidomic profiling is an MS-based approach that differs from regular whole-call proteomics in that it considers intact peptides that are bound to antigen-presenting molecules (151). In this way, immunopeptidomic MS data offer a systematic perspective of the antigen landscape in cells and are considered evidence of antigen presentation. Immunopeptidomic profiling has gained popularity in immune-oncology as a tool to profile and validate TAs (106,118,152) and is often paired with NGS sequencing data in a proteogenomic workflow. However, as an MS-based approach, immunopeptidomic profiling suffers from low sensitivity. Nevertheless, immunopeptidomics holds promise for many applications in antigen discovery and validation for T cell-based immunotherapies.

Surfaceomic data. The surfaceome, or cell-surface proteome, refers to the collection of proteins that are expressed on the cell surface. Specialized MS-based proteomic protocols enable the identification and quantification of cell surfaceomes (153-155). As CAR-T therapies target extracellular antigens on tumor cells, using MS-based surfaceomics to profile expression of tumor-related cell-surface proteins can provide insight into their potential targetability as TAs for CAR-T therapies. In fact, this strategy has been used to prioritize TAs in advanced prostate cancers (156). Although limited by low sensitivity, surfaceomic data could be useful for confirming dysregulated RNA-derived CAR-T targets.

Big data-informed discovery

Another path to gaining insights into the landscape of RNA-derived TAs is through integration of big data. Massive datasets containing NGS and proteomic data from normal tissues and cancer samples with rich annotations have been accumulated in publicly available repositories, offering a major resource for discovering tumor-related RNA events for therapeutic targets. Here, we summarize several large-scale multi-omics datasets that are available for cancer researchers. As shown in Table 1.1, these big-data repositories offer data for thousands of normal and tumor samples, with sequencing by various groups and consortiums. The availability of RNA-seq data across conditions, complemented by other omics data, enables RNA dysregulation-derived TAs and their targetability by immunotherapy to be systematically and comprehensively characterized.

Evaluating tumor association of RNA-derived TAs. Tumor samples may harbor thousands of aberrantly expressed or differentially expressed RNAs. Thus, TA candidates must be robustly and efficiently prioritized. Similarly, to how a reference genome aids somatic mutation discovery, compiling big RNA-seq datasets across normal and tumor conditions could aid understanding of the selective expression of tumor-related RNA dysregulations. Knowing the expression pattern in normal samples could help to determine the tumor specificity and potential toxicity when targeting the antigen. Knowing the abnormal RNA expression pattern in tumor samples could help in evaluating how generalizable the target is among tumor patients. Future work is needed to compile a standardized and comprehensive catalog of normal and tumor RNA events, which would improve the robustness and reproducibility of RNA-derived tumor antigens. Emerging single-cell RNA-seq approaches could elevate the resolution of reference data by adding cell-type-specific and spatial information.

Despite the significant benefits and premises, big data also present many difficulties and analytical challenges. Firstly, the datasets do not equally represent different phenotypes and conditions. Certain tumor or normal tissues are underrepresented in datasets, posing issues for data analysis, integration, and interpretation. For example, no large-scale normal pediatric sample repository is currently available, which is a bottleneck to understanding changes in pediatric cancer. Secondly, when integrating biological big data, batch effects caused by various factors need to be addressed. Data from different sources need to be uniformly processed, harmonized, and evaluated for potential artifacts. Thirdly, standardized reproducible programs and best practices for integrating RNA-seq big data for TA discovery are lacking. Without such procedures, results from individual studies are not directly comparable, making big-data discovery unreliable. Thus, to complement the ever-growing list of tumor transcriptomic sequencing projects, advanced statistical models, efficient algorithms, and standardized computational pipelines need to be developed to address these challenges.

1.6 Tables

Table 1.1: Publicly available genomic and transcriptomic datasets for cancer research

Project name	Sample source	Phenotype	Data types	Sample size
GTEx	Tissue-derived & cell lines	Adult normal	WGS, WES, RNA-seq	>10,000 (from >600 individuals)
TCGA	Tissue-derived	Adult tumor & adjacent normal	WES, RNA-seq, etc.	>10,000
CPTAC	Tissue-derived	A fraction of TCGA samples	MS proteomics	>1,000 samples (matched to TCGA)
CCLE	Cell line	Adult tumor	WES, RNA-seq	~1,000 samples
ICGC	Tissue-derived	Adult tumor		
TARGET	Tissue-derived	Pediatric tumor	WGS, WES, RNA-seq, etc.	
St. Jude PCGP(157)	Tissue-derived	Pediatric tumor	WGS, WES, RNA-seq	~2,000 samples
HTAN(158)	Tissue-derived	Adult tumor	scRNA-seq, etc.	

1.7 References

1. 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. *et al.* (2010) Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med*, **363**, 711-723.

2. Brown, S.D., Warren, R.L., Gibb, E.A., Martin, S.D., Spinelli, J.J., Nelson, B.H. and Holt, R.A. (2014) Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*, **24**, 743-750.

3. Topalian, S.L., Hodi, F.S., Brahmer, J.R., Gettinger, S.N., Smith, D.C., McDermott, D.F., Powderly, J.D., Carvajal, R.D., Sosman, J.A., Atkins, M.B. *et al.* (2012) Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med*, **366**, 2443-2454.

4. Sun, C., Mezzadra, R. and Schumacher, T.N. (2018) Regulation and Function of the PD-L1 Checkpoint. *Immunity*, **48**, 434-452.

5. Hinrichs, C.S. and Rosenberg, S.A. (2014) Exploiting the curative potential of adoptive T-cell therapy for cancer. *Immunol Rev*, **257**, 56-71.

6. Tout, M., Casasnovas, O., Meignan, M., Lamy, T., Morschhauser, F., Salles, G., Gyan, E., Haioun, C., Mercier, M., Feugier, P. *et al.* (2017) Rituximab exposure is influenced by baseline metabolic tumor volume and predicts outcome of DLBCL patients: a Lymphoma Study Association report. *Blood*, **129**, 2616-2623.

7. Dudley, M.E., Wunderlich, J.R., Robbins, P.F., Yang, J.C., Hwu, P., Schwartzentruber, D.J., Topalian, S.L., Sherry, R., Restifo, N.P., Hubicki, A.M. *et al.* (2002) Cancer regression and autoimmunity in patients after clonal

repopulation with antitumor lymphocytes. *Science*, **298**, 850-854.

8. Rosenberg, S.A., Yang, J.C., Sherry, R.M., Kammula, U.S., Hughes, M.S., Phan, G.Q., Citrin, D.E., Restifo, N.P., Robbins, P.F., Wunderlich, J.R. *et al.* (2011) Durable complete responses in heavily pretreated patients with metastatic melanoma using T-cell transfer immunotherapy. *Clin Cancer Res*, **17**, 4550-4557.

9. Madan, R.A., Gulley, J.L., Fojo, T. and Dahut, W.L. (2010) Therapeutic cancer vaccines in prostate cancer: the paradox of improved survival without changes in time to progression. *Oncologist*, **15**, 969-975.

10. Rosenberg, S.A. and Restifo, N.P. (2015) Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, **348**, 62-68.

11. Eshhar, Z., Waks, T., Gross, G. and Schindler, D.G. (1993) Specific activation and targeting of cytotoxic lymphocytes through chimeric single chains consisting of antibody-binding domains and the gamma or zeta subunits of the immunoglobulin and T-cell receptors. *Proc Natl Acad Sci U S A*, **90**, 720-724.

12. Morgan, R.A., Dudley, M.E., Wunderlich, J.R., Hughes, M.S., Yang, J.C., Sherry, R.M., Royal, R.E., Topalian, S.L., Kammula, U.S., Restifo, N.P. *et al.* (2006) Cancer regression in patients after transfer of genetically engineered lymphocytes. *Science*, **314**, 126-129.

13. Robbins, P.F., Morgan, R.A., Feldman, S.A., Yang, J.C., Sherry, R.M., Dudley, M.E., Wunderlich, J.R., Nahvi, A.V., Helman, L.J., Mackall, C.L. *et al.* (2011) Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J*

Clin Oncol, **29**, 917-924.

14. June, C.H., O'Connor, R.S., Kawalekar, O.U., Ghassemi, S. and Milone, M.C. (2018) CAR T cell immunotherapy for human cancer. *Science*, **359**, 1361-1365.

15. Maude, S.L., Frey, N., Shaw, P.A., Aplenc, R., Barrett, D.M., Bunin, N.J., Chew, A., Gonzalez, V.E., Zheng, Z., Lacey, S.F. *et al.* (2014) Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med*, **371**, 1507-1517.

16. Coulie, P.G., Van den Eynde, B.J., van der Bruggen, P. and Boon, T. (2014) Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer*, **14**, 135-146.

17. Yarchoan, M., Johnson, B.A., Lutz, E.R., Laheru, D.A. and Jaffee, E.M. (2017) Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer*, **17**, 209-222.

18. Schumacher, T.N. and Schreiber, R.D. (2015) Neoantigens in cancer immunotherapy. *Science*, **348**, 69-74.

19. Gubin, M.M., Artyomov, M.N., Mardis, E.R. and Schreiber, R.D. (2015) Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J Clin Invest*, **125**, 3413-3421.

20. Linnemann, C., van Buuren, M.M., Bies, L., Verdegaal, E.M., Schotte, R., Calis, J.J., Behjati, S., Velds, A., Hilkmann, H., Atmioui, D.E. *et al.* (2015) High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat Med*, **21**, 81-85.

21. Vigneron, N. (2015) Human Tumor Antigens and Cancer Immunotherapy. *Biomed Res Int*, **2015**, 948501.

22. Stone, J.D., Harris, D.T. and Kranz, D.M. (2015) TCR affinity for p/MHC formed by tumor antigens that are self-proteins: impact on efficacy and toxicity. *Curr Opin Immunol*, **33**, 16-22.

23. Tian, S., Maile, R., Collins, E.J. and Frelinger, J.A. (2007) CD8+ T cell activation is governed by TCR-peptide/MHC affinity, not dissociation rate. *J Immunol*, **179**, 2952-2960.

24. Melero, I., Gaudernack, G., Gerritsen, W., Huber, C., Parmiani, G., Scholl, S., Thatcher, N., Wagstaff, J., Zielinski, C., Faulkner, I. *et al.* (2014) Therapeutic vaccines for cancer: an overview of clinical trials. *Nat Rev Clin Oncol*, **11**, 509-524.

25. Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T. and Old, L.J. (2005) Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer*, **5**, 615-625.

26. Richters, M.M., Xia, H., Campbell, K.M., Gillanders, W.E., Griffith, O.L. and Griffith, M. (2019) Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med*, **11**, 56.

27. Lu, Y.C. and Robbins, P.F. (2016) Cancer immunotherapy targeting neoantigens. *Semin Immunol*, **28**, 22-27.

28. Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M.J., van de Haar, J., Engin, H.B., de Prisco, N., Ideker, T., Hildebrand, W.H., Font-Burgada, J. *et al.* (2017) MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*, **171**, 1272-1283 e1215.

29. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L. *et al.* (2017) An immunogenic personal

neoantigen vaccine for patients with melanoma. *Nature*, **547**, 217-221.

30. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Lower, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrors, B. *et al.* (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, **547**, 222-226.

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

32. Hundal, J., Carreno, B.M., Petti, A.A., Linette, G.P., Griffith, O.L., Mardis, E.R. and Griffith, M. (2016) pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med*, **8**, 11.

33. Zhang, J., Mardis, E.R. and Maher, C.A. (2017) INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, **33**, 555-557.

34. Keenan, T.E., Burke, K.P. and Van Allen, E.M. (2019) Genomic correlates of response to immune checkpoint blockade. *Nat Med*, **25**, 389-402.

35. Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G. and Serody, J.S. (2019) Alternative tumour-specific antigens. *Nat Rev Cancer*, **19**, 465-478.

36. Licatalosi, D.D. and Darnell, R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*, **11**, 75-87.

37. Consortium, G.T. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648-660.

38. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A. *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*, **44**, 1084-1089.

39. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743-747.

40. Zhang, Y., Chen, F., Fonseca, N.A., He, Y., Fujita, M., Nakagawa, H., Zhang, Z., Brazma, A., Group, P.T.W., Group, P.S.V.W. *et al.* (2020) High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by rearrangement-mediated cis-regulatory alterations. *Nat Commun*, **11**, 736.

41. Lim, Y.W., Chen-Harris, H., Mayba, O., Lianoglou, S., Wuster, A., Bhangale, T., Khan, Z., Mariathasan, S., Daemen, A., Reeder, J. *et al.* (2018) Germline genetic polymorphisms influence tumor gene expression and immune cell infiltration. *Proc Natl Acad Sci U S A*, **115**, E11701-E11710.

42. PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N.R., Demircioglu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.V., Liu, F., Shiraishi, Y. *et al.* (2020) Genomic basis for RNA alterations in cancer. *Nature*, **578**, 129-136.

43. Felsher, D.W. and Bishop, J.M. (1999) Reversible tumorigenesis by MYC in hematopoietic lineages. *Mol Cell*, **4**, 199-207.

44. Tay, Y., Rinn, J. and Pandolfi, P.P. (2014) The multilayered complexity of ceRNA crosstalk and competition. *Nature*, **505**, 344-352.

45. Carthew, R.W. and Sontheimer, E.J. (2009) Origins and Mechanisms of miRNAs

and siRNAs. *Cell*, **136**, 642-655.

46. Malone, C.D. and Hannon, G.J. (2009) Small RNAs as guardians of the genome. *Cell*, **136**, 656-668.

47. Cagan, R.L. and Ready, D.F. (1989) Notch is required for successive cell decisions in the developing Drosophila retina. *Genes Dev*, **3**, 1099-1112.

48. Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet*, **10**, 155-159.

49. Shyu, A.B., Wilkinson, M.F. and van Hoof, A. (2008) Messenger RNA regulation: to translate or to degrade. *EMBO J*, **27**, 471-481.

50. Baralle, F.E. and Giudice, J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*, **18**, 437-451.

51. Lee, Y. and Rio, D.C. (2015) Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem*, **84**, 291-323.

52. Park, E., Pan, Z., Zhang, Z., Lin, L. and Xing, Y. (2018) The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet*, **102**, 11-26.

53. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y. and Pritchard, J.K. (2016) RNA splicing is a primary link between genetic variation and disease. *Science*, **352**, 600-604.

54. Bonnal, S.C., Lopez-Oreja, I. and Valcarcel, J. (2020) Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol*.

55. Oltean, S. and Bates, D.O. (2014) Hallmarks of alternative splicing in cancer. *Oncogene*, **33**, 5311-5318.

56. Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.

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

58. Urbanski, L.M., Leclair, N. and Anczukow, O. (2018) Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA*, **9**, e1476.

59. Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A. and Eyras, E. (2017) The Functional Impact of Alternative Splicing in Cancer. *Cell Rep*, **20**, 2215-2226.

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

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

62. Hsu, T.Y., Simon, L.M., Neill, N.J., Marcotte, R., Sayad, A., Bland, C.S., Echeverria, G.V., Sun, T., Kurley, S.J., Tyagi, S. *et al.* (2015) The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature*, **525**, 384-388.

63. David, C.J. and Manley, J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev*, **24**, 2343-2364.

64. Sibley, C.R., Blazquez, L. and Ule, J. (2016) Lessons from non-canonical splicing. *Nat Rev Genet*, **17**, 407-421.

65. Jia, Y., Xie, Z. and Li, H. (2016) Intergenically Spliced Chimeric RNAs in Cancer. *Trends Cancer*, **2**, 475-484.

66. Velusamy, T., Palanisamy, N., Kalyana-Sundaram, S., Sahasrabuddhe, A.A., Maher, C.A., Robinson, D.R., Bahler, D.W., Cornell, T.T., Wilson, T.E., Lim, M.S. *et al.* (2013) Recurrent reciprocal RNA chimera involving YPEL5 and PPP1CB in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A*, **110**, 3035-3040.

67. Zhang, Y., Gong, M., Yuan, H., Park, H.G., Frierson, H.F. and Li, H. (2012) Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. *Cancer Discov*, **2**, 598-607.

68. Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Del Pozo, A., Tress, M., Johnson, R., Guigo, R. *et al.* (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res*, **22**, 1231-1242.

69. Kannan, K., Wang, L., Wang, J., Ittmann, M.M., Li, W. and Yen, L. (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc Natl Acad Sci U S A*, **108**, 9172-9177.

70. Wu, C.S., Yu, C.Y., Chuang, C.Y., Hsiao, M., Kao, C.F., Kuo, H.C. and Chuang, T.J. (2014) Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res*, **24**, 25-36.

71. Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y. *et al.* (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res*, **44**, 2859-2872.

72. 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.

73. Vo, J.N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., Wu, Y.M., Dhanasekaran, S.M., Engelke, C.G., Cao, X. *et al.* (2019) The Landscape of Circular RNA in Cancer. *Cell*, **176**, 869-881 e813.

74. Li, J., Sun, D., Pu, W., Wang, J. and Peng, Y. (2020) Circular RNAs in Cancer: Biogenesis, Function, and Clinical Significance. *Trends Cancer*, **6**, 319-336.

75. Eisenberg, E. and Levanon, E.Y. (2018) A-to-I RNA editing - immune protector and transcriptome diversifier. *Nat Rev Genet*, **19**, 473-490.

76. Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K. *et al.* (2017) Dynamic landscape and regulation of RNA editing in mammals. *Nature*, **550**, 249-254.

77. Hwang, T., Park, C.K., Leung, A.K., Gao, Y., Hyde, T.M., Kleinman, J.E., Rajpurohit, A., Tao, R., Shin, J.H. and Weinberger, D.R. (2016) Dynamic regulation of RNA editing in human brain development and disease. *Nat Neurosci*, **19**, 1093-1099.

78. Nishikura, K. (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem*, **79**, 321-349.

79. Savva, Y.A., Rieder, L.E. and Reenan, R.A. (2012) The ADAR protein family. *Genome Biol*, **13**, 252.

80. Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O'Connell, M.A. and Li, J.B. (2013) Identifying RNA editing sites using RNA sequencing data

alone. *Nat Methods*, **10**, 128-132.

81. Chen, L., Li, Y., Lin, C.H., Chan, T.H., Chow, R.K., Song, Y., Liu, M., Yuan, Y.F., Fu, L., Kong, K.L. *et al.* (2013) Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat Med*, **19**, 209-216.

82. Fumagalli, D., Gacquer, D., Rothe, F., Lefort, A., Libert, F., Brown, D., Kheddoumi, N., Shlien, A., Konopka, T., Salgado, R. *et al.* (2015) Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. *Cell Rep*, **13**, 277-289.

83. Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H.T., Danan-Gotthold, M., Knisbacher, B.A., Eisenberg, E. and Levanon, E.Y. (2015) Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. *Cell Rep*, **13**, 267-276.

84. Kung, C.P., Maggi, L.B., Jr. and Weber, J.D. (2018) The Role of RNA Editing in Cancer Development and Metabolic Disorders. *Front Endocrinol (Lausanne)*, **9**, 762.

85. Han, L., Diao, L., Yu, S., Xu, X., Li, J., Zhang, R., Yang, Y., Werner, H.M.J., Eterovic, A.K., Yuan, Y. *et al.* (2015) The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell*, **28**, 515-528.

86. Goodier, J.L. and Kazazian, H.H., Jr. (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell*, **135**, 23-35.

87. Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T. *et al.* (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell*, **141**, 1171-

1182.

88. Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M. and Devine, S.E. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, **141**, 1253-1261.

89. Cowley, M. and Oakey, R.J. (2013) Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet*, **9**, e1003234.

90. Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, **300**, 1288-1291.

91. Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*, **12**, 1060-1067.

92. Lin, L., Shen, S., Tye, A., Cai, J.J., Jiang, P., Davidson, B.L. and Xing, Y. (2008) Diverse splicing patterns of exonized Alu elements in human tissues. *PLoS Genet*, **4**, e1000225.

93. Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet*, **41**, 563-571.

94. Clayton, E.A., Rishishwar, L., Huang, T.C., Gulati, S., Ban, D., McDonald, J.F. and Jordan, I.K. (2020) An atlas of transposable element-derived alternative splicing in cancer. *Philos Trans R Soc Lond B Biol Sci*, **375**, 20190342.

95. Group, P.T.C., Calabrese, C., Davidson, N.R., Demircioglu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.V., Liu, F., Shiraishi, Y. *et al.* (2020) Genomic

basis for RNA alterations in cancer. *Nature*, **578**, 129-136.

96. Erson-Bensan, A.E. and Can, T. (2016) Alternative Polyadenylation: Another Foe in Cancer. *Mol Cancer Res*, **14**, 507-517.

97. Dai, D., Wang, H., Zhu, L., Jin, H. and Wang, X. (2018) N6-methyladenosine links RNA metabolism to cancer progression. *Cell Death Dis*, **9**, 124.

98. Rodriguez, A.J., Czaplinski, K., Condeelis, J.S. and Singer, R.H. (2008) Mechanisms and cellular roles of local protein synthesis in mammalian cells. *Curr Opin Cell Biol*, **20**, 144-149.

99. Houseley, J. and Tollervey, D. (2009) The many pathways of RNA degradation. *Cell*, **136**, 763-776.

100. Kochenderfer, J.N., Wilson, W.H., Janik, J.E., Dudley, M.E., Stetler-Stevenson, M., Feldman, S.A., Maric, I., Raffeld, M., Nathan, D.A., Lanier, B.J. *et al.* (2010) Eradication of B-lineage cells and regression of lymphoma in a patient treated with autologous T cells genetically engineered to recognize CD19. *Blood*, **116**, 4099-4102.

101. Shahid, K., Khalife, M., Dabney, R. and Phan, A.T. (2019) Immunotherapy and targeted therapy-the new roadmap in cancer treatment. *Ann Transl Med*, **7**, 595.

102. Morgan, R.A., Yang, J.C., Kitano, M., Dudley, M.E., Laurencot, C.M. and Rosenberg, S.A. (2010) Case report of a serious adverse event following the administration of T cells transduced with a chimeric antigen receptor recognizing ERBB2. *Mol Ther*, **18**, 843-851.

103. Bonifant, C.L., Jackson, H.J., Brentjens, R.J. and Curran, K.J. (2016) Toxicity and management in CAR T-cell therapy. *Mol Ther Oncolytics*, **3**, 16011.

104. Vauchy, C., Gamonet, C., Ferrand, C., Daguindau, E., Galaine, J., Beziaud, L., Chauchet, A., Henry Dunand, C.J., Deschamps, M., Rohrlich, P.S. *et al.* (2015) CD20 alternative splicing isoform generates immunogenic CD4 helper T epitopes. *Int J Cancer*, **137**, 116-126.

105. Barrett, C.L., DeBoever, C., Jepsen, K., Saenz, C.C., Carson, D.A. and Frazer, K.A. (2015) Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy. *Proc Natl Acad Sci U S A*, **112**, E3050-3057.

106. Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.K. and Van Allen, E.M. (2018) Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*, **36**, 1056-1058.

107. Frankiw, L., Baltimore, D. and Li, G. (2019) Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol*, **19**, 675-687.

108. Li, H., Wang, J., Mor, G. and Sklar, J. (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, **321**, 1357-1361.

109. Funnell, T., Tasaki, S., Oloumi, A., Araki, S., Kong, E., Yap, D., Nakayama, Y., Hughes, C.S., Cheng, S.G., Tozaki, H. *et al.* (2017) CLK-dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. *Nat Commun*, **8**, 7.

110. Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N. and Regev, A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*, **20**, 213.

111. Fotakis, G., Rieder, D., Haider, M., Trajanoski, Z. and Finotello, F. (2020)

NeoFuse: predicting fusion neoantigens from RNA sequencing data. *Bioinformatics*, **36**, 2260-2261.

112. 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.

113. Zhu, L.P., He, Y.J., Hou, J.C., Chen, X., Zhou, S.Y., Yang, S.J., Li, J., Zhang, H.D., Hu, J.H., Zhong, S.L. *et al.* (2017) The role of circRNAs in cancers. *Biosci Rep*, **37**.

114. Xu, Z., Li, P., Fan, L. and Wu, M. (2018) The Potential Role of circRNA in Tumor Immunity Regulation and Immunotherapy. *Front Immunol*, **9**, 9.

115. Peng, X., Xu, X., Wang, Y., Hawke, D.H., Yu, S., Han, L., Zhou, Z., Mojumdar, K., Jeong, K.J., Labrie, M. *et al.* (2018) A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell*, **33**, 817-828 e817.

116. Zhang, M., Fritsche, J., Roszik, J., Williams, L.J., Peng, X., Chiu, Y., Tsou, C.C., Hoffgaard, F., Goldfinger, V., Schoor, O. *et al.* (2018) RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat Commun*, **9**, 3919.

117. Larouche, J.D., Trofimov, A., Hesnard, L., Ehx, G., Zhao, Q., Vincent, K., Durette, C., Gendron, P., Laverdure, J.P., Bonneil, E. *et al.* (2020) Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med*, **12**, 40.

118. Laumont, C.M., Vincent, K., Hesnard, L., Audemard, E., Bonneil, E., Laverdure, J.P., Gendron, P., Courcelles, M., Hardy, M.P., Cote, C. *et al.* (2018) Noncoding

regions are the main source of targetable tumor-specific antigens. *Sci Transl Med*, **10**.

119. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-628.

120. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.

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

122. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, **111**, E5593-5601.

123. 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.

124. Picardi, E. and Pesole, G. (2013) REDItools: high-throughput RNA editing detection made easy. *Bioinformatics*, **29**, 1813-1814.

125. Gao, Y., Zhang, J. and Zhao, F. (2018) Circular RNA identification based on multiple seed matching. *Brief Bioinform*, **19**, 803-810.

126. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R.J., Green, R.E. and Vollmers, C. (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci U S A*, **115**, 9726-9731.

127. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138.

128. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat Biotechnol*, **26**, 1146-1153.

129. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, **21**, 30.

130. Sharon, D., Tilgner, H., Grubert, F. and Snyder, M. (2013) A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*, **31**, 1009-1014.

131. Byrne, A., Beaudin, A.E., Olsen, H.E., Jain, M., Cole, C., Palmer, T., DuBois, R.M., Forsberg, E.C., Akeson, M. and Vollmers, C. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun*, **8**, 16027.

132. Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S. *et al.* (2016) Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*, **7**, 12065.

133. Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., Buck, D. and Au, K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res*, **6**, 100.

134. Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for

noisy long sequences. *Bioinformatics*, **32**, 2103-2110.

135. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218-223.

136. Calviello, L. and Ohler, U. (2017) Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet*, **33**, 728-744.

137. Schafer, S., Adami, E., Heinig, M., Rodrigues, K.E.C., Kreuchwig, F., Silhavy, J., van Heesch, S., Simaite, D., Rajewsky, N., Cuppen, E. *et al.* (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat Commun*, **6**, 7200.

138. Andreev, D.E., O'Connor, P.B., Loughran, G., Dmitriev, S.E., Baranov, P.V. and Shatsky, I.N. (2017) Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res*, **45**, 513-526.

139. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*, **33**, 981-993.

140. Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F. and Baranov, P.V. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res*, **22**, 2219-2229.

141. Cravatt, B.F., Simon, G.M. and Yates, J.R., 3rd. (2007) The biological impact of mass-spectrometry-based proteomics. *Nature*, **450**, 991-1000.

142. Mann, M., Kulak, N.A., Nagaraj, N. and Cox, J. (2013) The coming age of

complete, accurate, and ubiquitous proteomes. *Mol Cell*, **49**, 583-590.

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

144. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382-387.

145. Aebersold, R. and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**, 347-355.

146. Jaffe, J.D., Berg, H.C. and Church, G.M. (2004) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, **4**, 59-77.

147. Nesvizhskii, A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat Methods*, **11**, 1114-1125.

148. Nesvizhskii, A.I. and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, **4**, 1419-1440.

149. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582-587.

150. Istrail, S., Florea, L., Halldorsson, B.V., Kohlbacher, O., Schwartz, R.S., Yap, V.B., Yewdell, J.W. and Hoffman, S.L. (2004) Comparative immunopeptidomics of humans and their pathogens. *Proc Natl Acad Sci U S A*, **101**, 13268-13272.

151. Caron, E., Vincent, K., Fortier, M.H., Laverdure, J.P., Bramoulle, A., Hardy, M.P., Voisin, G., Roux, P.P., Lemieux, S., Thibault, P. *et al.* (2011) The MHC I

immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol Syst Biol*, **7**, 533.

152. Khodadoust, M.S., Olsson, N., Wagar, L.E., Haabeth, O.A., Chen, B., Swaminathan, K., Rawson, K., Liu, C.L., Steiner, D., Lund, P. *et al.* (2017) Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*, **543**, 723-727.

153. Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R. and Watts, J.D. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol*, **27**, 378-386.

154. Schiess, R., Mueller, L.N., Schmidt, A., Mueller, M., Wollscheid, B. and Aebersold, R. (2009) Analysis of cell surface proteome changes via label-free, quantitative mass spectrometry. *Mol Cell Proteomics*, **8**, 624-638.

155. DeVeale, B., Bausch-Fluck, D., Seaberg, R., Runciman, S., Akbarian, V., Karpowicz, P., Yoon, C., Song, H., Leeder, R., Zandstra, P.W. *et al.* (2014) Surfaceome profiling reveals regulators of neural stem cell function. *Stem Cells*, **32**, 258-268.

156. Lee, J.K., Bangayan, N.J., Chai, T., Smith, B.A., Pariva, T.E., Yun, S., Vashisht, A., Zhang, Q., Park, J.W., Corey, E. *et al.* (2018) Systemic surfaceome profiling identifies target antigens for immune-based therapy in subtypes of advanced prostate cancer. *Proc Natl Acad Sci U S A*, **115**, E4473-E4482.

157. Downing, J.R., Wilson, R.K., Zhang, J., Mardis, E.R., Pui, C.H., Ding, L., Ley, T.J. and Evans, W.E. (2012) The Pediatric Cancer Genome Project. *Nat Genet*,

44, 619-622.

158. Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E. *et al.* (2020) The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell*, **181**, 236-249.

Chapter 2 Pathway-guided analysis identifies Myc-dependent alternative pre-mRNA splicing in aggressive prostate cancers

2.1 Introduction

Alternative pre-mRNA splicing is a regulated process that governs exon choice and greatly diversifies the proteome. It is an essential process that contributes to development, tissue specification, and homeostasis and is often dysregulated in disease states (1). In cancer, this includes growth signaling, epithelial-to-mesenchymal transition, resistance to apoptosis, and treatment resistance (2). In prostate cancer, our area of interest, the most notable splicing change is the emergence of the ligand-independent androgen receptor ARV7 isoform in response to hormone deprivation (3). Other examples include proangiogenic splice variants of VEGFA (4), tumorigenic variants of the transcription factors ERG and KLF6 (5,6), and antiapoptotic splicing of BCL2L2 (7,8). However, the intersection of upstream oncogenic signaling, pre-mRNA splicing, and the biological processes affected by those splicing events has not been defined at a global level.

Prostate cancers progress from hormone-responsive, localized disease to hormone-independent, metastatic disease accompanied by changes in gene expression and mutations that confer cell-autonomous growth and therapeutic resistance (9). The study of disease progression from primary prostate adenocarcinoma (PrAd) to metastatic, castration-resistant prostate cancer (mCRPC) and treatment-related neuroendocrine prostate cancer (NEPC) has been aided by large-scale genomic and transcriptomic studies of patient samples representing each form of the disease (10-13). Examples of driver alterations found in precursor lesions and primary tumors include TMPRSS2-ERG translocations and PTEN loss (14). Metastatic tumors are characterized by Myc and AR amplification (15,16). NEPC includes near-universal loss of TP53 signaling by inactivating mutation as well as chromosomal loss of RB1 (17). Sequencing efforts and subsequent functional experiments have identified prostate cancer driver alterations and defined the impact of gene expression networks on prostate cancer phenotypes. These studies have led to the successful development of new therapeutics targeting AR signaling and DNA repair in advanced disease (18,19).

Prostate cancer progression is also associated with shifts in alternative pre-mRNA splicing patterns, but this process is not well understood (20). Investigations of global changes in exon usage in prostate cancer have focused on stage- or race-specific comparisons (21-25). Comparisons of tumor-adjacent benign material and PrAd identified intron retention and exon skipping events in the biomarkers KLK3 and AMACR, respectively (22). Others studying NEPC and PrAd have shown that a network of splicing events controlled by the serine–arginine RNA-binding protein SRRM4 contributes to the neuroendocrine phenotype (26-28). Comparisons of European American and African American (AA) PrAd samples identified an AA-specific splice variant of PIK3CD that enhanced AKT/mTOR signaling (23). How these splicing alterations connect to the driver alterations described above remains to be explored.

The accumulation of RNA-sequencing (RNA-Seq) data in large databases presents a unique opportunity to conduct an analysis of alternative splicing across the full range of prostate cancer disease states. For our study, we prepared a unified dataset of large, publicly available RNA-Seq datasets representing normal tissue, tumor-adjacent benign tissue, primary adenocarcinoma, metastatic castration-resistant adenocarcinoma, and treatment-related metastatic NEPC. However, handling datasets of this size requires splicing analysis software with greater efficiency than what is currently available. To analyze these hundreds of datasets, we created an improved version of our rMATS software (dubbed rMATS-turbo) that can handle this volume of RNA-Seq data (29,30).

We identify a high-confidence set of exons whose incorporation varies across prostate cancer disease states. By combining expression-level and exon-level analyses, we developed a pathway-guided strategy to examine the impact of oncogenic pathways on incorporation of these exons. This correlational analysis implicates Myc, mTOR, and E2F signaling in the control of exon choice in spliceosomal proteins. To further investigate the contributions of Myc signaling to exon choice, we developed unique engineered human prostate cell lines with regulated Myc expression. Functional experiments in these cell lines identify Myc-dependent exons and experimentally confirm that cassette exon choice in many splicing regulatory proteins is responsive to Myc expression level. These exons often encode frameshifts or premature termination codons (PTCs) that would result in nonsense-mediated decay (NMD). We show that an ultraconserved, NMD-determinant exon in the RNA-binding protein SRSF3 is particularly responsive to Myc signaling. Our results implicate Myc signaling as a regulator of alternative splicing-coupled NMD (AS-NMD) as part of a program of growth control.

2.2 Results

2.2.1 Exon-Level Analysis Defines the Landscape of Alternative Pre-mRNA Splicing Across the Prostate Cancer Disease Spectrum

We combined RNA-Seq data from disparate published datasets representing 876 samples of normal tissue, benign tumor-adjacent material, primary adenocarcinoma, metastatic castration-resistant adenocarcinoma (mCRPC), and treatment-related NEPC (**Figure 2.1A**) (10-13,31,32). Metaanalyses of RNA-Seq data with gene- or isoform-level counts are subject to confounding batch effects and rely on existing isoform annotation (33). Exon-level analysis, however, uses a ratio-based methodology to estimate exon incorporation, which may be more robust against batch effects and confounding factors in large-scale RNA-Seq datasets (34-37). In addition, exon-level analysis can detect novel exon–exon junctions and is thus independent of previous annotation.

To facilitate alternative splicing analysis in this and other large RNA-Seq datasets, we developed rMATS-turbo (also known as rMATS 4.0.2), a computational pipeline that permits the efficient capture, storage, and analysis of splicing information from very large-scale raw RNA-Seq data. This improved pipeline refactors the original ratio-based rMATS software that we developed for splicing analysis in RNA-Seq data to optimize it for very large-scale RNA-Seq datasets and is now available for public use (29,30). It offers significant improvements in speed and data storage efficiency.

We applied rMATS-turbo to the combined RNA-Seq dataset and identified over 330,000 different cassette exons across all prostate samples. Previous estimates of the diversity of splicing events in human cells vary, but are generally of the same order of magnitude (38). We also identified tens of thousands of additional alternative splicing events (**Figure 2.1A**), including alternative 5′ and 3′ splice sites, mutually exclusive exons, and retained introns. For this study, we focused on cassette exons, as these are the most well-defined type of alternative splicing event. We should note that although the rMATS-turbo software detected numerous mutually exclusive exons, most of these events were in fact part of more complex alternative splicing events; thus, we did not include these mutually exclusive exons in downstream analyses.

Filtering of these exons for coverage (≥10 splice junction reads per event), cross-sample variance (range of percent-spliced-in [PSI] > 5%; mean skipping or inclusion > 5%) and commonality (events detected in ≥1% of all samples) produced a set of 13,149 high-confidence exons with variable incorporation across samples (see **2.4 Methods**). Principal-component analysis (PCA) of this exon usage matrix grouped samples of the same disease phenotype regardless of dataset (**Figure 2.1B**). By comparison, a similar unsupervised analysis of isoform-level count-based metric from the same metadataset grouped samples more by dataset of origin than disease phenotype (**Supplementary Figure 2.7 A and B**). This result is consistent with prior observations that the exon-level splicing analysis is more robust against batch effects and other confounding factors in large-scale RNA-Seq datasets (35-37).

2.2.2 Combining Gene Pathway Analysis and Exon Usage Identifies Exon Correlates of Oncogenic Signaling

Genomic studies of prostate cancer have identified driver alterations associated with disease progression (39). We sought to define how the variable cassette exons we identified and the biological processes they participate in might relate to these oncogenic signals. Instead of selecting single oncogenes for study, we developed PEGASAS (pathway enrichment-guided activity study of alternative splicing), a pathway-guided analytic strategy that uses gene signatures to estimate the activities of signaling pathways and to discover potential downstream exon changes (**Figure 2.2A**). Gene signature-based analyses use an ensemble of features (a set of genes collectively) to estimate pathway activity and outperform single-gene measurements (40). To mitigate potential batch effects in the expression data, we utilized a rank-based metric to calculate the signature score, providing a more robust measure of pathway activity as it is in essence normalized on a per-sample basis (41).

We employed the hallmark gene signature sets maintained by the Molecular Signatures Database (MSigDB) (42). These 50 sets represent a diverse and well-validated array of cellular functions and signaling pathways. To assess the performance of these signatures in our combined dataset, we examined signature scores for the AR, Myc Targets V2, and MTOR gene sets across five different prostate phenotypes. Consistent with previously reported observations of pathway activation in prostate cancer progression, the androgen response gene signature scores we measured were lowest in NEPC samples (**Supplementary Figure 2.8A**). Similarly, MTOR and Myc signature scores were higher in mCRPC samples than in normal tissues. The Myc and MTOR signature scores increased between normal healthy donors (Genotype-Tissue Expression [GTEx]) and tumor-adjacent normal (TCGA-PRAD), consistent with field cancerization and tumor–stromal interaction effects on gene expression reported previously by others (43).

We then scored each sample in our metadataset for all 50 pathways and correlated this score with the data matrix of over 13,000 variable cassette exons (**Dataset S 2.1**). After filtering for correlation strength and false-discovery rate (FDR), each pathway returned between 11 and 1,330 exon correlates (**Dataset S 2.1**). The 10 gene sets that returned the greatest number of exon correlates with a Pearson's correlation coefficient greater than 0.3 or less than −0.3 are shown (**Figure 2.2B**). Nine out of 10 of these gene sets had exon correlates found in genes with strong functional enrichment by gene ontology (adjusted *P* value < 0.05).

2.2.3 Cassette Exons Correlating with Myc, E2F, and MTOR Signaling Are Enriched in Splicing-Related Genes

We next examined the biological processes specified by the genes containing the variant exons correlated with prostate cancer-relevant hallmark gene sets (**Figure 2.2C**). We also added a signature that describes transcriptional activity due to TMPRSS signaling as this common prostate cancer alteration is not represented by a hallmark gene set (44). Here, we represent the network of data as a hive plot to show how exons (left axis) correlate with signaling pathways (middle axis) and the functional enrichment of genes containing those correlated exons (right axis) (45). Gene ontology analysis indicated that the relatively small number of exons correlated with AR or Notch were modestly enriched in cell adhesion and chromatin remodeling processes. Surprisingly, the numerous exon correlates of Myc, E2F, and MTOR were strongly enriched in genes related to the spliceosome and alternative pre-mRNA splicing. In addition, the overlap in the exon sets correlated with Myc, E2F, and MTOR was striking, with 50 to 60% of exons held in common (**Figure 2.2D**). These pathways play central roles in growth control and are frequently codysregulated in human cancers, so a shared set of exons might be expected from a correlation analysis.

2.2.4 Myc-Correlated Exons Are Found in the Oncogenes SRSF3 and HRAS

Given the centrality of Myc signaling in tumorigenesis, tumor maintenance, and tumor progression in a multitude of tissue lineages (46,47) including the prostate, this pathway was selected for further investigation (15,48,49). The validity of these correlational results critically depends on the integrity of the underlying gene signature used to produce them. We therefore performed additional validation steps on the "MYC Targets V2" hallmark gene set by examining its performance in The Cancer Genome Atlas prostate adenocarcinoma RNA-Seq dataset (TCGA-PRAD) that has accompanying patient outcomes data (32). We noted that samples with genomic amplifications of Myc had higher signature scores on average, as did samples that overexpressed Myc at the mRNA level (**Supplementary Figure 2.9A**). To examine whether these relatively small changes in signature score had clinical relevance, we performed Kaplan–Meier survival analyses using the "MYC Targets V2" signature, Myc genomic amplification status, or Myc single-gene overexpression status as strata. The Myc gene signature was equally predictive of overall survival as genomic amplification status and outperformed single-gene expression stratification (**Supplementary Figure 2.9B**).

Convinced of the performance of the Myc signature by these additional tests, we performed further analysis of the 1,039 Myc-correlated exons we identified in the prostate metadataset (**Figure 2.3A and Dataset S 2.1**). Unsupervised clustering of these 1,039 exons also grouped the samples by phenotype (**Supplementary Figure 2.9C**), identifying patterns in Myc-dependent exon incorporation that varied accordingly.

Two examples among the most strongly Myc-correlated cassette exons from our analysis are found in SRSF3 and HRAS (**Figure 2.3B**). Incorporation of the identified alternative exon in SRSF3 is anticorrelated with the Myc signature score (**Figure 2.3B, Left**). When examined by cancer phenotype, incorporation of this exon decreases as prostate cancer progresses from normal tissue to primary tumor and is even lower in mCRPC samples (**Figure 2.3C, Left**). Incorporation of this exon in NEPC is slightly higher, consistent with the Myc signature scores in these samples (**Supplementary Figure 2.8A**).

SRSF3 is a serine–arginine splicing factor that can act as a proto-oncogene and also participates in transcription termination and DNA repair (50-53). The exon in question is ultraconserved throughout evolution and contains an in-frame stop codon. Also known as a poison exon, this sequence functions as a PTC (**Supplementary Figure 2.9D, Top**). Incorporation of this PTC has been shown previously to reduce SRSF3 expression levels by inducing NMD of the transcript (54,55). These data suggest increased Myc signaling leads to increased exon skipping, reduced NMD, and increased expression of SRSF3.

A cassette exon in HRAS was also anticorrelated with Myc activity (**Figure 2.3B, Right**). When examined by cancer phenotype, exon skipping increased with tumor progression (**Figure 2.3C, Right**). HRAS is a well-known oncogene that cooperates with Myc to induce carcinogenesis in multiple tissues (56,57). Inclusion of the cassette exon and the stop codon it contains results in the truncated HRAS p19 product instead of the p21 form (58). HRASp19 lacks the cysteine residues in the carboxyl-terminal domain of HRASp21 required for nuclear translocation and RAS-driven transformation and may function instead as a tumor suppressor (58,59). This exon is conserved in mammals (**Supplementary Figure 2.9D, Bottom**). Incorporation of this exon is anticorrelated with Myc activity, suggesting that Myc can drive increased expression of oncogenic HRAS by affecting its splicing.

2.2.5 Myc-Correlated Exons in Prostate Cancers Are Highly Conserved in Breast and Lung Adenocarcinomas

To determine whether the observed effects of Myc activity on splicing were prostate cancer specific, we performed a similar correlation analysis on a second hormone-dependent malignancy, breast adenocarcinoma, as well as on a hormone-independent epithelial malignancy, lung adenocarcinoma. The normal tissue and cancer RNA-Seq datasets for this analysis were drawn from TCGA (TCGA-BRCA and TCGA-LUAD) datasets and the GTEx collection of normal tissue (31,60,61). We performed a similar correlation between Myc signature score and exon usage as described above (**Figure 2.3D**). The Myc signature scores in breast and lung tissues behaved similarly to those in the prostate tissues, with increases in score at each step when moving from normal to tumor-adjacent normal to carcinoma (**Supplementary Figure 2.9E**). We identified 2,852 Myc-correlated cassette exons in breast samples and 2,465 in lung samples using the same filtering criteria for the prostate study (**Supplementary Figure 2.9F**). The exon list includes the same anticorrelated exon in SRSF3, as shown for lung samples (**Figure 2.3D, fourth panel**). Intersecting this set with our previously defined set of Myc-responsive prostate cancer exons (**Figure 2.3A**), we found extensive overlap and similar exon incorporation behavior in the three sets (**Figure 2.3E**). The triple intersection was even more strongly enriched for RNA-binding proteins (**Figure 2.3F**). Our analysis suggests the exon incorporation response to Myc overexpression is conserved across these cancers.

2.2.6 Creation of an Engineered Model of Advanced Prostate Cancer with

Regulated Myc Expression from Benign Human Prostate Cells to Define Myc-Dependent Exon Events

Correlation analysis strongly implicates Myc, E2F, and MTOR signaling in the control of exons related to alternative pre-mRNA splicing but cannot define the individual contribution of each pathway to the observed phenotype. We therefore sought to determine whether the Myc-correlated splicing effects we observed were indeed Myc dependent.

Numerous studies of the effect of Myc overexpression have described large numbers of Myc target genes with significant tissue heterogeneity (62,63). The presence of complex background genetics, undefined driver alterations, and tissue culture-specific phenomena further complicate the study of Myc biology (64). We therefore constructed a model of advanced prostate cancer by the transformation of benign human prostate epithelial cells with defined oncogenes (**Figure 2.4A**) (65). We have previously shown that the enforced expression of Myc and myristoylated (activated) AKT1 (myrAKT1) generates androgen receptor-independent adenocarcinoma (66,67). MyrAKT1 is included to phenocopy the activation of AKT1 that follows deletion of the tumor suppressor PTEN, a common event in prostate cancer tumorigenesis. Here, we cloned the Myc cDNA into a doxycycline-inducible promoter lentiviral construct, whereas MyrAKT1 was constitutively expressed (**Figure 2.4B** and **2.6 Appendix**).

After lentiviral transduction of isolated human prostate basal cells (**Supplementary Figure 2.10A**), we initiated the organoid culture and subsequent subcutaneous xenograft tumor outgrowth in immunocompromised mice in the constant presence of the drug (**Supplementary Figure 2.10 B and C**). As previously reported, only doubly transduced cells resulted in tumor outgrowth (**Figure 2.4C**). The histologic appearance and marker expression patterns of the xenograft outgrowths were similar to those previously published with constitutive constructs (**Figure 2.4D** and **Supplementary Figure 2.10D**). The xenograft outgrowths were dissociated, and plated in tissue culture conditions with doxycycline to initiate autonomously growing cell lines (**Figure 2.4E**). We repeated the entire procedure to generate three independent cell lines from the prostate epithelium of three different human specimens.

2.2.7 Myc Withdrawal Affects Expression of Splicing-Related Genes

Withdrawal of doxycycline from the Myc/myrAKT1 cell lines resulted in the rapid, dose-dependent loss of Myc protein expression, consistent with its previously reported short half-life (**Figure 2.5A and Supplementary Figure 2.11A**) (68). The cells also rapidly slowed their growth with increased G_0/G_1 fraction at 24 h (**Supplementary Figure 2.11 B and C**). They adopted a senescent-like phenotype after prolonged Myc withdrawal with up-regulation of P21 (**Figure 2.5A**). A similar consequence of Myc withdrawal in oncogene-addicted transformed cells has been previously reported (69).

We performed RNA-Seq on samples from Myc-high and Myc-low conditions to define Myc-dependent genes and exons in our model system. These samples were sequenced with high read depth (>100 M reads) to enable accurate quantification of alternative splicing in downstream analysis. Primary analysis of the RNA expression data showed that thousands of genes were highly responsive to Myc withdrawal (CuffDiff *q*-value < 0.05) (**Figure 2.5B**). Gene ontology analysis identified enrichment of several growth-related biological processes among the Myc-dependent genes (**Figure 2.5C**). Of note, genes involved in RNA processing were among the most highly enriched in this subset. This is consistent with previous reports of Myc's broad control of the growth phenotype. The regulated Myc expression system also allowed us to independently validate the Myc signature score we used in our correlation analysis (**Figure 2.5D**).

2.2.8 Experimentation Confirms Myc-Regulated Exons Are Enriched in Splicing-Related Proteins and Often Encode PTCs

We applied rMATS-turbo to analyze Myc-regulated exon usage in our engineered cell lines. To accommodate the paired nature of the dataset (comparing Myc-high and Myc-low conditions for each), we employed the PAIRADISE statistical test to the rMATS-turbo output (70). After filtering for coverage (≥10 splice junction reads per event), effect size (|deltaPSI| > 5%), and FDR < 5%, this analysis yielded 1,970 cassette exons that significantly changed incorporation in response to Myc withdrawal (**Figure 2.6 A and B** and **Dataset S 2.1**). We note that, among the Myc-dependent exons, we again identified the alternative exons in SRSF3 and HRAS described above, experimentally demonstrating that their incorporation is dependent on Myc signaling (**Figure 2.6C**). The relative incorporation of the poison exon in SRSF3 increased when Myc was withdrawn, which would act to decrease the amount of SRSF3 protein in response to oncogene loss. We confirmed by immunoblotting that SRSF3 protein levels decreased relative to the housekeeping protein GAPDH in this experimental setting (**Supplementary Figure 2.12A**).

Similar to the correlational data from the patient specimens, the Myc-dependent exons were strikingly enriched in genes affecting RNA splicing-related processes (**Figure 2.6D**). Intersecting this set of exons with the Myc-correlated exons in patient tissue identified 147 common exons (**Figure 2.6E**), a highly significant overlap (*P* = 1.03 × 10-90). The remaining exons may not be responsive to short-term withdrawal of Myc in the cell line model or may be correlated with other signaling derangements that often accompany Myc deregulation in patient cancers (e.g., E2F or MTOR).

Alternative pre-mRNA splicing can regulate transcript levels through the incorporation or skipping of NMD-determinant exons (71). We hypothesized that Myc-driven exon choice in splicing proteins could contribute to the regulation of their expression levels. To examine the functional outcome of Myc-driven splicing changes on NMD, we annotated the 147 exons in the patient data–cell line intersection for PTCs and frameshifts (**Figure 2.6F and Dataset S 2.1**). These 147 exons correspond to 124 genes, 30 of which were RNA-binding proteins by gene ontology designation. We annotated all these exons using the Ensembl database to identify those that contained verified PTCs. We supplemented this annotation by parsing the remaining exons to identify those predicted to produce a frameshift within the coding sequence of the parent mRNA transcript. We found that 36 of the 43 exons in RNA-binding genes encode a PTC, a frameshift, or both (**Dataset S 2.1**). These exons represent a set of Myc-responsive sequences that act to regulate transcript abundance of proteins involved in alternative pre-mRNA splicing.

2.3 Discussion

This analysis was powered by rMATS-turbo, a fast, flexible, and extensible software package that allows rigorous examination of exon usage across disparate datasets. These public datasets have moderate read depth (50 to 75 M reads) and variable read length (50 to 75 bp). Here, we have used rMATS-turbo to perform a comprehensive survey of exon usage across the entire spectrum of prostate cancer disease progression. This exon-level analysis allows the correlation of exon matrices with any continuous metadata of interest. Our PEGASAS methodology identifies putative exon targets of cancer signaling networks. Its successful application to prostate, breast, and lung cancer datasets suggests that pathway-driven analysis of alternative splicing in pancancer data will also be of interest.

The engineered human prostate cell lines we developed with regulated Myc expression represent a unique opportunity to examine the consequences of Myc withdrawal on a defined genetic background. We employed them to identify over a thousand exons that significantly altered incorporation rates in response to Myc withdrawal, again with a striking enrichment for splicing-related proteins. The effects of Myc overexpression have been shown in other cancer contexts to have deleterious effects on splicing (72,73). In Eu-Myc lymphoma cells, a Myc-target gene, PRMT5, is essential for maintaining splicing fidelity. Similarly, a component of the core spliceosome, BUD31, was shown to be a MYC-synthetic lethal gene in a human mammary transformation model. Others have shown that Myc-driven changes in splicing are in part accomplished by the induction of the canonical serine–arginine splicing factor SRSF1 (74). Further elucidation of the events downstream from Myc overexpression that lead to splicing changes is needed.

We note that Myc dysregulates the splicing of the PTC-containing exon in the serine–arginine protein SRSF3 (54,55). This exon is Myc-correlated in both the prostate and breast cancer datasets, Myc-regulated in our tissue culture model, and ultraconserved. SRSF3 is known to alter the splicing of a number of downstream targets, as well as to autoregulate its own splicing. In a feedback loop, high levels of SRSF3 protein bind to its pre-mRNA transcript and promote inclusion of the poison exon (55). However, in the transformed setting we examined, Myc-high states were associated with high levels of SRSF3 expression and low levels of poison exon incorporation. This suggests Myc signaling may allow escape from this autoregulatory mechanism and stabilize SRSF3 transcripts despite high SRSF3 protein levels. SRSF3 itself has been recently shown to regulate splicing of NMD-determinant exons in chromatin modifier proteins during the induction of pluripotent stem cells (75). Given the role of Myc signaling in the acquisition of stem-like phenotypes and the stem-like state of advanced cancers, the mechanism that connects Myc overexpression to splicing changes in SRSF3 deserves further exploration (76,77).

Furthermore, the phenomenon of Myc-regulated poison exons is not limited to SRSF3. We identified a number of exons in splicing proteins from patient tissues with experimentally validated Myc dependence in vitro that also contained PTCs. Alternative splicing coupled to NMD has been widely described as a mechanism controlling levels of splicing factors and other RNA-binding proteins (78). These splicing events are often autoregulated by the encoded protein or cross-regulated by a related paralog (79). Our data on Myc regulation indicate that this system of AS-NMD is also more globally regulated as part of a program of growth control. We postulate that these exons and regulation of them by Myc may be part of an adaptive response to alter spliceosomal throughput in response to high transcriptional flux.

One limitation of our study is that RNA and protein levels of the same genes are often poorly correlated (80). The potential for premature stop codons introduced by alternative splicing to induce NMD could further skew this relationship. Further studies of the relationship between Myc levels and NMD-determinant exons in splicing-related proteins should include proteomic measurements.

Our study provides further insight into the relationship between Myc signaling and alternative splicing changes that could be used to guide the development of splicing-targeted cancer therapy (81). Future work will need to establish the specificity of these exon events for cells with oncogenic levels of Myc expression to avoid simultaneously targeting rapidly dividing normal cell types.

2.4 Methods

Descriptions of the gene ontology analysis, overlap enrichment assessment, lentiviral constructs, organotypic human prostate transformation assay, xenograft outgrowth, cell line derivation, and other tissue culture experiments are available in **2.6.1 Supplementary Methods.**

2.4.1 RNA-Seq Data Processing Framework

A comprehensive RNA-Seq dataset was compiled from published prostate cancer and normal prostate datasets that reflect the full progression of prostate cancer. In total, 876 samples were downloaded from different sources. RNA-Seq Fastq files of normal prostate samples [GTEx Consortium (31)] and prostate cancer samples [Beltran et al. study (10), Robinson et al. study (11), and Stand-Up-To-Cancer study (12)] were downloaded from dbGAP (82,83) via *fastq-dump* in SRA toolkit. RNA-Seq Fastq files from TCGA primary prostate cancer and adjacent benign samples were downloaded from GDC via gdc-client

(84).

A unified RNA-Seq processing framework was constructed to perform read mapping as well as gene and isoform quantification on the collected multiphenotypic prostate RNA-Seq samples. Specifically, read mapping was done by STAR 2.5.3a (85) with a STAR 2-pass function enabled to improve the detection of splicing junctions. The STAR genome index was built with–sjdbOverhang 100 as a generic parameter to handle differences in read length of RNA-Seq samples from various sources. The genome annotation file was downloaded from GENCODE V26 (86) under human genome version hg19 (GRCh37). The subsequent gene/isoform expression quantification is performed by Cufflinks (87) with default parameters.

RNA-Seq alternative splicing quantification is conducted uniformly with a newly engineered version (version 4.0.2) of the rMATS-turbo software package (29,30). An exon-based ratio metric, commonly defined as PSI ratio, was employed to measure the alternative splicing events. The PSI ratio is calculated as follows:

$$\psi = \frac{I/L_I}{S/L_S + I/L_I},$$

where *S* and *I* are the numbrs of reads mapped to the junction supporting skipping and inclusion form, respectively. Effective length L is used for normalization.

Customized scripts were applied to calculate PSI value for each individual alternative splicing event from the rMATS-turbo junction count output. To build a confident set of exon events, the splice junction of each event was required to be covered by no less than 10 splice junction reads. Additionally, each event was required to have a PSI range greater than 5% across the entire dataset (|maxPSI − minPSI| > 5%), with a mean skipping or inclusion value over 5%. Events with missing values in the majority (over 99%) of samples were removed.

2.4.2 Analysis and Evaluation of Alternative Splicing Profile of Prostate Cancer Metadataset

PCA was applied to inspect the RNA-Seq–derived gene expression/alternative splicing profiles of our multiphenotypic prostate cancer dataset. First, the matrix of sample vs. fragments per kilobase of transcript per million mapped reads/PSI value was produced by customized scripts. Then, the matrix was completed and imputed by KNN method (*knnImputation* in *DMwR* package) (88) for missing values. Last, the matrix was mean centered and scaled (PSI matrix is not scaled). PCA was conducted via prcomp function in R. The top five PCs were inspected, but only the first two that describe the highest percentage of the variance are shown.

In addition, silhouette width was applied to assess the fitness of PCA clustering results derived from alternative splicing or gene/isoform expression measurements (89). Specifically, disease conditions were used as sample labels to compute the silhouette width of each cluster. Average silhouette widths were compared between PCA clustering results with different metrics (90). The R package *cluster* (91) was used for Silhouette calculation based on PCA results and disease phenotype labels.

2.4.3 PEGASAS

In order to identify exon incorporation shifts that could correspond to oncogenic pathway alterations during tumor progression, a correlation-based analysis was developed to define signaling pathway correlated alternative splicing events. It involves two major steps.

The first step is to define signaling pathway activity and alternative splicing levels. The quantification of gene expression and alternative splicing is detailed in ***2.4.1 RNA-Seq Data Processing Framework***. Signaling pathway activity can be characterized by assessing the expression level of its target genes as a set relative to other genes (42). The MSigDB (92) has compiled gene sets (42) for the use with gene set enrichment analysis (GSEA) (93) software or similar applications. Here, a group of well-defined gene sets, known as hallmarks (42), was selected to assess a wide range of pathways in prostate cancers. To measure the activity of a given signaling pathway gene set, all genes (both genes within the gene set as well as those not in the gene set) were ranked according to their gene expression values, then a weight was assigned to each gene based on the number of genes in the set (pathway or nonpathway) they belonged to. This was used to construct empirical distributions for both sets, and a two-sample Kolmogorov–Smirnov test statistic, which is the supremum of the differences between the two distributions, was computed as a measure of the activity of the signaling pathway, i.e., an “activity score.” Given the same gene set and gene annotation, the higher the score, the higher the activity of a signaling pathway in a sample. Note that the score should not be used to compare across signaling pathways as each gene set has distinct number of genes, which affects the score.

The second step is to identify pathway activity-correlated alternatively spliced exons. For each pathway, the pathway activity score defined above was correlated with all of the AS events identified by rMATS-turbo. The Pearson correlation coefficient was computed for each pathway–exon pair across samples in the dataset. A Pearson correlation coefficient with an absolute value >0.3 was considered as correlated. Data points for each pathway–exon pair were permutated 5,000 times locally to produce empirical P values to filter out faulty correlations caused by data structure or missing data points. A stringent empirical *P* value < 2 × 10-4 was required for this analysis. The analytical framework performs streamlined analysis of multiple gene sets (e.g., 50 hallmark gene sets). Customized scripts were implemented to generate the summary plot.

2.4.4 Cell Line Gene Expression and Alternative Splicing Differential Analysis

The same RNA-Seq processing framework described above was applied to quantify gene expression and alternative splicing of Myc cell line samples. Differentially expressed genes were identified and visualized by the Cuffdiff and cummeRbund pipeline with a threshold of *q*-value < 0.05. Skipped exon events quantified by rMATS-turbo were analyzed by the PAIRADISE statistical model for conducting paired tests of between Myc +/− conditions (70,87). PAIRADISE with equal.variance = TRUE was used to perform the test. The resulting events were first filtered by the coverage and deltaPSI requirements (≥10 splice junction reads per event, |deltaPSI| > 0.05). Then, an FDR 5% cutoff was applied to identify significant differential alternative splicing events between the on and off states of the engineered Myc cell line.

2.4.5 Code availability

The computational pipeline of PEGASAS is available at https://github.com/Xinglab/PEGASAS (94), and custom scripts used to perform filtering, analysis, and visualization have been deposited separately at https://github.com/Xinglab/Myc-regulated_AS_PrCa_paper (95)

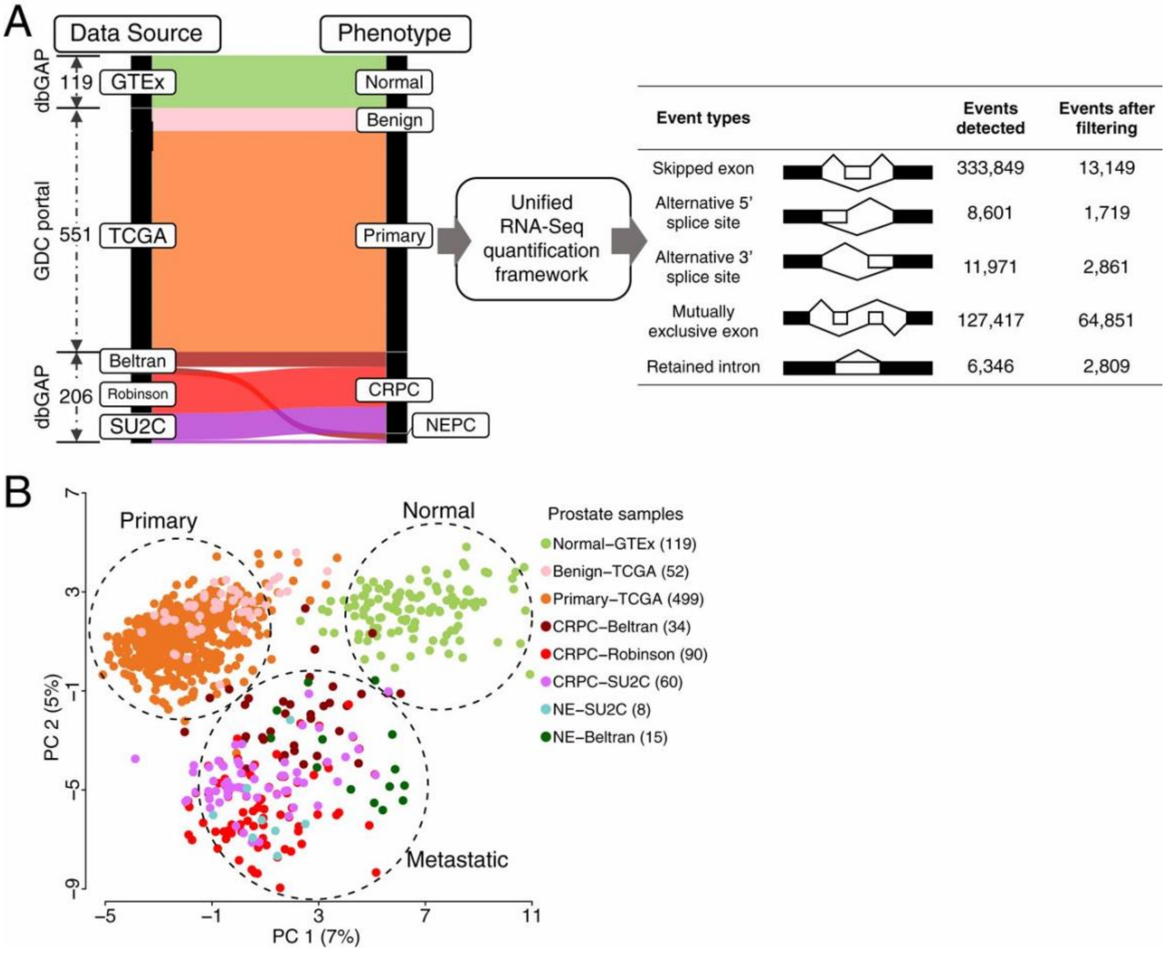
2.4.6 Data availability

Raw sequencing files (fastq) from the engineered cell lines and gene expression matrices are available through Gene Expression Omnibus (accession no. GSE141633) (96). The PSI and gene expression matrices for the prostate metadataset are also available from the same source. The normal prostate expression data from GTEx used for the analyses described in the manuscript were obtained from dbGaP (https://www.ncbi.nlm.nih.gov/gap) accession no. phs000424 (accessed 1 October 2018). Data on primary prostate cancers were obtained from the TCGA Research Network and downloaded from the Genomic Data Commons (http://portal.gdc.cancer.gov/projects/TCGA-PRAD) accession no. phs000178 (accessed 1 October 2017). Additional datasets on metastatic prostate cancers are available by controlled access through dbGaP with accession nos. phs000909, phs000673, and phs000915 (accessed 1 October 2018).

Acknowledgements

This research was supported by funds from NIH/National Cancer Institute under Awards R01CA220238, U01CA233074, U24CA232979, and P50CA092131; the Parker Institute for Cancer Immunotherapy, Grant 20163828; the University of California, Los Angeles, Tumor Cell Biology Training Grant T32CA009056; and the Office of the Assistant Secretary of Defense for Health Affairs through the Prostate Cancer Research Program under Award W81XWH-16-1-0216. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense. The results published here are in part based upon data generated by the TCGA Research Network (http://cancer.gov/tcga) and data generated by the Genotype-Tissue Expression (GTEx) Project (https://commonfund.nih.gov/gtex). The GTEx Project is supported by the Common Fund of the Office of the Director of the NIH, and by National Cancer Institute, National Human Genome Research Institute, National Heart, Lung, and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke.

2.5 Figures



Event types	Events detected	Events after filtering
Skipped exon	333,849	13,149
Alternative 5' splice site	8,601	1,719
Alternative 3' splice site	11,971	2,861
Mutually exclusive exon	127,417	64,851
Retained intron	6,346	2,809

Figure 2.1 A global, exon-level analysis of alternative pre-mRNA splicing in normal prostate and prostate cancers identifies patterns of exon usage in RNA-binding proteins

(A) Schematic with alluvial plot depicting the data-processing workflow combining RNA-Seq data from various prostate tissue disease states (*Left*) and summary table depicting various exon events detected by rMATS-turbo before and after filtering for splice junction reads coverage, PSI range, and commonality (*Right*). The alluvial plot depicts the sorting of patient RNA-Seq datasets from individual studies on the Left into prostate phenotypes on the *Right*. **(B)** Scatter plot depiction of an unsupervised PCA of exon usage matrices from eight different prostate datasets representing healthy tissue, tumor-adjacent benign tissue, primary prostate cancer, metastatic castration-resistant prostate cancer (mCRPC), and treatment-associated neuroendocrine prostate cancer (NEPC).

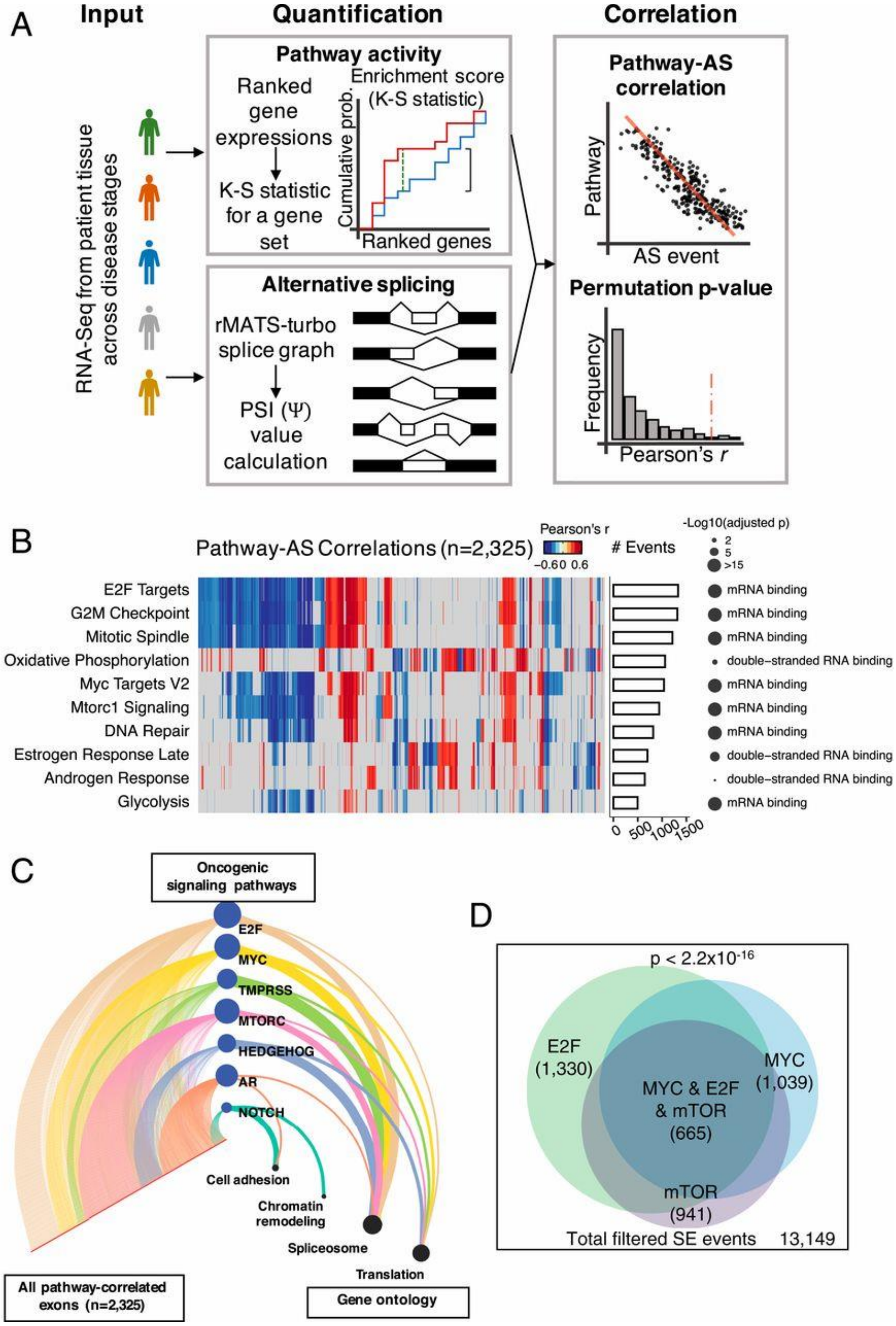


Figure 2.2 Pathway enrichment-guided activity study of alternative splicing (PEGASAS) analysis identifies exon correlates of oncogenic signaling in prostate cancers

(A) Workflow diagram describing PEGASAS correlation of gene signature score with exon usage. Each sample is scored for a gene expression signature of interest. Gene signature scores are correlated with exon usage matrices to identify pathway-correlated exon incorporation changes. **(B)** Heatmap of the correlation coefficients of the exon changes correlated with gene signatures in the Molecular Signatures Database (MSigDB) hallmark gene sets as generated by PEGASAS. The 10 signatures that returned the highest number of exon correlates are shown here. Each row depicts the results of the correlation to a single hallmark signature. Each column represents a single exon. The color represents the strength and direction of the correlation (red positive, blue negative) of a single exon with each pathway. Columns are sorted by hierarchical clustering. Rows are ranked by total number of exon correlates passing statistical metrics for each pathway (# Events, bar chart). The gene ontology term with the highest enrichment for the genes containing pathway-correlated exons and the corresponding *P* value are also depicted. The *P* values correspond to the gene ontology enrichment and are not a measure of significance of pathway–exon correlation. (C) Hive plot depiction of exons correlated with selected prostate cancer-related gene signatures and the biological processes associated with genes containing those exons. All pathway-correlated exons are displayed on the left axis. Seven well-known prostate cancer driver pathways are represented as nodes on the middle axis. The area of these nodes reflects the number of exons correlated with each pathway. The right axis depicts four summary gene ontology terms. The width of the edges connecting the nodes on the middle axis to the nodes on the right axis is proportional to the enrichment of each pathway for each biological process. The size of the nodes on the right axis is proportional to the total number of exons associated with each biological process. (D) Area-proportional Venn diagram depicting the intersection of Myc-, E2F-, and

MTOR-correlated exons in prostate cancer. Exons must share the same correlation direction (positive or negative) to appear in the intersection. AS, alternative splicing; K-S, Kolmogorov–Smirnov; SE, skipped exon.

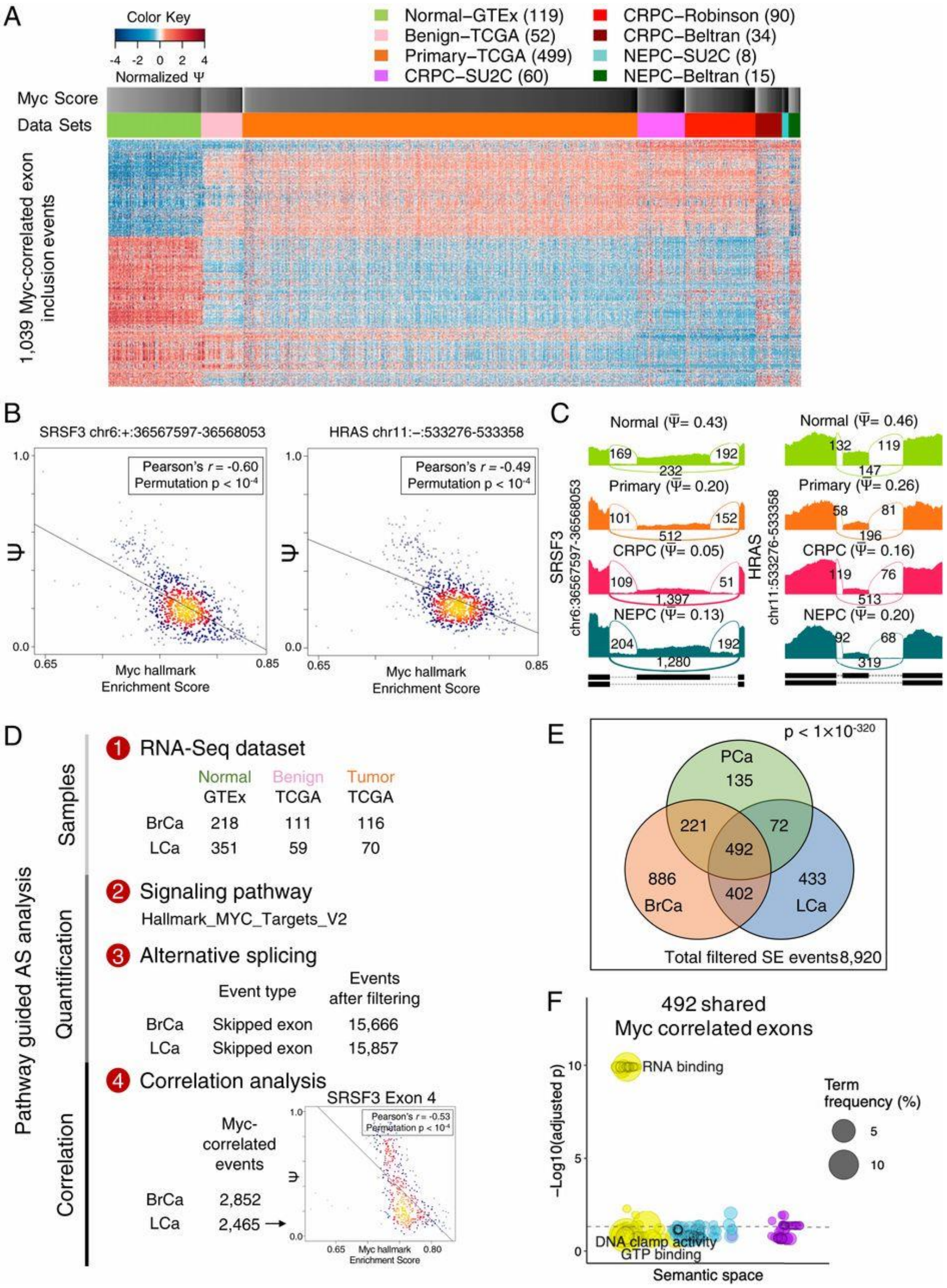


Figure 2.3 Exon incorporation events correlated with Myc activity are strongly enriched in RNA-binding proteins and are conserved in prostate and breast cancers

(A) Heatmap depiction of exon usage of 1,039 Myc-correlated exons across prostate cancer datasets in healthy tissue, primary adenocarcinoma, metastatic adenocarcinoma, and neuroendocrine prostate cancer (NEPC). Columns represent samples ordered by disease phenotype and sorted by Myc Targets V2 signature score within each group. The Myc score annotation is colored from white (low) to black (high) based on the rank-transformed signature score of patient samples across the datasets. Rows represent exon inclusion events ordered by hierarchical clustering. **(B)** Scatterplots depicting examples of cassette exons in SRSF3 and HRAS transcripts whose incorporation is negatively correlated with Myc gene signature score. **(C)** Sashimi plots depicting average cassette exon incorporation levels of exons in SRSF3 and HRAS in prostate cancer datasets separated by cancer phenotype. Sashimi plots depict density of exon-including and exon-skipping reads as determined by rMATS-turbo analysis. **(D)** Workflow diagram for performing pathway-guided alternative splicing analysis on normal and cancerous breast and lung tissues. Each sample is scored for the Myc Targets V2 signature and correlated with the exon usage matrix to identify pathway-correlated exon incorporation changes. **(E)** Venn diagram indicating the intersection between Myc-correlated exon sets in prostate cancers with breast and lung adenocarcinomas. Exons must share the same correlation direction (positive or negative) to appear in the intersection. **(F)** REVIGO chart depicting the gene ontology of genes containing the 492 Myc-correlated exons from the triple intersection described above. SE, skipped exon.

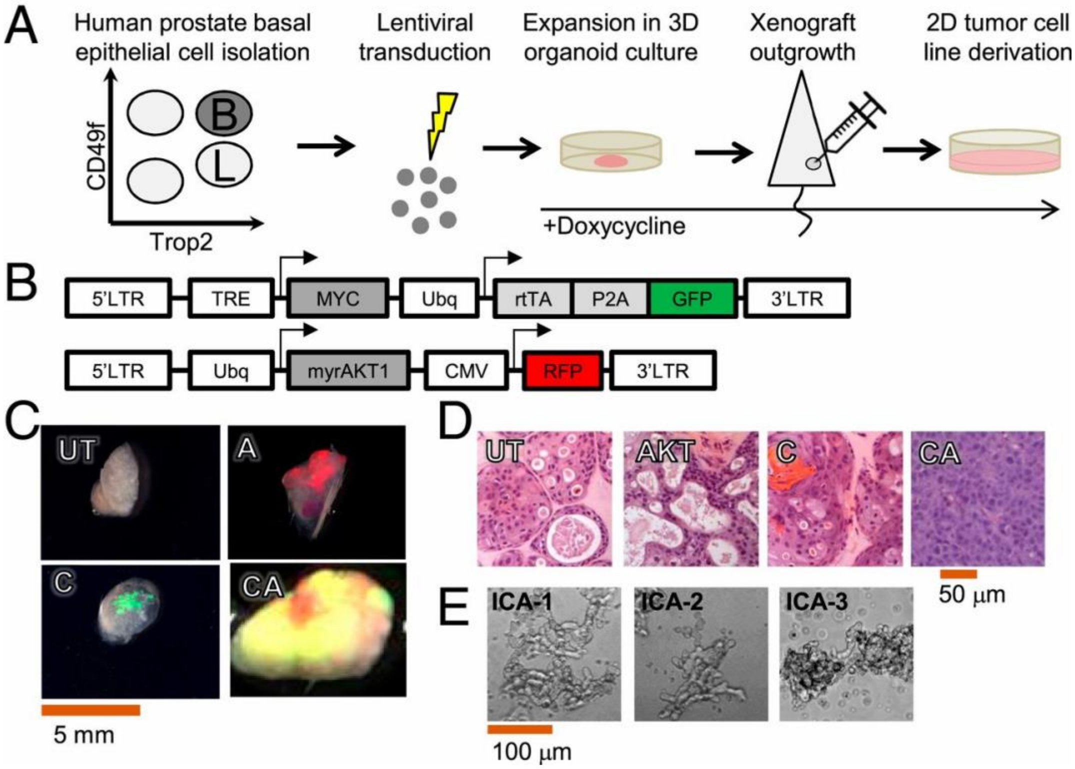


Figure 2.4 Enforced expression of activated AKT1 and doxycycline-regulated c-Myc initiates AR-negative PrAd in human prostate cells

(A) Workflow diagram for derivation of Myc/myrAKT1-transformed human prostate cells from benign epithelium. “B” = Trop2+/CD49fhi basal cells; “L” = Trop2+/CD49flo luminal cells. **(B)** Depiction of lentiviral vectors used to enforce doxycycline-regulated expression of Myc and constitutive expression of myrAKT1. Histologic sections of transduced organoids. **(C)** Photomicrographs and fluorescent overlay of recovered grafts and tumor outgrowth after lentiviral transduction and subcutaneous implantation in NSG mice. A, myrAKT1 transduction (RFP); C, c-Myc transduction (GFP); CA, dual transduction with c-Myc and myrAKT1 (GFP and RFP merge depicted as yellow); UT, untreated. **(D)** Hematoxylin and eosin (H&E) stain of histologic sections of recovered grafts and tumor outgrowths. **(E)** Photomicrographs of cell lines ICA-1, ICA-2, and ICA-3 derived from tumor outgrowths growing as suspended rafts in tissue culture.

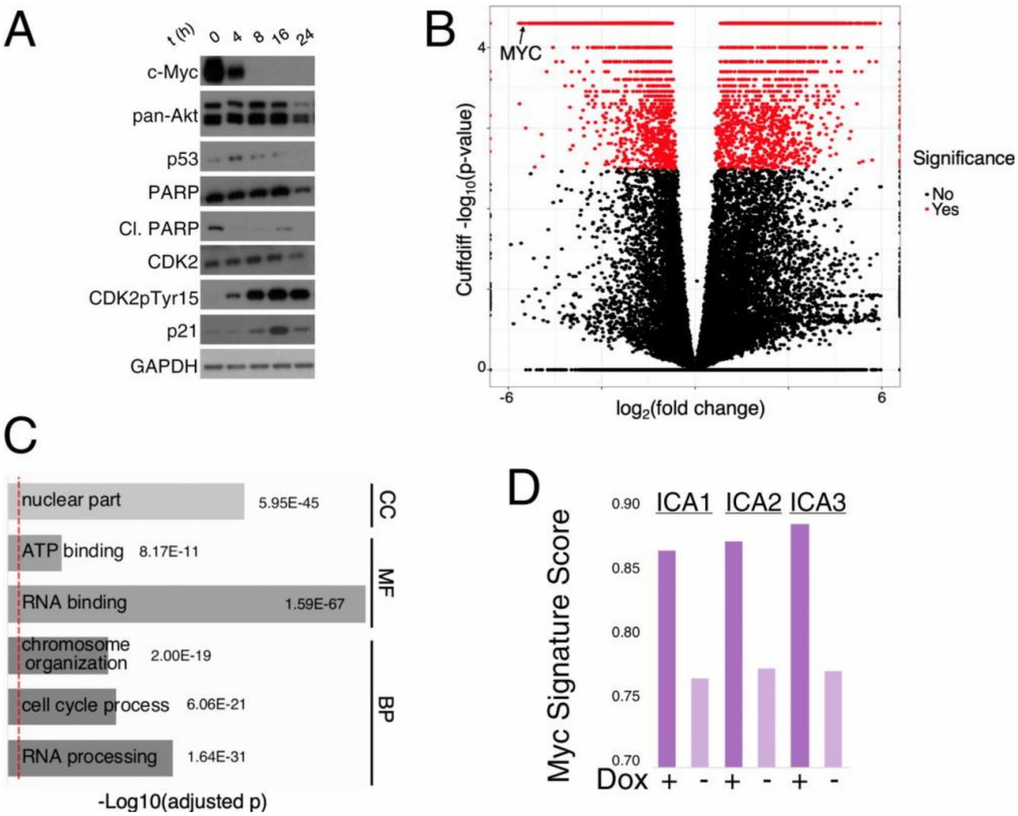
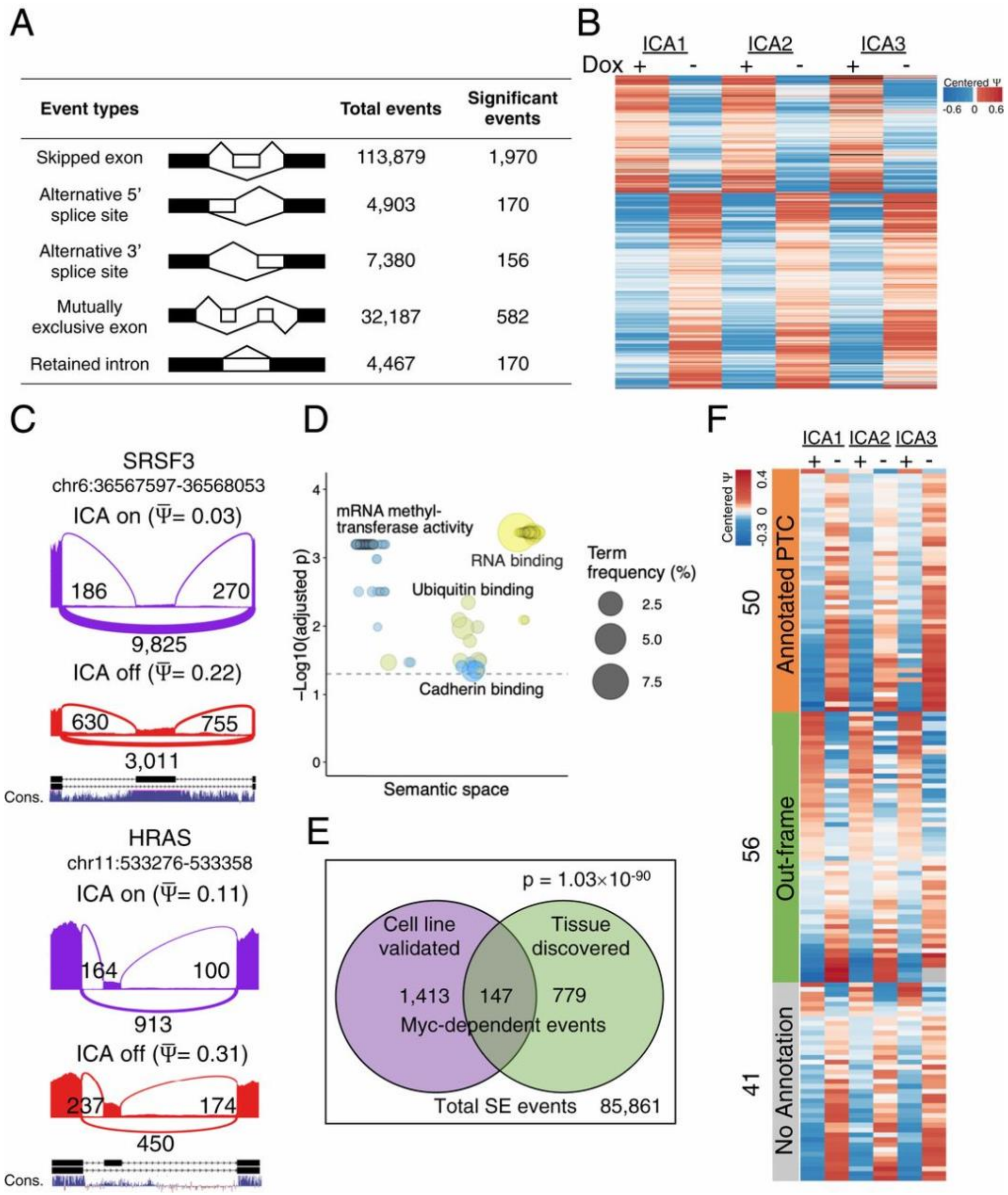


Figure 2.5 Myc loss in the engineered cell lines produces a senescent-like phenotype and strongly affects the expression of RNA binding proteins

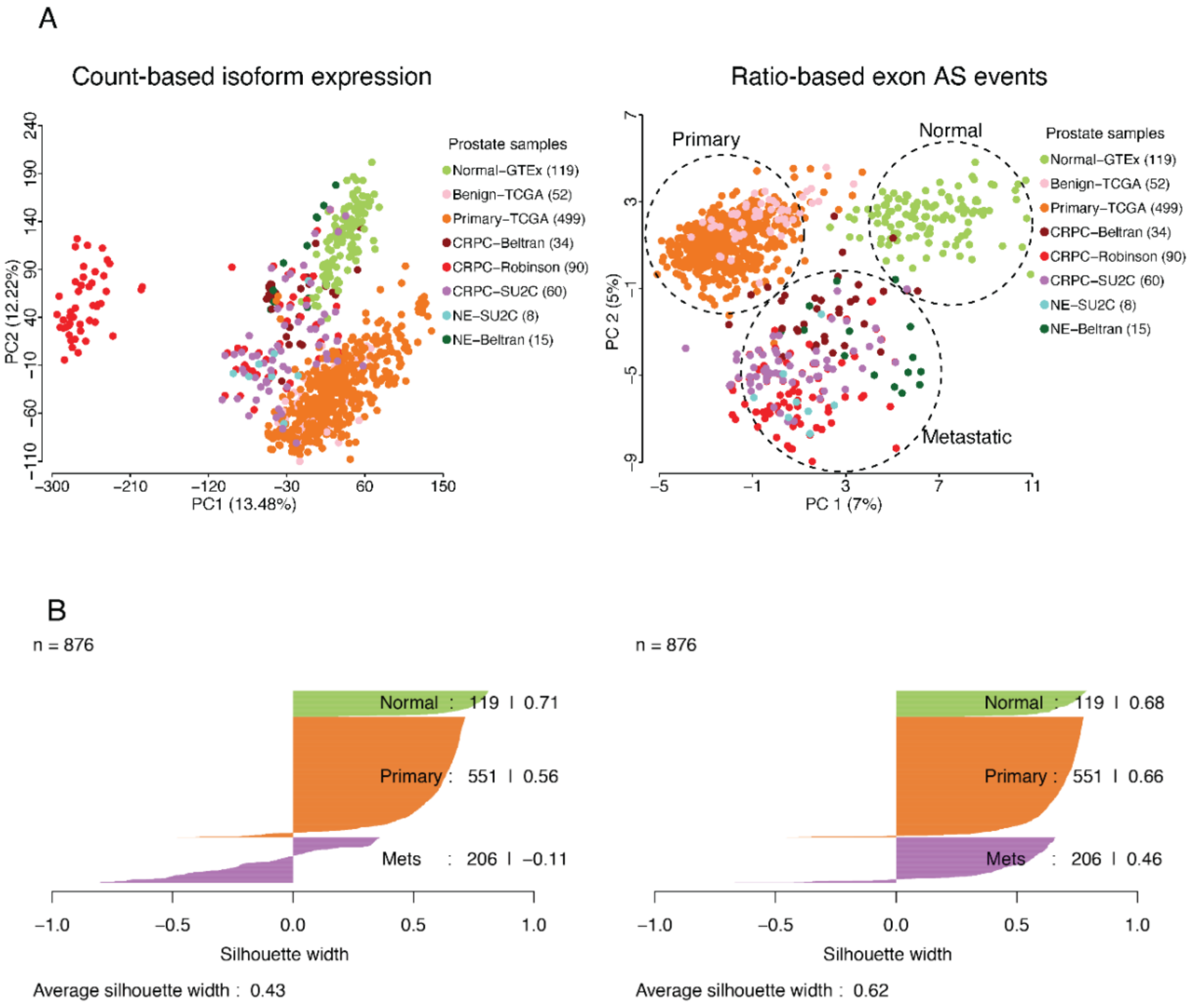
(A) Western blot of lysates from ICA1 cells withdrawn from doxycycline in a time course examining Myc expression and changes in proteins related to cell cycle state. Each of the three cell lines was examined in this manner, and the data shown are representative of all three. **(B)** Volcano plot of gene-level expression changes after Myc withdrawal. Genes down-regulated upon Myc loss appear on the left-hand side of the plot. Gene expression changes with the Cuffdiff *q*-value of <0.05 appear red. **(C)** Selected top gene ontology terms from the gene ontology analysis of Myc-dependent gene expression changes displaying strong enrichment for RNA binding. BP, Biological Process; CC, Cellular Component; MF, Molecular Function**. (D)** Comparison of Myc Targets V2 signature score levels in engineered cell lines in the presence and absence of doxycycline.



Event types	Total events	Significant events
Skipped exon	113,879	1,970
Alternative 5' splice site	4,903	170
Alternative 3' splice site	7,380	156
Mutually exclusive exon	32,187	582
Retained intron	4,467	170

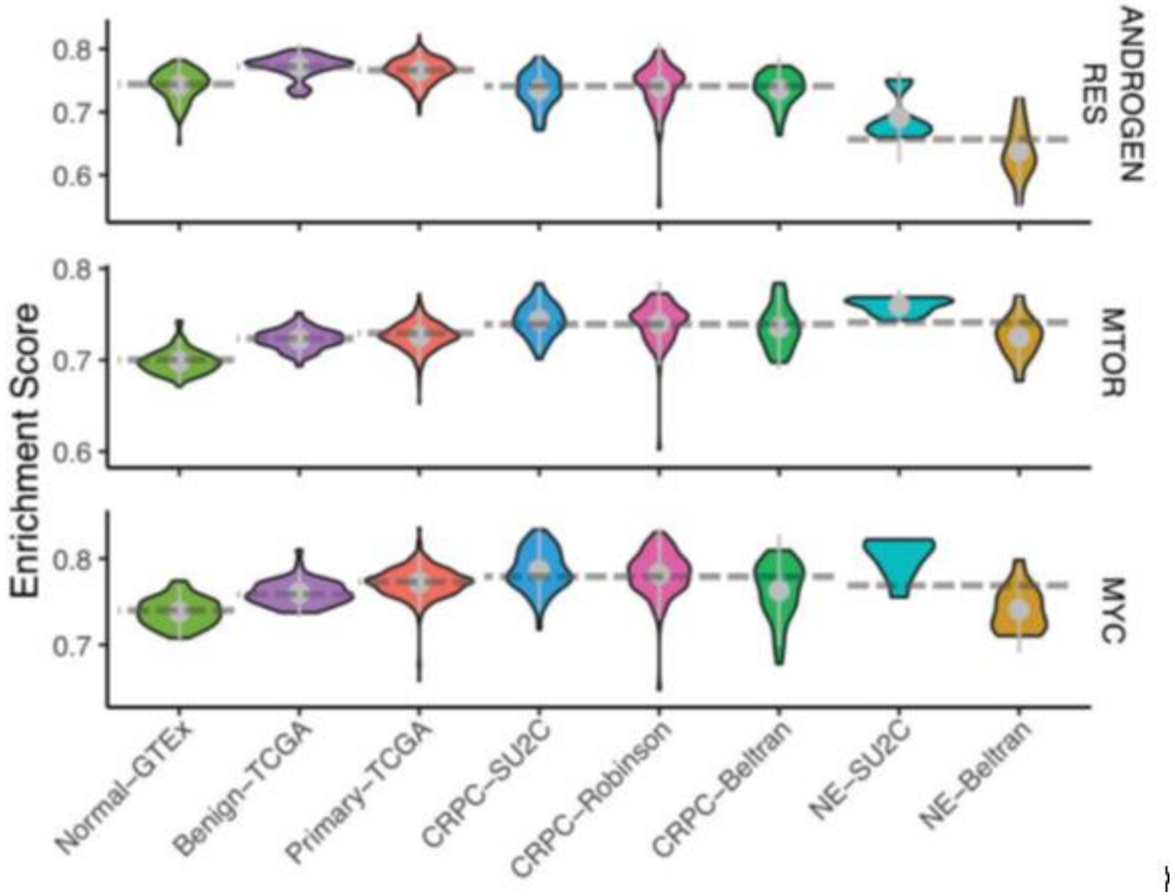
Figure 2.6 Exon-level splicing analysis of c-Myc/myrAKT1 transformed human prostate cells identifies Myc-dependent exon incorporation events in splicing regulatory proteins

(A) Summary table of exon incorporation changes occurring after Myc withdrawal. **(B)** Heatmap depicting changes in exon incorporation of 1,970 Myc-dependent cassette exons in three independent engineered cell lines. **(C)** Sashimi plots depicting the change in splice junction RNA-Seq reads in SRSF3 and HRAS exons in the engineered cell lines following Myc withdrawal. Sashimi plots depict density of exon-including and exon-skipping reads as determined by rMATS-turbo analysis. **(D)** REVIGO scatter plot depicting gene ontology terms enriched among genes containing exons whose incorporation is responsive to Myc withdrawal. Semantic distance is a measurement of relatedness between gene ontology terms calculated by REVIGO. Representative gene ontology terms have been selected to describe each cluster. The dashed line indicates adjusted *P* = 0.05. **(E)** Venn diagram depicting the overlap between Myc-dependent exons (purple) and Myc-correlated exons identified in patient tissues (green). Exons must change incorporation level with Myc in the same direction as the correlation (positive or negative) in order to appear in the intersection of the two sets. **(F)** Heatmap depicting the annotated outcome of exon changes in validated Myc-dependent exons. The annotation identifies exons likely to produce PTCs (orange) or frameshifts (green). SE, skipped exon.



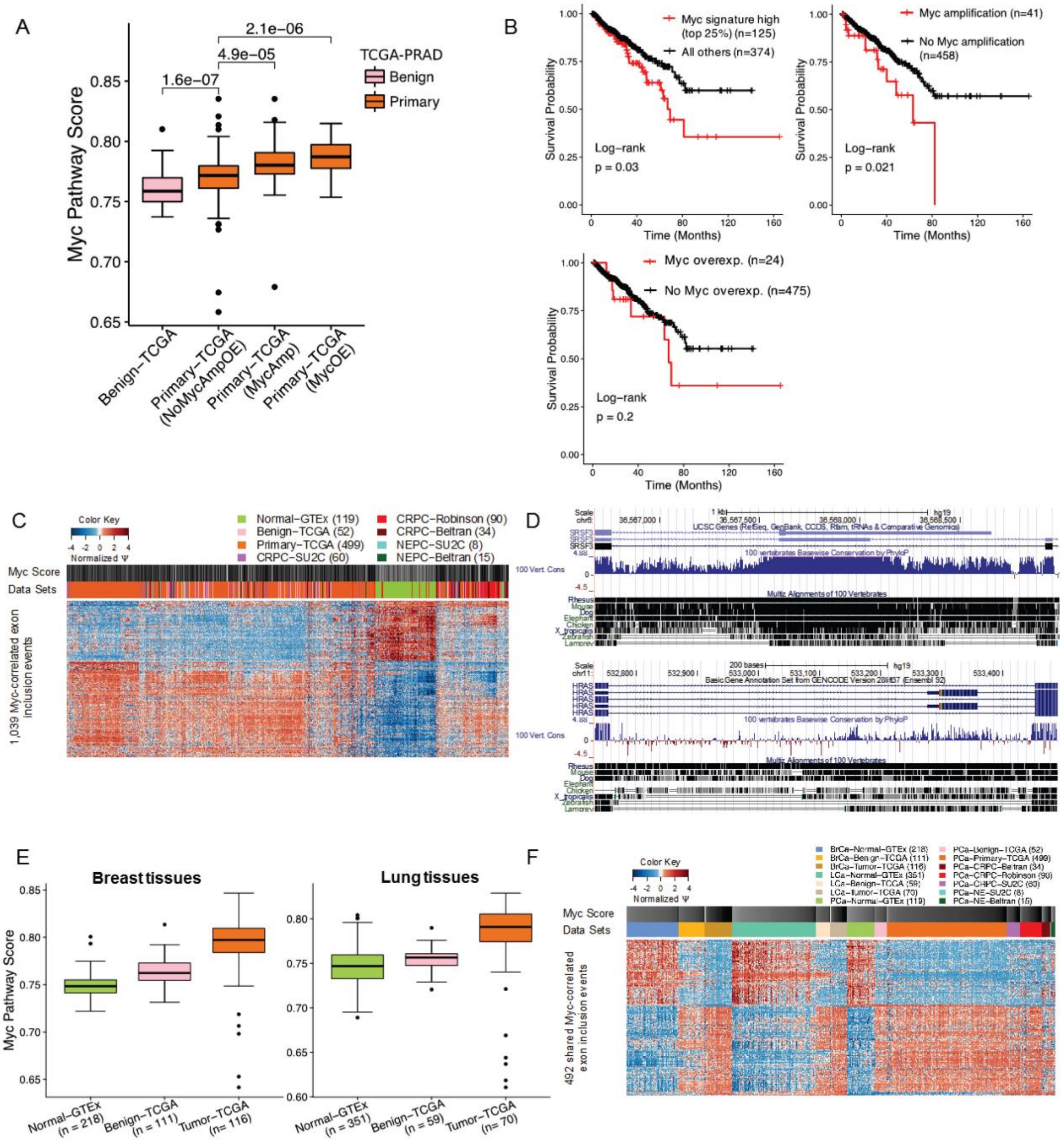
Supplementary Figure 2.7 Comparison of count-based and ratio-based isoform-level analyses of prostate RNA-Seq datasets

(A) Unsupervised analysis of count-based isoform expression from a combined prostate cancer dataset (left panel). The same methodology applied to the ratio-based alternative splicing approach from Figure 2.1B in the main text is shown for comparison (right panel). **(B)** Silhouette width-based comparison of clustering fitness for each of the principle component analyses shown above. Mets, metastatic.



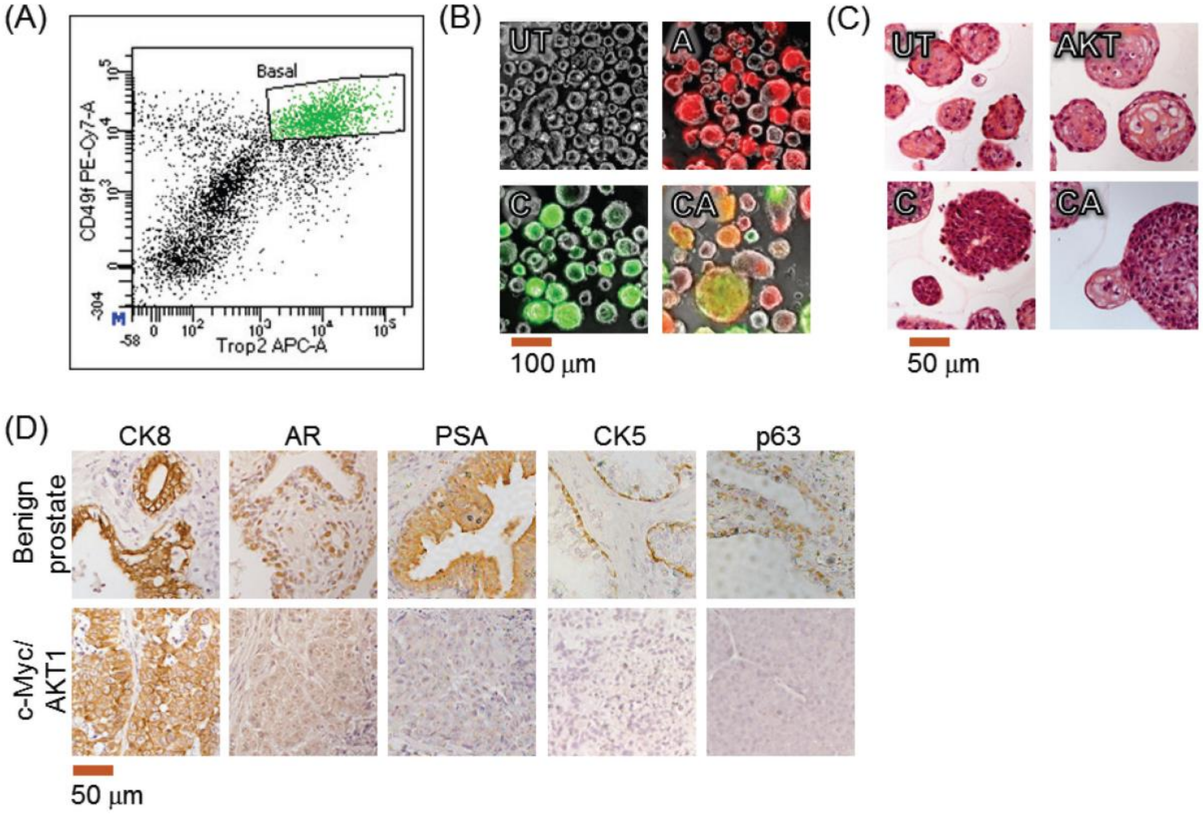
Supplementary Figure 2.8 Gene signature analysis identifies a common set of exons correlated with Myc, E2F, or mTOR pathways

Violin plot depiction of gene signature scores of AR, Myc Targets V2, and mTOR sets across prostate cancer datasets. Dashed lines indicate averages across datasets profiling a disease phenotype (normal prostate, benign prostate, primary prostate cancer, mCRPC, and NEPC).



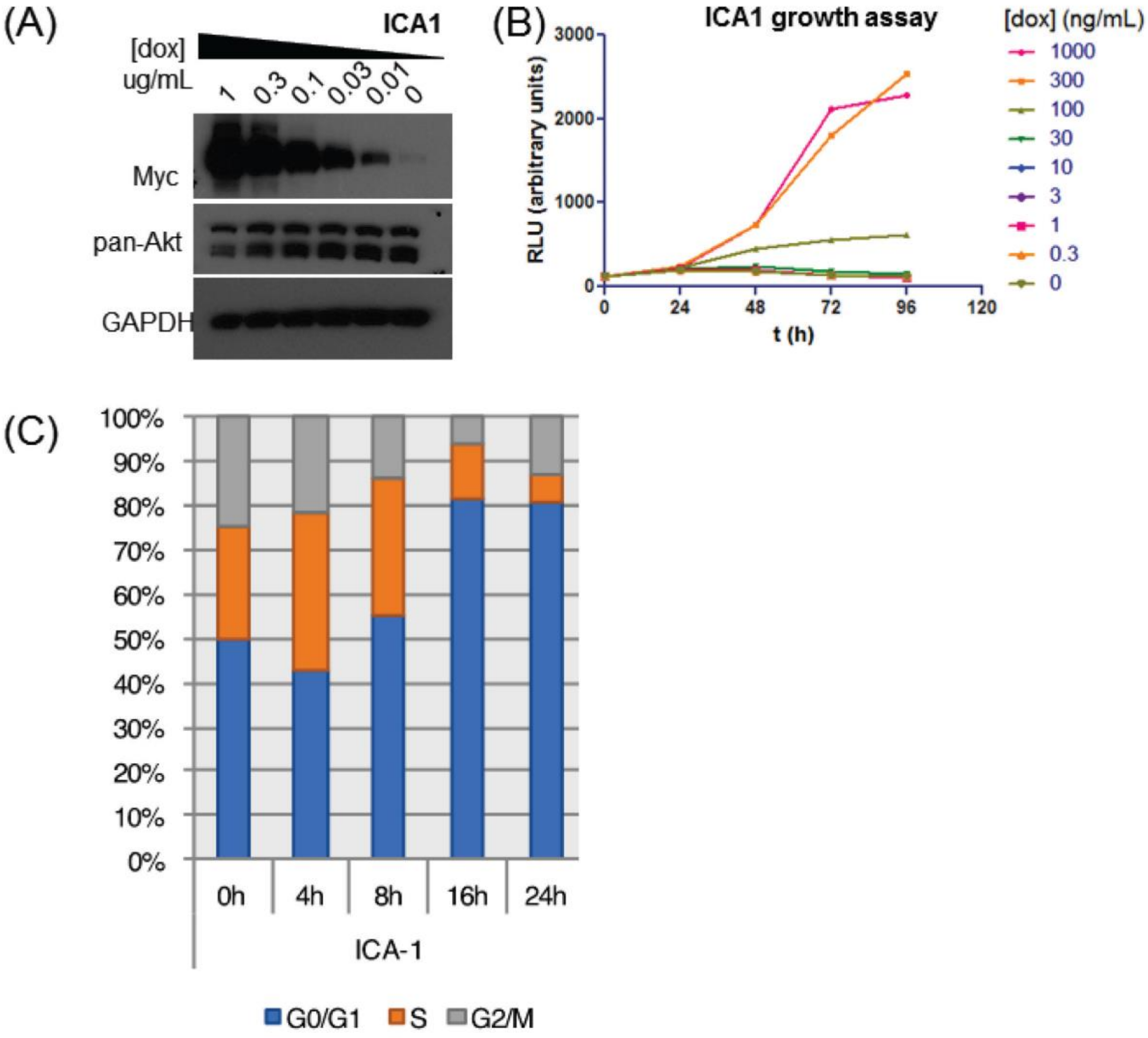
Supplementary Figure 2.9 Validation of Myc signature score and exon conservation across phylogeny and tumor type

(A) Box-and-whisker plot depiction of Myc signature scores in benign prostate tissues and primary prostate cancers stratified by Myc status. Samples with genomic amplifications of the Myc locus or single-gene overexpression are compared to samples without these alterations and adjacent benign tissues. **(B)** Kaplan-Meier disease-free survival plots of prostate cancers stratified by Myc signature score (first panel), Myc amplification status (second panel), or single-gene Myc expression (third panel). **(C)** Unsupervised two-way hierarchical clustering heatmap depiction of exon usage of 1,039 Myc-correlated exons across prostate cancer datasets in healthy tissue, and in primary, metastatic, and neuroendocrine prostate cancers. Columns depict patient samples. The Myc score annotation is colored from white (low) to black (high) based on the rank-transformed signature score of patient samples across the data sets. Rows represent exon inclusion events. Both are ordered by hierarchical clustering. **(D)** UCSC Genome Browser tracks depicting ultraconservation of Myc-regulated exons in SRSF3 (top panel) and HRAS (bottom panel) from humans to lamprey. **(E)** Box-and-whisker plot depiction of the Myc Targets V2 signature scores for breast and lung tissues. Left panel depicts normal breast (GTEx), tumor-adjacent normal breast (TCGA-BRCA), and breast adenocarcinomas (TCGA-BRCA). Right panel depicts normal lung (GTEx), tumor-adjacent normal lung (TCGA-LUAD), and lung adenocarcinomas (TCGA-LUAD). **(F)** Heatmap of Myc-correlated exons in the prostate meta-dataset alongside tissues from normal breast and lung as well as breast and lung adenocarcinomas. Dashed line indicates separation between two cancer types. The Myc score annotation is colored from white (low) to black (high) based on the rank-transformed signature of patient samples across the datasets.



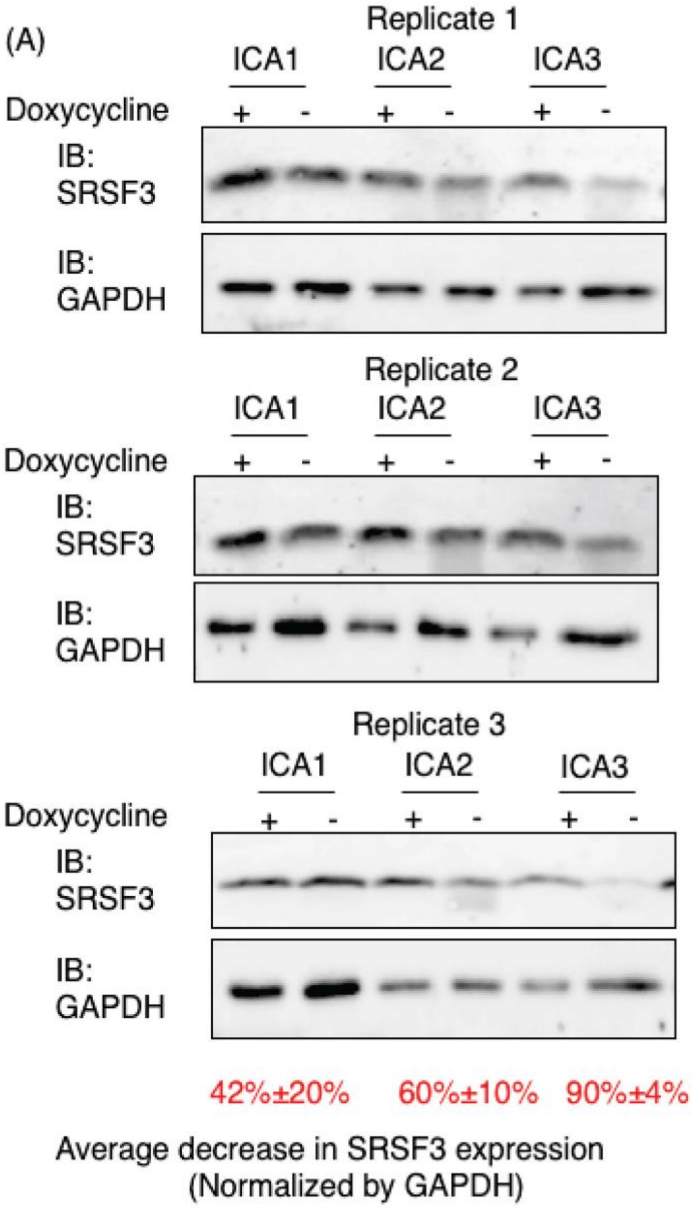
Supplementary Figure 2.10 Establishment of engineered human tumor model with regulated Myc expression

(A) Representative scatterplot from florescence-activated cell sorting isolation of CD49f-high/Trop2-high basal cells from total dissociated benign human prostate. **(B)** Florescent photomicrograph of doubly transduced prostate organoids as well as single and untransduced controls. “UT” = untreated, “C” = c-Myc transduction (GFP), “A” = myrAKT1 (RFP), “CA” = c-Myc and myrAKT1 (merge = yellow). **(C)** Photomicrograph of fixed organoids to show histology. Hematoxylin and eosin staining. **(D)** Immunohistochemical staining of transformed xenograft outgrowth compared to normal prostate tissue controls.



Supplementary Figure 2.11 Characterization of the response to Myc withdrawal in vitro

(A) Immunoblot of Myc expression levels in engineered cell line ICA1 in response to doxycycline titration. Data are representative of all three cell lines**. (B)** Growth response of ICA1 cell line in response to doxycycline titration as measured in a luciferase-based assay. **(C)** Stacked column chart depicting the change in cell cycle distribution over time after doxycycline withdrawal as measured by flow cytometry.



Supplementary Figure 2.12 Individual exon incorporation changes in response to Myc withdrawal

(A) Semi-quantitative immunoblot of SRSF3 protein levels in response to Myc withdrawal for 24 h. Quantitation is the average reduction in SRSF3 levels measured in each cell line over the three independent replicates shown.

2.6 Appendix

2.6.1 Supplementary Methods

Gene Ontology (GO) analysis with background correction for expressed genes

The GO annotation was queried via the EnrichR API in R (97). A customized background gene list is required for the proper calculation of over- and under-representation of a GO term (98). For the alternative splicing analysis in this study, the background genes were selected by having sufficient coverage at splice junctions to meet the filtering criteria described above. With this customized background list, a corrected p-value can be computed using the hypergeometric test. The Benjamini-Hochberg procedure was used to control for the false discovery rate (FDR) at 5%. To reduce complexity, the resulting GO terms were required to contain at least 10 genes, with an exception for **Figure 2.2B**, where the minimum term size was increased to 100 to display the most representative terms. To visualize GO results, the REVIGO web server was employed with customized R plotting scripts for **Figures 2.3 and 2.6** (99).

Overlap enrichment assessment

Hypergeometric test p-value is used to measure the significance of the overlap between two groups of alternative splicing events. The triple intersection p-value is calculated by R package “SuperExactTest” based on hypergeometric test (100).

Breast cancer and lung cancer Myc-correlated alternative splicing analysis

The RNA-Seq processing framework described above was applied to quantify gene expression and alternative splicing of GTEx normal breast and lung samples, and TCGA BRCA and LUAD tumor-adjacent normal samples and tumor samples that are matched to tumor-adjacent normal samples. These datasets are de-identified. The Myc pathway-dependent splicing analysis was performed as described above.

Lentiviral constructs

The myrAKT1 lentiviral vector has been described previously (101). The inducible Myc lentiviral vector was cloned by inserting MYC into the BamHI site of the PSTV lentiviral backbone. Lentiviruses were prepared and titered as described (101).

Organotypic human prostate transformation assay

This assay was conducted as previously described (65,102) with de-identified human prostate samples. Doxycycline (1 ug/mL, Calbiochem 324385) was added to all culture media and renewed every 3 days.

Xenograft outgrowth of transformed cells and cell line derivation

The xenograft and cell line derivation protocols have been previously described and were modified only to accommodate the doxycyline-inducible vector (65,102). Mice were fed sterile doxycycline chow (Bio-Serv S3888) continuously starting 3 days before xenograft implantation. Cell line initiation was performed on harvested tumors with the addition of 1 ug/mL doxycycline to all media.

Cell line exon annotations

Exon annotations of known stop codons and the middle exon length were generated based on the same GENCODE gene annotation file used for alignment. Potential frameshift annotation is determined if the middle exon length cannot be divided by three. Potential RNA binding proteins were labeled according to the GO annotation term ‘RNA binding’.

Cell line propagation

The engineered cell lines were grown in stem cell media, composed of advanced DMEM/F12K (Gibco 12634028) base media with addition of B27 (Gibco 17504044), EGF (10 ng/mL, Peprotech 100-47), and FGF2 (10 ng/mL, Peprotech 100-18B) as well as Glutamax (Gibco 35050061). Doxycycline (1 ug/mL) was added to cultures to maintain MYC expression. Media was renewed every 3 days.

Myc withdrawal experiments

Cells were collected by centrifugation and washed with media three times to remove doxycycline. 1 million cells were plated for each condition. Doxycycline was added back to the appropriate wells and then harvested at the appropriate time point (0-24 h).

Histology

Portions of xenograft outgrowths were fixed in formalin overnight and transferred to 70% ethanol solution before submission for further processing by the Tissue Procurement Core Laboratory at UCLA (TPCL). Organoids were collected by dispase dissociation from Matrigel, washed three times with PBS, and then formalin-fixed for 30 min at room temperature. The fixed organoids were again collected by centrifugation and resuspended in HistoGel and submitted to TPCL. All samples were paraffin-embedded, sectioned at 4 μm, and mounted on glass slides. Hematoxylin and eosin staining was conducted according to standard protocols.

Immunohistochemistry

Immunohistochemical studies were conducted as previously described (102). Briefly, unstained slides were subjected to deparaffinization, rehydration, and heat-activated citric acid antigen retrieval. Rehydrated slides were blocked with 1% horse serum in PBS before overnight incubation with primary antibodies also diluted in 1% horse serum/PBS. Primary and secondary antibodies and their dilutions are listed below. Antibody binding was detected using an HRP-conjugated secondary antibody and a chromogenic substrate.

Immunoblotting

Portions of tumor xenografts or 10 million cultured cells were placed in 8M urea lysis buffer with protease inhibitors (Sigma-Aldrich 4693159001) and homogenized with a Dounce apparatus. The lysate was cleared by ultracentrifugation at 30,000 x g for 30 min.

Samples were denatured by boiling in SDS loading buffer under reducing conditions for 1 min and subjected to polyacrylamide gel electrophoresis. Wet transfer to nitrocellulose membrane was followed by blocking in 1% milk/0.1% Tween/PBS and overnight primary antibody incubation at 5 °C in the same buffer. HRP-conjugated secondary antibodies were applied after washing and the blot visualized with a pro-luminescent substrate. Semi-quantitative blots of SRSF3 protein levels used PVDF membrane. Fluorescence levels were measured by Typhoon scanner and normalized to GAPDH levels. Antibody sources and dilutions are described below.

Antibodies for flow cytometry, immunohistochemistry and immunoblotting

Antibodies used for flow cytometry were the fluorochrome conjugates CD49f-PE (12-0495-82; eBiosciences) and Trop2-APC (FAB650A; R&D Systems).

Primary antibodies used for immunohistochemistry included CK8 (1:1,000, Covance MMS-162P), AR (1:250, Santa Cruz sc-816), PSA (KLK3) (1:2000, Dako A0562), CK5 (1:1000, Covance PRB-160P), and p63 (1:250, Santa Cruz sc-8431). Secondary antibodies used were ImmPRESS anti-rabbit Ig peroxidase and anti-mouse Ig peroxidase (Vector Labs). Liquid DAB+ substrate reagent (Dako) was used to perform direct chromogenic visualization.

The following primary antibodies were used for immunoblotting (all at 1:1000 dilution, unless otherwise noted): Myc (Abcam ab32072), pan-AKT (Cell Signaling 4691), p53 (Cell Signaling Technology 2527), PARP1 (AbCam ab32138), cleaved PARP1 (AbCam, ab32064), anti-Cdk2 (AbCam ab32147), anti-Cdk2 (phospho Y15) (AbCam ab76146), p21 anti-p21 [EPR3993] (ab109199), and GAPDH (1:5,000, GeneTex GT239). HRP-conjugated goat anti-rabbit and goat-anti-mouse secondary antibodies (BioRad) were used for luminescent detection. For semi-quantitative Western blots, goat anti-mouse-cy5 (1:5000, Sigma-Aldrich GEPA45009) was used.

Cell cycle analysis

One million cells were withdrawn from doxycycline as described above and harvested by centrifugation at the appropriate time-point. Cell pellets were washed three times with PBS and then singly dissociated with trypsin prior to fixation in 10% cold ethanol. After overnight fixation at 5° C, cells were pelleted and rehydrated in PBS. RNAse was added and the suspension incubated at room temperature for 4 h before staining with 20 ng/mL 7AAD and analysis by flow cytometry.

Cell growth assay

Cells were washed with PBS, withdrawn from doxycycline, and plated at a density of 100,000 cells per well. Cells were lysed with CellTiterGlo luciferase reagent at the appropriate time and submitted for luminometry.

Whole transcriptome sequencing analysis

Total RNA was isolated by guanidinium thiocyanate-phenol-chloroform extraction, followed by column clean-up. Isolated RNA was submitted for RNA integrity number (RIN) analysis. Only samples with RIN > 9 were carried forward. cDNA libraries were prepared from isolated RNA after poly-A selection using the TruSeq RNA Sample Prep Kit v2 (Illumina). High-throughput sequencing with 150 bp paired-end reads was performed using an Illumina HiSeq 2500. At least 100 million reads were collected for each sample.

Cell line exon annotations

Exon annotations of known stop codons and the middle exon length were generated based on the same GENCODE gene annotation file used for alignment. Potential frameshift annotation is determined if the middle exon length cannot be divided by three. Potential RNA binding proteins were labeled according to the GO annotation term 'RNA binding'.

2.6.2 Supplementary Datasets

Dataset S 2.1

Excel spreadsheet with four tabs. (A) Matrix of pathway correlation scores for skipped exon events in the prostate metadataset. (B) Skipped exon events from prostate tissues with corresponding delta PSI in engineered cell lines alongside annotation for stop codon and NMD prediction. (C) Myc-dependent skipped exon events in engineered cell lines with NMD annotation. (D) Intersection of Myc-correlated skipped exon events in prostate tissues with Myc-dependent events in engineered cell lines with NMD annotation.

2.7 References

1. Baralle, F.E. and Giudice, J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*, **18**, 437-451.

2. Liu, S. and Cheng, C. (2013) Alternative RNA splicing and cancer. *Wiley Interdiscip Rev RNA*, **4**, 547-566.

3. Ho, Y. and Dehm, S.M. (2017) Androgen Receptor Rearrangement and Splicing Variants in Resistance to Endocrine Therapies in Prostate Cancer. *Endocrinology*, **158**, 1533-1542.

4. Catena, R., Muniz-Medina, V., Moralejo, B., Javierre, B., Best, C.J., Emmert-Buck, M.R., Green, J.E., Baker, C.C. and Calvo, A. (2007) Increased expression of VEGF121/VEGF165-189 ratio results in a significant enhancement of human prostate tumor angiogenesis. *Int J Cancer*, **120**, 2096-2109.

5. Narla, G., DiFeo, A., Fernandez, Y., Dhanasekaran, S., Huang, F., Sangodkar, J., Hod, E., Leake, D., Friedman, S.L., Hall, S.J. *et al.* (2008) KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis. *J Clin Invest*, **118**, 2711-2721.

6. Hagen, R.M., Adamo, P., Karamat, S., Oxley, J., Aning, J.J., Gillatt, D., Persad, R., Ladomery, M.R. and Rhodes, A. (2014) Quantitative analysis of ERG expression and its splice isoforms in formalin-fixed, paraffin-embedded prostate cancer samples: association with seminal vesicle invasion and biochemical recurrence. *Am J Clin Pathol*, **142**, 533-540.

7. Mercatante, D.R., Bortner, C.D., Cidlowski, J.A. and Kole, R. (2001) Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells.

analysis of apoptosis and cell death. *J Biol Chem*, **276**, 16411-16417.

8. Antonopoulou, E. and Ladomery, M. (2018) Targeting Splicing in Prostate Cancer. *Int J Mol Sci*, **19**.

9. Arora, K. and Barbieri, C.E. (2018) Molecular Subtypes of Prostate Cancer. *Curr Oncol Rep*, **20**, 58.

10. Beltran, H., Prandi, D., Mosquera, J.M., Benelli, M., Puca, L., Cyrta, J., Marotz, C., Giannopoulou, E., Chakravarthi, B.V., Varambally, S. *et al.* (2016) Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med*, **22**, 298-305.

11. 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. *et al.* (2015) Integrative clinical genomics of advanced prostate cancer. *Cell*, **161**, 1215-1228.

12. Robinson, D.R., Wu, Y.M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V. *et al.* (2017) Integrative clinical genomics of metastatic cancer. *Nature*, **548**, 297-303.

13. Cancer Genome Atlas Research, N. (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**, 1011-1025.

14. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666-677.

15. Jenkins, R.B., Qian, J., Lieber, M.M. and Bostwick, D.G. (1997) Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization. *Cancer Res*, **57**, 524-531.

16. Linja, M.J., Savinainen, K.J., Saramaki, O.R., Tammela, T.L., Vessella, R.L. and Visakorpi, T. (2001) Amplification and overexpression of androgen receptor gene in hormone-refractory prostate cancer. *Cancer Res*, **61**, 3550-3555.

17. Chen, H., Sun, Y., Wu, C., Magyar, C.E., Li, X., Cheng, L., Yao, J.L., Shen, S., Osunkoya, A.O., Liang, C. *et al.* (2012) Pathogenesis of prostatic small cell carcinoma involves the inactivation of the P53 pathway. *Endocr Relat Cancer*, **19**, 321-331.

18. Tran, C., Ouk, S., Clegg, N.J., Chen, Y., Watson, P.A., Arora, V., Wongvipat, J., Smith-Jones, P.M., Yoo, D., Kwon, A. *et al.* (2009) Development of a second-generation antiandrogen for treatment of advanced prostate cancer. *Science*, **324**, 787-790.

19. Mateo, J., Carreira, S., Sandhu, S., Miranda, S., Mossop, H., Perez-Lopez, R., Nava Rodrigues, D., Robinson, D., Omlin, A., Tunariu, N. *et al.* (2015) DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med*, **373**, 1697-1708.

20. Paschalis, A., Sharp, A., Welti, J.C., Neeb, A., Raj, G.V., Luo, J., Plymate, S.R. and de Bono, J.S. (2018) Alternative splicing in prostate cancer. *Nat Rev Clin Oncol*.

21. Thorsen, K., Sorensen, K.D., Brems-Eskildsen, A.S., Modin, C., Gaustadnes, M., Hein, A.M., Kruhoffer, M., Laurberg, S., Borre, M., Wang, K. *et al.* (2008) Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol Cell Proteomics*, **7**, 1214-1224.

22. Ren, S., Peng, Z., Mao, J.H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K.,

Xu, W. *et al.* (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res*, **22**, 806-821.

23. Wang, B.D., Ceniccola, K., Hwang, S., Andrawis, R., Horvath, A., Freedman, J.A., Olender, J., Knapp, S., Ching, T., Garmire, L. *et al.* (2017) Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer. *Nat Commun*, **8**, 15921.

24. Li, H.R., Wang-Rodriguez, J., Nair, T.M., Yeakley, J.M., Kwon, Y.S., Bibikova, M., Zheng, C., Zhou, L., Zhang, K., Downs, T. *et al.* (2006) Two-dimensional transcriptome profiling: identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. *Cancer Res*, **66**, 4079-4088.

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

26. Gan, Y., Li, Y., Long, Z., Lee, A.R., Xie, N., Lovnicki, J.M., Tang, Y., Chen, X., Huang, J. and Dong, X. (2018) Roles of Alternative RNA Splicing of the Bif-1 Gene by SRRM4 During the Development of Treatment-induced Neuroendocrine Prostate Cancer. *EBioMedicine*, **31**, 267-275.

27. Lee, A.R., Li, Y., Xie, N., Gleave, M.E., Cox, M.E., Collins, C.C. and Dong, X. (2017) Alternative RNA splicing of the MEAF6 gene facilitates neuroendocrine prostate cancer progression. *Oncotarget*, **8**, 27966-27975.

28. Li, Y., Donmez, N., Sahinalp, C., Xie, N., Wang, Y., Xue, H., Mo, F., Beltran, H.,

Gleave, M., Wang, Y. *et al.* (2017) SRRM4 Drives Neuroendocrine Transdifferentiation of Prostate Adenocarcinoma Under Androgen Receptor Pathway Inhibition. *Eur Urol*, **71**, 68-78.

29. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, **111**, E5593-5601.

30. Xie, Z. and Xing, Y. (2019) rMATS-turbo. *http://rnaseq-rmats.sourceforge.net/rmats4.0.2*.

31. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.

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

33. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. and Irizarry, R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, **11**, 733-739.

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

35. Shen, S., Wang, Y., Wang, C., Wu, Y.N. and Xing, Y. (2016) SURVIV for survival analysis of mRNA isoform variation. *Nat Commun*, **7**, 11548.

36. Park, E., Pan, Z., Zhang, Z., Lin, L. and Xing, Y. (2018) The Expanding

Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet*, **102**, 11-26.

37. Johnson, N.T., Dhroso, A., Hughes, K.J. and Korkin, D. (2018) Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? *RNA*, **24**, 1119-1132.

38. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.

39. Frank, S., Nelson, P. and Vasioukhin, V. (2018) Recent advances in prostate cancer research: large-scale genomic analyses reveal novel driver mutations and DNA repair defects. *F1000Res*, **7**.

40. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267-273.

41. Qiu, X., Wu, H. and Hu, R. (2013) The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*, **14**, 124.

42. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**, 417-425.

43. Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M. and Butte, A.J. (2017) Comprehensive analysis of normal adjacent

to tumor transcriptomes. *Nat Commun*, **8**, 1077.

44. Setlur, S.R., Mertz, K.D., Hoshida, Y., Demichelis, F., Lupien, M., Perner, S., Sboner, A., Pawitan, Y., Andren, O., Johnson, L.A. *et al.* (2008) Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl Cancer Inst*, **100**, 815-825.

45. Krzywinski, M., Birol, I., Jones, S.J. and Marra, M.A. (2012) Hive plots--rational approach to visualizing networks. *Brief Bioinform*, **13**, 627-644.

46. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, **45**, 1134-1140.

47. Dang, C.V. (2012) MYC on the path to cancer. *Cell*, **149**, 22-35.

48. Gurel, B., Iwata, T., Koh, C.M., Jenkins, R.B., Lan, F., Van Dang, C., Hicks, J.L., Morgan, J., Cornish, T.C., Sutcliffe, S. *et al.* (2008) Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis. *Mod Pathol*, **21**, 1156-1167.

49. Koh, C.M., Bieberich, C.J., Dang, C.V., Nelson, W.G., Yegnasubramanian, S. and De Marzo, A.M. (2010) MYC and Prostate Cancer. *Genes Cancer*, **1**, 617-628.

50. Urbanski, L.M., Leclair, N. and Anczukow, O. (2018) Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA*, **9**, e1476.

51. Cui, M., Allen, M.A., Larsen, A., Macmorris, M., Han, M. and Blumenthal, T. (2008) Genes involved in pre-mRNA 3'-end formation and transcription

termination revealed by a lin-15 operon Muv suppressor screen. *Proc Natl Acad Sci U S A*, **105**, 16665-16670.

52. He, X. and Zhang, P. (2015) Serine/arginine-rich splicing factor 3 (SRSF3) regulates homologous recombination-mediated DNA repair. *Mol Cancer*, **14**, 158.

53. Jia, R., Li, C., McCoy, J.P., Deng, C.X. and Zheng, Z.M. (2010) SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. *Int J Biol Sci*, **6**, 806-826.

54. Corbo, C., Orru, S. and Salvatore, F. (2013) SRp20: an overview of its role in human diseases. *Biochem Biophys Res Commun*, **436**, 1-5.

55. Jumaa, H. and Nielsen, P.J. (1997) The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J*, **16**, 5077-5085.

56. Land, H., Parada, L.F. and Weinberg, R.A. (1983) Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature*, **304**, 596-602.

57. Wang, C., Lisanti, M.P. and Liao, D.J. (2011) Reviewing once more the c-myc and Ras collaboration: converging at the cyclin D1-CDK4 complex and challenging basic concepts of cancer biology. *Cell Cycle*, **10**, 57-67.

58. Cohen, J.B., Broz, S.D. and Levinson, A.D. (1989) Expression of the H-ras proto-oncogene is controlled by alternative splicing. *Cell*, **58**, 461-472.

59. Camats, M., Kokolo, M., Heesom, K.J., Ladomery, M. and Bach-Elias, M. (2009) P19 H-ras induces G1/S phase delay maintaining cells in a reversible quiescence state. *PLoS One*, **4**, e8513.

60. Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J. *et al.* (2016) The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*, **7**, 11479.

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

62. Ji, H., Wu, G., Zhan, X., Nolan, A., Koh, C., De Marzo, A., Doan, H.M., Fan, J., Cheadle, C., Fallahi, M. *et al.* (2011) Cell-type independent MYC target genes reveal a primordial signature involved in biomass accumulation. *PLoS One*, **6**, e26057.

63. Zeller, K.I., Jegga, A.G., Aronow, B.J., O'Donnell, K.A. and Dang, C.V. (2003) An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol*, **4**, R69.

64. Chandriani, S., Frengen, E., Cowling, V.H., Pendergrass, S.A., Perou, C.M., Whitfield, M.L. and Cole, M.D. (2009) A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. *PLoS One*, **4**, e6693.

65. Park, J.W., Lee, J.K., Phillips, J.W., Huang, P., Cheng, D., Huang, J. and Witte, O.N. (2016) Prostate epithelial cell of origin determines cancer differentiation state in an organoid transformation assay. *Proc Natl Acad Sci U S A*, **113**, 4482-4487.

66. Stoyanova, T., Cooper, A.R., Drake, J.M., Liu, X., Armstrong, A.J., Pienta, K.J., Zhang, H., Kohn, D.B., Huang, J., Witte, O.N. *et al.* (2013) Prostate cancer

originating in basal cells progresses to adenocarcinoma propagated by luminal-like cells. *Proc Natl Acad Sci U S A*, **110**, 20111-20116.

67. Bluemn, E.G., Coleman, I.M., Lucas, J.M., Coleman, R.T., Hernandez-Lopez, S., Tharakan, R., Bianchi-Frias, D., Dumpit, R.F., Kaipainen, A., Corella, A.N. *et al.* (2017) Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. *Cancer Cell*, **32**, 474-489 e476.

68. Dani, C., Blanchard, J.M., Piechaczyk, M., El Sabouty, S., Marty, L. and Jeanteur, P. (1984) Extreme instability of myc mRNA in normal and transformed human cells. *Proc Natl Acad Sci U S A*, **81**, 7046-7050.

69. Gartel, A.L., Ye, X., Goufman, E., Shianov, P., Hay, N., Najmabadi, F. and Tyner, A.L. (2001) Myc represses the p21(WAF1/CIP1) promoter and interacts with Sp1/Sp3. *Proc Natl Acad Sci U S A*, **98**, 4510-4515.

70. Demirdjian, L., Shen, S., Wu, Y.N. and Xing, Y. (2019) PAIRADISE: Paired analysis of differential isoform expression. *https://bioconductor.org/packages/release/bioc/html/PAIRADISE.html*.

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

72. Koh, C.M., Bezzi, M., Low, D.H., Ang, W.X., Teo, S.X., Gay, F.P., Al-Haddawi, M., Tan, S.Y., Osato, M., Sabo, A. *et al.* (2015) MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature*, **523**, 96-100.

73. Hsu, T.Y., Simon, L.M., Neill, N.J., Marcotte, R., Sayad, A., Bland, C.S.,

Echeverria, G.V., Sun, T., Kurley, S.J., Tyagi, S. *et al.* (2015) The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature*, **525**, 384-388.

74. Das, S., Anczukow, O., Akerman, M. and Krainer, A.R. (2012) Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell Rep*, **1**, 110-117.

75. Ratnadiwakara, M., Archer, S.K., Dent, C.I., Ruiz De Los Mozos, I., Beilharz, T.H., Knaupp, A.S., Nefzger, C.M., Polo, J.M. and Anko, M.L. (2018) SRSF3 promotes pluripotency through Nanog mRNA export and coordination of the pluripotency gene expression program. *Elife*, **7**.

76. Smith, B.A., Balanis, N.G., Nanjundiah, A., Sheu, K.M., Tsai, B.L., Zhang, Q., Park, J.W., Thompson, M., Huang, J., Witte, O.N. *et al.* (2018) A Human Adult Stem Cell Signature Marks Aggressive Variants across Epithelial Cancers. *Cell Rep*, **24**, 3353-3366 e3355.

77. Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q. and Plath, K. (2009) Role of the murine reprogramming factors in the induction of pluripotency. *Cell*, **136**, 364-377.

78. Nasif, S., Contu, L. and Muhlemann, O. (2018) Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. *Semin Cell Dev Biol*, **75**, 78-87.

79. Zhou, Z. and Fu, X.D. (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, **122**, 191-207.

80. Liu, Y., Beyer, A. and Aebersold, R. (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, **165**, 535-550.

81. Martinez-Montiel, N., Rosas-Murrieta, N.H., Anaya Ruiz, M., Monjaraz-Guzman, E. and Martinez-Contreras, R. (2018) Alternative Splicing as a Target for Cancer Treatment. *Int J Mol Sci*, **19**.

82. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*, **39**, 1181-1186.

83. Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M. *et al.* (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*, **42**, D975-979.

84. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A. and Staudt, L.M. (2016) Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*, **375**, 1109-1112.

85. 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.

86. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760-1774.

87. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562-578.

88. Torgo, L.s. (2017) *Data mining with R : learning with case studies*. Second edition. ed. CRC Press, Taylor & Francis Group, Boca Raton.

89. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.P. (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun*, **9**, 284.

90. Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

91. Mächler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2018) *cluster: Cluster Analysis Basics and Extensions*, R package version 2.0.7-1.

92. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.

93. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

94. Pan, Y., Xing, Y. (2020) Pathway Enrichment-Guided Activity Study of Alternative Splicing (PEGASAS). https://github.com/Xinglab/PEGASAS.

95. Pan, Y., Xing, Y. (2019) Myc-regulated alternative splicing events in aggressive prostate cancers. https://github.com/Xinglab/Myc-regulated_AS_PrCa_paper.

96. Phillips, J.W., et al. (2019) The landscape of alternative splicing in aggressive prostate cancers.

97. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, **44**, W90-97.

98. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587-3595.

99. Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.

100. Wang, M., Zhao, Y. and Zhang, B. (2015) Efficient Test and Visualization of Multi-Set Intersections. *Sci Rep*, **5**, 16923.

101. Xin, L., Lawson, D.A. and Witte, O.N. (2005) The Sca-1 cell surface marker enriches for a prostate-regenerating cell subpopulation that can initiate prostate tumorigenesis. *Proc Natl Acad Sci U S A*, **102**, 6942-6947.

102. Park, J.W., Lee, J.K., Sheu, K.M., Wang, L., Balanis, N.G., Nguyen, K., Smith, B.A., Cheng, C., Tsai, B.L., Cheng, D. *et al.* (2018) Reprogramming normal human epithelial tissues to a common, lethal neuroendocrine cancer lineage. *Science*, **362**, 91-95.

Chapter 3 IRIS: Big data-informed discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing

3.1 Introduction

Cancer immunotherapy has gained tremendous momentum in the past decade. The clinical effectiveness of checkpoint inhibitors, such as neutralizing antibodies against PD-1 and CTLA-4, is thought to result from their ability to reactivate tumor-specific T cells (1). Meanwhile, adoptive cell therapies use genetically modified T-cell receptors (TCRs) or synthetic chimeric antigen receptor T cells (CAR-T) for tumor-specific antigen recognition (2). The finding that cancer cells express specific T-cell-reactive antigens has galvanized epitope discovery in recent years (3-6). Nevertheless, the identification of tumor antigens remains a major challenge (7,8). Although somatic mutation-derived antigens have been successfully targeted by cancer therapies (9-12), this approach remains largely ineffective for tumors with low or moderate mutation loads (7,13).

3.2 Results

3.2.1 IRIS: A big data-powered platform for discovering AS-derived cancer immunotherapy targets

Various types of dysregulation at the RNA level can generate immunogenic peptides in cancer cells (13-15). Notably, tumors harbor up to 30% more alternative splicing (AS) events than normal tissues, and the resulting peptides are predicted to be presented by human leukocyte antigen (HLA) (16). However, there are no integrated methods to systematically identify AS-derived tumor antigens. Therefore, we leveraged tens of thousands of normal and tumor transcriptomes generated by large-scale consortium studies (e.g. GTEx, TCGA) (17,18) to build a versatile, big data-informed platform for discovering AS-derived immunotherapy targets. Our *in silico* platform, named 'IRIS' (Isoform peptides from RNA splicing for Immunotherapy target Screening), incorporates three main components: processing of RNA-Seq data, *in silico* screening of tumor AS isoforms, and integrated prediction and prioritization of TCR and CAR-T targets (**Figure 3.1a**).

IRIS's RNA-Seq data-processing module uses standard input data to discover and quantify AS events in tumors using our ultra-fast rMATS-turbo software (19,20). Identified AS events are fed to the *in silico* screening module, which statistically compares AS events against any combination of samples selected from large-scale (>10,000) reference RNA-Seq samples of normal and tumor tissues (**Supplementary Figure 3.3**) to identify AS events that are tumor-associated, tumor-recurrent, and potentially tumor-specific (**3.4 Methods**). Tumor specificity is a key metric for evaluating potential tissue toxicity, which is an important side effect of targeting lineage-specific antigens that are expressed by both tumor and normal cells (21). In addition to screening multiple patient samples simultaneously in the default ‘group mode’, IRIS can be performed in the ‘personalized mode’ to identify targets for a specific patient sample (**3.4 Methods**). Potential false-positive events are removed by using a blacklist of AS events whose quantification across diverse RNA-Seq datasets is error-prone due to technical variances such as read length (**3.4 Methods** and **Supplementary Figure 3.4**). IRIS’s target prediction module first constructs splice-junction peptides of predicted tumor isoforms and then predicts AS-derived targets for TCR/CAR-T therapies (**3.4 Methods**). This module performs tumor HLA typing using RNA-Seq data and then integrates multiple HLA-binding prediction algorithms for predicting TCR targets and/or peptide vaccines. In parallel, protein extracellular domain annotations are used for predicting CAR-T targets (**Supplementary Figure 3.5**). IRIS also includes the option to confirm predicted AS-derived targets using mass spectrometry (MS) data via proteo-transcriptomics data integration. This option provides an orthogonal approach for target discovery and validation by integrating RNA-Seq data with various types of MS data, such as whole-cell proteomics, surfaceomics, or immunopeptidomics data (**3.4 Methods** and **Supplementary Figure 3.6a**).

3.2.2 Proteo-transcriptomic analysis of HLA presentation of AS-derived epitopes in normal and tumor cell lines

We performed a proof-of-concept analysis and preliminary confirmation of AS-derived epitopes by applying IRIS to RNA-Seq and MS-based immunopeptidomics data of cancer and normal cell lines. We identified hundreds of AS-derived epitopes that were supported by both RNA-Seq and MS data (**Supplementary Figure 3.6b, Supplementary Data 3.1**).

MS-supported epitopes were enriched for transcripts with high expression levels and peptides with strong predicted HLA-binding affinities (**Supplementary Figure 3.6c-e**), consistent with the expected pattern of HLA-epitope binding (22).

3.2.3 Identification of AS-derived cancer immunotherapy targets from 22 GBM samples

To explore IRIS's ability to discover AS-derived immunotherapy targets in clinical samples, we generated RNA-Seq data from 22 resected glioblastomas (GBMs) and analyzed these data by IRIS. Candidate epitopes were then validated based on their recognition by patient T cells. **Figure 3.1b (top)** summarizes the stepwise IRIS results. After uniform processing of RNA-Seq data by rMATS-turbo, IRIS discovered 190,232 putative skipped exon (SE) events from the 22 GBM samples. Using the *in silico* screening module, we compared these AS events against reference normal and tumor panels to evaluate tumor association, recurrence, and specificity (**3.4 Methods**). Specifically, AS events were compared against: normal brain samples from GTEx (tissue-matched normal panel, for evaluating tumor association), two cohorts of brain tumor samples - GBM and lower-grade glioma (LGG) - from TCGA (tumor panel, for evaluating tumor recurrence), and 11 other selected normal (nonbrain) tissues from GTEx (normal panel, for evaluating tumor specificity). After initially screening against the tissue-matched normal panel and removing blacklisted events, IRIS identified 6,276 tumor-associated AS events in the 22 GBM samples ('Primary' set, **Figure 3.1b**). Of these, 1,738 events were identified as tumor-recurrent and tumor-specific based on comparison with the tumor panel and normal panel, respectively ('Prioritized' set, **Figure 3.1b; Supplementary Data 3.2**).

Next, for each AS event, splice junctions of the tumor isoform (i.e. the isoform that was more abundant in the tumor samples than in the tissue-matched normal panel) were translated into peptides, followed by TCR/CAR-T target prediction (**Figure 3.1b**). For the GBM dataset, IRIS predicted 4,153 ‘primary’ tumor-associated epitope-producing splice junctions. Of these, 1,127 were tumor-recurrent and tumor-specific compared to the tumor panel and normal panel, respectively, and were therefore predicted to be ‘prioritized’ TCR targets. In parallel, IRIS identified 416 ‘primary’ tumor-associated extracellular peptide-producing splice junctions, of which 87 were predicted to be ‘prioritized’ CAR-T targets.

IRIS generates an integrative report for predicted immunotherapy targets (**Supplementary Data 3.3**). Representative examples for six prioritized TCR targets are shown in the bottom panel of **Figure 3.1b** (see **Supplementary Figure 3.5b** for prioritized CAR-T target examples). Violin plots depict exon inclusion levels across the 22 GBM samples (‘GBM-input’) and different sets of reference panels using the percent-spliced-in (PSI) metric (23). Tumor isoforms can be either the exon-skipped (low PSI) or the exon-included (high PSI) isoform compared to the tissue-matched normal panel. As illustrated by the darker dots in the ‘Summary’ column, all six epitope-producing splice junctions were tumor-associated compared to the tissue-matched normal panel (‘Brain’), and tumor-recurrent compared to the tumor panel (‘GBM’ and ‘LGG’). Two AS events (in *TRIM11* and *FAM76B*) consistently showed distinct PSI values in tumors compared to normal brain and nonbrain tissues, indicating high tumor specificity. For candidate splice junctions, IRIS also calculates the fold-change (FC) of tumor isoforms between tumor samples and the tissue-matched normal panel (**3.4 Methods**). For example, the tumor isoform in *TRIM11* had an average isoform proportion of 8.60% in the 22 GBM samples and 0.13% in normal brain samples, representing an FC of 65.6 in tumor samples versus the tissue-matched normal panel. We should note that, as shown under ‘Predicted HLA-epitope binding’, a single splice junction can give rise to multiple putative epitopes with distinct peptide sequences and HLA binding affinities.

3.2.4 IRIS-predicted AS-derived TCR targets recognized by CD3+CD8+ T cells in tumors and peripheral blood from patients

Finally, we sought to validate the immunogenicity and T-cell recognition of IRIS-identified candidate TCR targets using an MHC class I dextramer-based assay(12,24). We focused on predicted AS-derived tumor epitopes with strong putative HLA-binding affinity to common HLA types found in at least five of the 22 patients. We selected seven AS-derived tumor-associated epitopes (five HLA-A02:01 and two HLA-A03:01) for dextramer-based T-cell recognition testing (**Supplementary Data 3.4**). All but one epitope (last one in the table) showed some degree of tumor specificity when evaluated in normal (nonbrain) tissues (‘vs. Normal’, see **Figure 3.2a**). We obtained customized HLA-matched, fluorescently labeled MHC class I dextramer:peptide (pMHC) complexes for each candidate epitope. We conducted flow cytometry to detect CD8$_{+}$ T-cell binding with the pMHC complexes using available peripheral blood mononuclear cells (PBMCs) and/or *ex vivo*-expanded tumor-infiltrating lymphocytes (TILs). Based on the binding of each AS-derived tumor epitope to a patient’s CD3$_{+}$CD8$_{+}$ T cells, we classified epitope reactivity as ‘positive’ (binding > 0.1% of cells), ‘marginal’ (binding 0.01-0.1% of cells), or ‘negative’ (binding < 0.01% of cells). Epitopes that showed at least marginal reactivity were considered to be ‘recognized’ by patient T cells. We analyzed samples from two HLA-

A02:01 and four HLA-A03:01 patients, as well as samples from three HLA-A02:01 and three HLA-A03:01 healthy donors (**Supplementary Data 3.5**).

Both predicted HLA-A03:01 tumor epitopes were recognized by patient T cells. In particular, one epitope (KIGRLVTRK, in *PLA2G6*) was recognized by T cells from all four tested patients but only one of the three tested healthy donors. In one patient (LB2867), recognition of tumor epitope KIGRLVTRK was marginal in PBMCs but positive in the expanded TIL population, with epitope-reactive T cells representing 0.03% of T cells in PBMCs and 1.69% of T cells in TILs. This patient had been previously treated with neoadjuvant anti-PD-1 and anti-CTLA-4 checkpoint blockade immunotherapy. These results suggest epitope KIGRLVTRK as a promising immunotherapy target in HLA-A03 patients from our GBM cohort. T cells from another patient (LB2907) showed positive reactivity to both tested HLA-A03:01 epitopes. All four predicted HLA-A02:01 epitopes were recognized by T cells from tested patients and healthy donors. The non-tumor-specific epitope (YAIVWVNGV, bottom row in **Figure 3.2a**) was tested in two patients and three healthy donors and was recognized by T cells in only one healthy donor (marginal reactivity, 0.013% of $CD3_+CD8_+$ T cells). Taken together, our dextramer-based assay results indicate that the AS-derived TCR targets predicted by IRIS can be recognized by tumor-infiltrating and peripheral $CD3_+CD8_+$ T cells.

Dextramer-positive T cells are expected to contain many clonotypes, only a few of which are dominant. To discover and quantify which TCR clonotypes comprise the epitope-reactive T cells, we sorted the TILs from one patient (LB2867) for cells that reacted positively with the KIGRLVTRK pMHC complex (**Figure 3.2b**), and performed V(D)J immune profiling using single-cell RNA-Seq (scRNA-Seq) on the sorted population

(**Figure 3.2c**). Of the 325 unique TCR clonotypes, the 10 most abundant TCRs represented 86.3% of all clonotypes (**Supplementary Data 3.6**), with the most frequent clonotype comprising 38.9% of all epitope-reactive T cells. This result suggests that there was clonal expansion of a select few dominant TCR clones within the tumor that were able to recognize the AS-derived epitope. To further validate our findings using complementary approaches, we analyzed bulk expanded TILs using immunoSEQ and pairSEQ assays (**Figure 3.2c, Supplementary Figure 3.7**). We confirmed that the top 10 reported clonotypes from scRNA-Seq were present in the bulk TIL population based on the TCR β-chain CDR3 region. In addition, the pairSEQ assay, which uses statistical modeling to predict pairing of TCR α and β chains, found identically paired TCRs for seven of the top 10 TCRs from scRNA-Seq. Together, these data suggest that a select few TCR clones dominantly recognize the AS-derived epitope KIGRLVTRK in this patient.

3.2.5 Discover diverse forms of AS-derived tumor antigens from 22 GBM samples using upgraded IRIS platform

Since the last release, IRIS has been updated with improved reference databases and additional functionalities. In addition to existed tumor reference for GBM and LGG, 14 more major TCGA cancer types are included to the reference database using the uniform RNA-seq processing pipeline provided by IRIS for generating reference database (**3.4 Methods**). Collectively, splicing pattern of 30 normal tissue types and 16 tumor types based on more than 17,000 transcriptomes were summarized in IRIS reference database (Supplementary Table 3.1-3.2). The global landscape of AS in normal tissues and tumors are profiled (Supplementary Figure 3.8). An unsupervised representation showed PSI value based AS quantification can capture differences between tissues or tumors, with brain and brain tumors standing out from other tissue and tumors. This underlies the importance of utilizing AS pattern references to determine the frequency and specificity of a splicing events in cancer.

As shown in the Supplementary Figure 3.9, stepwise results of updated IRIS to identify AS-derived cancer immunotherapy targets from 22 GBM samples are summarized. In addition to skipped-exon (SE) events, 5,919 alternative 5’ splice site (A5SS), 8,619 alternative 3’ splice site (A3SS), and 5,285 retained intron (RI) events were identified with proper filtering by IRIS (**3.4 Methods**), including both annotated and unannotated AS events in the genome annotation file. Using the *in silico* screening module, we compared these AS events against reference normal and tumor panels to evaluate tumor association, recurrence, and specificity (**3.4 Methods**). Specifically, AS events were compared against: normal brain samples from GTEx (tissue-matched normal panel, for evaluating tumor association), two cohorts of brain tumor samples - GBM and lower-grade glioma (LGG) - from TCGA (tumor panel, for evaluating tumor recurrence), and 11 other selected normal (nonbrain) tissues from GTEx (normal panel, for evaluating tumor specificity). After initially screening against the tissue-matched normal panel and removing blacklisted events, IRIS identified 9,945 tumor-associated SE events, 919 A5SS events, 1,384 A3SS events and 1,780 RI events in the 22 GBM samples (‘Primary’ set, **Figure 3.1b**). Of these, 1,184 SE events, 273 A5SS events, 493 A3SS events and 1,014 RI events were identified as tumor-recurrent and tumor-specific based on comparison with the tumor panel and normal panel, respectively (‘Prioritized’ set, **Figure 3.1b; Supplementary Data 3.2**).

Followed by inference based on normal and tumor reference, two translation strategies were used: translating based on known ORFs in proteome database or using all three ORFs. For each AS event, splice junctions of the tumor isoform (i.e. the isoform that was more abundant in the tumor samples than in the tissue-matched normal panel) were translated into peptides.

IRIS predicted 6,140 SE-derived, 265 A5SS-derived, 953 A3SS-derived and 1,281 RI-derived 'primary' tumor-associated epitope-producing splice junctions. Of these, 713 SE-derived, 80 A5SS-derived, 334 A3SS-derived and 784 RI-derived epitope-producing splice junctions were tumor-recurrent and tumor-specific compared to the tumor panel and normal panel, respectively, and were therefore predicted to be 'prioritized' TCR targets. In parallel, IRIS identified 655 SE-derived, 39 A5SS-derived, 84 A3SS-derived and 107 RI-derived 'primary' tumor-associated extracellular peptide-producing splice junctions, of which 60 SE-derived, 14 A5SS-derived, 22 A3SS-derived and 60 RI-derived junctions were predicted to be 'prioritized' CAR-T targets.

3.2.6 An independent cohort of 53 GBM samples were sequenced and analyzed by IRIS to cross-validate discovered tumor antigens

To evaluate IRIS-predicted targets from 22 GBM samples, we replicate the analysis using an independent cohort of 53 GBM samples (Supplementary Figure 3.10). RNA-seq data from 53 GBM samples were sequenced and processed through the same uniform pipeline to serve as a validation cohort to evaluate targets discovered by the discovery cohort of 22 GBM samples. Subjected IRIS to analyze this validation cohort, we identified 1,187 SE-derived epitope-producing splice junctions as 'Prioritized' TCR targets. Due to the difference of HLA types in patients from the two cohorts, the number of TCR targets are not directly comparable. Aiming to search for generic tumor antigens for TCR therapies, we focused on common HLA types (HLA*01:01, HLA*02:01, HLA*03:01). This reduced number of SE-derived epitope-producing splice junctions to 249 for the discovery cohort and 439 for the validation cohort. A shared set of 219 SE-derived epitope-producing splice junctions were identified for further selection. Next, important features for tumor targets like tumor-specificity were leveraged along with expression level of the gene of AS events, HLA types and HLA predicted binding affinities to rank the 219 shared TCR target candidates by the two cohorts. As shown in Supplementary Figure 3.10 b, top ten candidates are selected for dextramer-based T-cell recognition assay. All ten selected candidates are predicted binding to HLA*02:01 in order to minimize the variance introduced by the HLA type of the dextramers. This study is still ongoing and the result of this experiment will help to evaluate IRIS-predicted TCR targets and inform the design of future target discovery procedures.

3.3 Discussion

In summary, we have developed IRIS, a big data-powered platform for discovering AS-derived tumor antigens as an underexploited source of immunotherapy targets. Using IRIS followed by a dextramer-based assay, we discovered and validated AS-derived tumor epitopes recognized by T cells in patients. Our results provide experimental evidence for the immunogenicity of tumor antigens arising from AS and reveal novel potential targets for TCR and CAR-T therapies. The IRIS software can be downloaded from https://github.com/Xinglab/IRIS.

3.4 Methods

3.4.1 IRIS module for RNA-Seq data processing

IRIS accepts standard formats of raw RNA-Seq FASTQ files and/or tab-delimited files of quantified AS events (from rMATS-turbo) as input data (**Figure 3.1a**). For raw RNA-Seq data, IRIS provides a standalone pipeline that aligns RNA-Seq reads to the reference human genome hg19 using the STAR 2.5.3a (25) two-pass mode, followed by Cufflinks v2.2.1 (26) and rMATS v4.0.2 (rMATS-turbo) (19,20) for quantification of gene expression and AS events, respectively, based on the GENCODE (V26) (27) gene annotation. To quantify AS events, we converted splice-junction counts in rMATS-turbo output into PSI (23) values. For each dataset, we removed low-coverage AS events, defined as events with an average count of less than 10 reads for the sum of all splice junctions across all samples in that dataset (tissue/tumor type). We applied this procedure to the 22 GBM samples from the UCLA cohort (BioProject: PRJNA577155), as well as to the normal and tumor samples of the reference panels used by IRIS. For the GTEx normal samples, aligned BAM files downloaded from the dbGAP repository were used directly for AS quantification.

3.4.2 Constructing big-data reference panels of AS events across normal human tissues and tumor samples

IRIS's big-data reference panels of normal and tumor samples are available as pre-processed, pre-indexed databases for fast retrieval by the IRIS program (**Supplementary Figure 3.3**). Specifically, 9,662 normal samples from the GTEx project (V7) (17) representing 53 tissue types of 30 histological sites were uniformly processed as described above. As shown in **Supplementary Figure 3.3a-b**, exon-based quantification of AS events was able to distinguish samples by tissue type. Selected TCGA (16,28) tumor samples (**Supplementary Figure 3.3c**) were processed similarly to form the tumor panel. Additionally, IRIS provides a stand-alone indexing function for users to include custom normal and tumor samples in their reference panels.

3.4.3 IRIS module for *in silico* screening of tumor AS events

IRIS performs *in silico* screening using two-sided and one-sided *t*-tests to identify tumor-associated, tumor-recurrent, and tumor-specific AS events in group comparisons. To define an AS event as significantly different from a reference group (i.e., to identify tumor-associated/tumor-specific events), IRIS sets two requirements: 1) a significant p-value from the two-sided *t*-test (default: $p < 0.01$), and 2) a threshold of PSI value difference (default: abs(Δψ) > 0.05). With a minor modification, to define an AS event as tumor-recurrent, IRIS compares a reference tumor panel with the tissue-matched normal panel and requires: 1) a significant p-value from the one-sided *t*-test in the same direction as the corresponding 'tumor-associated' event (default: p < 0.01/number of 'tumor-associated' events; Bonferroni correction applied due to large sample sizes in reference panels), and 2) a threshold of PSI value difference (default: abs(Δψ) > 0.05). In addition, as the normal or tumor reference panel may include multiple individual groups (e.g. tissue types), a threshold of the number of significant comparisons against groups in the normal or tumor reference panel is used to determine whether AS-derived antigens are tumor-specific or tumor-recurrent. For each AS event, IRIS defines the 'tumor isoform' as the isoform that is more abundant in tumors than in the tissue-matched normal panel. Optionally, to rank or filter targets, IRIS estimates the ‘fold-change (FC) of tumor isoform’ as the FC of the tumor isoform’s proportion in tumors compared to the tissue-matched normal panel. In addition to the default ‘group mode’, IRIS can be used to screen targets for a specific patient sample through the ‘personalized mode’. This mode uses an outlier detection approach, combining a modified Tukey’s rule (29) and a user-defined threshold of PSI value difference.

3.4.4 Identification of AS events that are prone to measurement errors due to technical variances across big-data reference panels

IRIS’s big-data reference panels were constructed by integrating various large-scale datasets with distinct technical conditions, such as RNA-Seq read length (30). Such technical variances across datasets could introduce discrepancies in the quantification of AS events (30). To identify error-prone AS events, we employed a data-based heuristic strategy to assess the effects of RNA-Seq read length (48 bp vs. 76 bp) and aligner (STAR vs. Tophat) on AS quantification (PSI value) (**Supplementary Figure 3.4a**). For a given tissue type (in this study, brain tissue), 10 randomly selected 76-bp RNA-Seq files from GTEx were artificially trimmed to 48 bp, and both 76- and 48-bp RNA-Seq files were aligned with STAR2.5.3a. Corresponding Tophat (v.1.4.1)-aligned 76-bp BAM files were directly downloaded from GTEx. AS events were quantified by rMATS-turbo. Events with significantly different PSI values ($p < 0.05$, abs($\Delta\psi$) > 0.05 from paired *t*-test) among RNA-Seq datasets with distinct technical conditions were included in a blacklist. Results of this analysis for GTEx normal brain samples are shown in **Supplementary Figure 3.4b**.

3.4.5 IRIS module for predicting AS-derived TCR and CAR-T targets

To obtain protein sequences of AS-derived tumor isoforms, IRIS generates peptides by translating splice-junction sequences into amino-acid sequences using known ORFs from the UniProtKB (31) database. Within each AS event, the splice-junction peptide sequence for the tumor isoform is compared to that of the alternative normal isoform, to ensure that the tumor isoform splice junction produces a distinct peptide.

For TCR target prediction, IRIS employs seq2HLA (32), which uses RNA-Seq data to characterize HLA class I alleles for each tumor sample. IRIS then uses IEDB API (33) predictors to obtain the putative HLA binding affinities of candidate epitopes. The IEDB 'recommended' mode runs several prediction tools to generate multiple predictions of binding affinity, which IRIS summarizes as a median IC_{50} value. By default, a threshold of median (IC_{50}) < 500 nM denotes a positive prediction for an AS-derived TCR target.

For CAR-T target prediction, IRIS maps AS-derived tumor isoforms to known protein extracellular domains (ECDs), as potential candidates for CAR-T therapy (**Supplementary Figure 3.5a**). Specifically, IRIS generates pre-computed annotations of protein ECDs. First, protein cellular localization information was retrieved from the UniProtKB (31) database (flat file downloaded in April 2018). ECD information was retrieved by searching for the term 'extracellular' in topological annotation fields, including 'TOPO_DOM', 'TRANSMEM', and 'REGION', in the flat file. Second, BLAST (34) was used to map individual exons in the gene annotation (GENCODE V26) to proteins with topological annotations. Third, the BLAST result was parsed to create annotations of the mapping between exons and ECDs in proteins. These pre-computed annotations are queried to search for AS-derived peptides that can be mapped to protein ECDs as potential CAR-T targets.

3.4.6 Proteo-transcriptomics data integration for MS validation

IRIS includes an optional proteo-transcriptomics data integration function that incorporates various types of MS data, such as whole-cell proteomics, surfaceomics, or immunopeptidomics data, to validate RNA-Seq-based target discovery at the protein level (**Supplementary Figure 3.6a**). Specifically, sequences of AS-derived peptides are added to canonical and isoform sequences of the reference human proteome (downloaded from UniProtKB in September 2018). For immunopeptidomics data, fragment MS spectra are searched against the RNA-Seq-based custom proteome library with no enzyme specificity using MSGF+(35). The search length is limited to 7-15 amino acids. The target-decoy approach is employed to control the false discovery rate (FDR) or ‘QValue’ at 5%.

3.4.7 IRIS analysis of immunopeptidomics data

Published matching RNA-Seq and MS immunopeptidomics data of B-LCL-S1 and B-LCL-S2 cell lines (B lymphoblastoid cell lines from two individual donors) were retrieved from Laumont *et al*.(36) (GEO: GSM1641206, GSM1641207, and PRIDE: PXD001898). Raw RNA-Seq data of the JeKo-1 lymphoma cell line were obtained from the Cancer Cell Line Encyclopedia via the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/legacy-archive/). Corresponding immunopeptidomics MS data of JeKo-1 were retrieved from

Khodadoust *et al*. (37) (PRIDE: PXD004746).

RNA-Seq data of the normal (B-LCL-S1, B-LCL-S2) and cancer (JeKo-1) cell lines were analyzed by IRIS as described above, with minor modifications. Specifically, AS events identified by the IRIS RNA-Seq data processing module were not subjected to the *in silico* screening module, but instead were directly used for the MS search. For MSGF+, FDR was set at 5%, which had the best concordance with predicted binding affinities (**Supplementary Figure 3.6c-d**). For comparison of predicted HLA binding and nonbinding peptides (**Supplementary Figure 3.6d**), a set of nonbinding peptides was created by randomly selecting peptides with median(IC_{50}) > 500 nM to the same number of binding peptides (median(IC_{50}) < 500 nM).

3.4.8 IRIS discovery of candidate TCR and CAR-T targets from 22 GBM samples

RNA-Seq samples were processed by IRIS. Detected skipped exon (SE) events were analyzed by using the IRIS screening and target prediction modules with the aforementioned default parameters. For reference panels, the ‘tissue-matched normal panel’ comprised normal brain tissue samples from GTEx; the ‘normal panel’ comprised other normal (nonbrain) tissue samples of 11 selected vital tissues (heart, skin, blood, lung, liver, nerve, muscle, spleen, thyroid, kidney and stomach) from GTEx; and the ‘tumor panel’ comprised two cohorts of brain tumor samples (GBM and LGG) from TCGA. The blacklist of AS events created for brain was applied before *in silico* screening by IRIS to eliminate error-prone AS events (**Supplementary Figure 3.4**).

In screening for the ‘Primary’ set of AS events, we considered an event to be ‘tumor-associated' if it was significantly different from the tissue-matched normal panel, using the default criteria described in 'IRIS module for *in silico* screening of tumor AS events'. In screening for the 'Prioritized' set, we prioritized an AS event if it was both 'tumor-recurrent' (significantly different from the tissue-matched normal panel, in the same direction as input GBM samples, in at least 1 of 2 groups in the GBM/LGG tumor panel) and 'tumor-specific' (significantly different from multiple of 11 groups in the normal panel in the same direction as the tissue-matched normal panel). Here, we used at least 2 groups but this threshold can be user-defined to allow for higher stringency.

When selecting potential TCR targets for dextramer validation, we applied three additional criteria: 1) predicted median(IC_{50}) ≤ 300 nM; 2) predicted binding to common HLA types, including HLA-A02:01 and HLA-A03:01; and 3) predicted binding to at least five patients in the GBM cohort. After excluding targets with low gene expression (average FPKM < 5), we selected seven epitopes to test for T-cell recognition by dextramer assays.

3.4.9 Patients.

Tumor specimens were collected from 22 consenting patients with GBM who underwent surgical resection for tumor removal at the University of California, Los Angeles (UCLA; Los Angeles, CA). From these patients, we also obtained PBMCs and TILs from two HLA-A02:01+ and four HLA-A03:01+ patients. All patients provided written informed consent, and this study was conducted in accordance with established Institutional Review Board-approved protocols.

3.4.10 PBMC collection.

Peripheral blood was drawn from patients before surgery and diluted 1:1 in RPMI media (Thermo Fisher Scientific, cat. no. MT10041CV). PBMCs, extracted by Ficoll gradient (Thermo Fisher Scientific, cat. no. 45-001-750), were washed twice in RPMI media. Collected PBMCs were frozen in 90% human AB serum (Thermo Fisher Scientific, cat. no. MT35060CI) and 10% DMSO (Sigma, cat. no. C6295-50ML) and stored in liquid nitrogen. In parallel, PBMCs from healthy HLA-A02:01 and HLA-A03:01 donors were purchased from Bloodworks Northwest (Seattle, WA) or Astarte Biologics (Bothell, WA).

3.4.11 TIL collection.

Surgically resected tumor samples were digested with a brain tumor dissociation kit (Miltenyi Biotec, cat. no. 130-095-42) and gentle MACS dissociator (cat. no. 130-093-235). After digestion and myelin depletion, collected cells were labeled with CD45 microbeads (cat. no. 130-045-801) and separated on Miltenyi LS columns (cat. no. 130-042-401) and MidiMACS Separator (cat no. 130-042-302). Collected CD45$^+$ cells were cultured at 1×10^6 cells/mL in X-VIVO 15 Media (Fisher Scientific, cat. no. BW04-418Q) containing 2% human AB serum with 50 ng/mL anti-CD3 antibody (BioLegend, cat. no. 317304), 1 µg/mL anti-CD28 antibody (BD Biosciences, cat. no. 555725), 1 µg/mL anti-CD49d antibody (BD Biosciences, cat. no. 555501), 300 IU/mL IL-2 (NIH, cat. no. 11697), and 10 ng/mL IL-15 (BioLegend, cat. no. 570302). Cells were expanded for 3-4 weeks and replenished with fresh media and cytokines every 2-3 days. Before freezing, expanded cells were placed in media containing 50 IU/mL IL-2 for 1-2 days and then frozen in the same freezing media as PBMCs.

3.4.12 Collection of tumor RNAs and RNA sequencing.

RNA from freshly collected or flash-frozen tumor specimens was extracted by using the RNeasy Mini Kit (Qiagen, cat. no. 74014). Paired-end RNA-Seq was performed at the UCLA Clinical Microarray Core using an Illumina HiSeq 3000 at a read length of 2×100 bp or 2×150 bp.

3.4.13 Dextramer flow-cytometric analysis of PBMCs and TILs

For each AS-derived peptide selected for validation, custom-made HLA-matched MHC Class I dextramer:peptide (pMHC) complexes were purchased from Immudex (Copenhagen, Denmark). Immudex also provided pMHC complexes for common cytomegalovirus (CMV) epitopes (cat. nos. WB2132 and WC2197) and for a nonhuman epitope (NI3233) as a negative control. Each pMHC complex was purchased with two separate tags for APC or PE fluorescence labeling, to increase specificity to targeted T cells with dual labeling.

To facilitate proper gating of CD8$_+$ T cells from PBMC and TIL populations, the following panel of antibodies (from BioLegend) was set up: CD3 BV605 (cat. no. 300460), CD8 FITC (cat. no. 344704), CD4 BV421 (cat. no. 317434), CD19 BV421 (cat. no. 302234), CD56 BV421 (cat. no. 362552), and CD14 BV421 (cat. no. 301828). For single-color compensation controls, OneComp eBeads were used (Thermo Fisher Scientific, cat. no. 01-1111-41).

For each set of pMHC complexes, at least 3×10$_6$ cells were stained according to manufacturer’s guidelines. Briefly, cells were thawed in a 37°C water bath and washed with RPMI and D-PBS (Fisher Scientific, cat. no. MT21031CV) before staining for cell viability with the Zombie Violet Viability Kit (BioLegend, cat. no. 423113). Next, the appropriate amount of each pMHC complex in a staining buffer of D-PBS with 5% fetal bovine serum (Fisher Scientific, cat. no. MT35016CV) was added to each sample. After 10 min, the aforementioned antibody cocktail was added. After a 30-min incubation period, cells were washed twice in the same staining buffer. All samples were tested in a BD LSRII flow cytometer, and data were analyzed with FlowJo (Treestar). For gating, the lymphocyte population was first selected using forward and side scatter, and then the BV421-negative population was gated out (i.e. excluding dead cells and the CD14, CD19, CD56, and CD4 populations) before selecting the $CD3_+CD8_+$ population. To set for proper gating of dextramer-positive cells, we used cells that were stained with the full antibody panel but no pMHC complexes, and cells that were given the nonhuman pMHC complex.

3.4.14 TCR sequencing using scRNA-Seq

Cells were stained by following the dextramer procedure with PE-conjugated pMHC complexes only. Cells were sorted by using the BD FACSAria flow cytometer, and PE_+ cells were collected. V(D)J immune profiling of sorted cells was done with scRNA-Seq, using the 10X Genomics Chromium Single Cell Immune Profiling Workflow at the UCLA Clinical Microarray Core. Each T cell was encapsulated in an oil emulsion droplet with a barcoded gel bead, and reverse transcription was performed to create a barcoded cDNA library. The V(D)J-enriched and gene expression libraries were sequenced using the 10X Genomics Chromium Controller. After sequencing, the Cell Ranger pipeline was used to align reads, filter, count barcodes and assign unique molecular identifiers.

3.4.15 Next-generation immune repertoire sequencing using the immunoSEQ platform

To assess the T-lymphocyte repertoire of bulk expanded TIL populations, we used the immunoSEQ assay (Adaptive Biotechnologies). This multiplex PCR system uses a mixture of primers that target the rearranged V and J segments of the CDR3 region to assess TCR diversity within a given sample. Genomic DNA from each sample was extracted by using the QIAamp DNA Blood Midi Kit (Qiagen, cat. no. 51185). We provided at least 1 µg of DNA (~60,000 cells) from each sample to Adaptive Biotechnologies for sequencing at a deep resolution. Resulting sequencing data were analyzed with the immunoSEQ Analyzer Platform (Adaptive Biotechnologies).

3.4.16 High-throughput αβ TCR pairing using the pairSEQ platform.

We provided Adaptive Biotechnologies with frozen bulk expanded TIL samples for their pairSEQ assay, to predict which α and β chains may pair to form a functional TCR. Briefly, T cells were randomly distributed into wells of a 96-well plate. The mRNA was extracted, converted to cDNA, and amplified by using TCR-specific primers. The cDNA of T cells from each well was given a specific barcode, and all wells were pooled together for sequencing. Each TCR sequence was mapped back to the original well through computational demultiplexing. Putative TCR pairs were identified by examining whether a sequenced TCR α chain was frequently seen to share the same well with a specific sequenced TCR β chain, above statistical noise.

3.4.17 Updated IRIS module for RNA-seq data processing

The snakemake-based framework is used to build IRIS. IRIS offers pipelines to perform streamlined RNA-seq processing, HLA inference, target screening and prediction.

The updated IRIS RNA-seq processing module offers a new option allowing filtering AS events based on junction read coverage by individual samples. For each sample in a dataset, we removed low-coverage AS events, defined as events with splice junction reads less than 10. In comparison to the existing group/dataset-based average read coverage filtering, this new option gives flexibility of examining events with selective expression in a subset of dataset, which may be removed by group-based filtering.

We applied this procedure to normal and tumor samples of the reference panels used by IRIS, increasing the number of AS events in the reference. The 22 GBM samples from the UCLA cohort (BioProject: PRJNA577155) remain as the previous group filtering mode described in the above section to the small sample size.

FASTQ files of GTEx V7 RNA-seq samples and TCGA tumor samples were downloaded and uniformly processed and quantified by STAR aligner, Cufflinks and rMATS-turbo (4.0.2) as described above.

For PCA, IRIS reference database is further filtered by group mode, plus a PSI range greater than 5% across the entire reference dataset, with a mean skipping or inclusion value over 5% missing value less than 5%.

3.4.18 Updated IRIS AS reference database

The updated IRIS's big-data reference panels of normal and tumor samples are available as pre-processed, pre-indexed databases for fast retrieval by the IRIS program (**Supplementary Table 3.1-3.2**). Specifically, 9,561 normal tissue samples from the GTEx project (V7) (17) representing 30 histological sites and additional three cell lines and 7,900 tumor samples from TCGA(18) (major tumor types with tumor sample n > 300) were uniformly processed as described above to form normal and tumor panels. Additionally, IRIS provides a stand-alone indexing and formatting function for users to include custom normal and tumor samples in their reference panels.

3.4.19 Additional options for IRIS module for *in silico* screening of tumor AS events

In addition to existing options, the IRIS screening module now supports both parametric and non-parametric tests in order to account for outliers in AS pattern in both input data and reference panels. AS events unique in input samples will be output in a separate file due to now available information in the reference.

IRIS translation can now be performed as an integrated step during IRIS screening. IRIS translation supports translating based on either known ORF in UniProtKB or three ORF translation, which can be useful in searching for AS events from unknown ORF or novel transcripts. An option is provided to remove the entire truncated splice junction peptides due to stop codons.

3.4.20 Updated to stringent threshold for IRIS discovery of candidate TCR and CAR-

T targets from 22 GBM samples

As we noted in the previous analysis, the threshold we used for prioritizing targets with high tumor specificity can be user-defined to allow for higher stringency. In the updated analysis, we applied more stringent threshold:

In screening for the ‘Primary’ set of AS events, we considered an event to be ‘tumor-associated’ if it was significantly different from the tissue-matched normal panel, using the default criteria described in ‘IRIS module for *in silico* screening of tumor AS events’. In screening for the ‘Prioritized’ set, we prioritized an AS event if it was both ‘tumor-recurrent’ (significantly different from the tissue-matched normal panel, in the same direction as input GBM samples, in at least 1 of 2 groups in the GBM/LGG tumor panel) and ‘tumor-specific’ (significantly different from 7 of 11 groups in the normal panel in the same direction as the tissue-matched normal panel). Here, we used at least 7 groups, but this threshold can be user-defined to allow for higher stringency.

3.4.21 Replication analysis using 53 GBM samples

53 GBM samples were sequenced to generate RNA-seq data with paired reads of 150bp read length. The same RNA-seq processing pipeline described above in IRIS RNA-seq data processing module was applied. IRIS analysis was performed using the same parameters for 22 GBM analysis.

The shared events of SE-derived epitope-producing splice junctions between 53 GBM cohort and 22 GBM cohort was defined by AS events with the same tumor isoform (skipping or inclusion).

3.4.22 Code Availability

IRIS source code is accessible on GitHub at https://github.com/Xinglab/IRIS.

3.4.23 Data Availability

The 22 UCLA GBM RNA-Seq data generated for this study were uploaded to BioProject database (BioProject: PRJNA577155). RNA-Seq data used to construct IRIS's normal and tumor reference panels of AS events are available from the GTEx project (https://gtexportal.org/) and The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/legacy-archive/). For the IRIS proteo-transcriptomics analysis, matching RNA-Seq data and MS immunopeptidomics data of B-LCL-S1 and B-LCL-S2 cell lines were retrieved from Laumont et al. (GEO: GSM1641206, GSM1641207 and PRIDE: PXD001898). Raw RNA-Seq data of the JeKo-1 lymphoma cell line were obtained from the Cancer Cell Line Encyclopedia via the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/legacy-archive/). Corresponding immunopeptidomics MS data of JeKo-1 were retrieved from Khodadoust *et al*. (PRIDE: PXD004746).

Acknowledgements

A.R. was funded by the Parker Institute for Cancer Immunotherapy (PICI), the National Institutes of Health (NIH; R35CA197633), and the Ressler Family Fund. O.N.W. was funded by the NIH (R01CA220238, U01CA233074, and P50CA092131) and PICI. Y. Xing was funded by the NIH (R01CA220238 and U01CA233074) and PICI. R.M.P. was funded by the NIH (P50CA211015 and R01CA222695), PICI, and Cancer Research Institute. A.H.L. is a predoctoral fellow supported by the UCLA Tumor Immunology Training Grant (T32CA009120). Flow cytometry and scRNA-Seq experiments were supported in part by funding from the NIH National Center for Advancing Translational Science - UCLA CTSI (UL1TR001881). The authors thank Eric Kutschera (CHOP) for suggestions and assistance with software development, and Life Science Editors for assistance with manuscript editing.

3.5 Figures

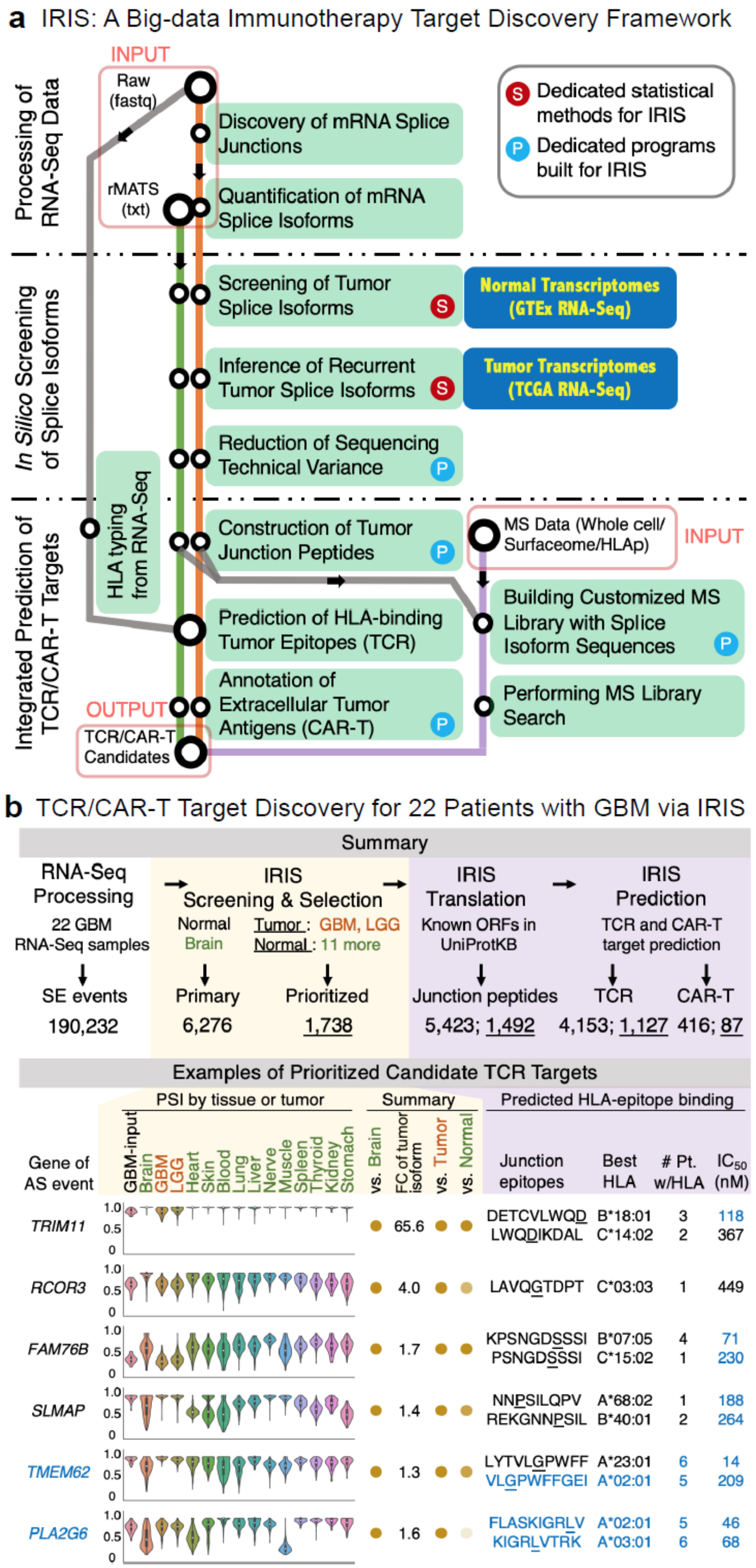
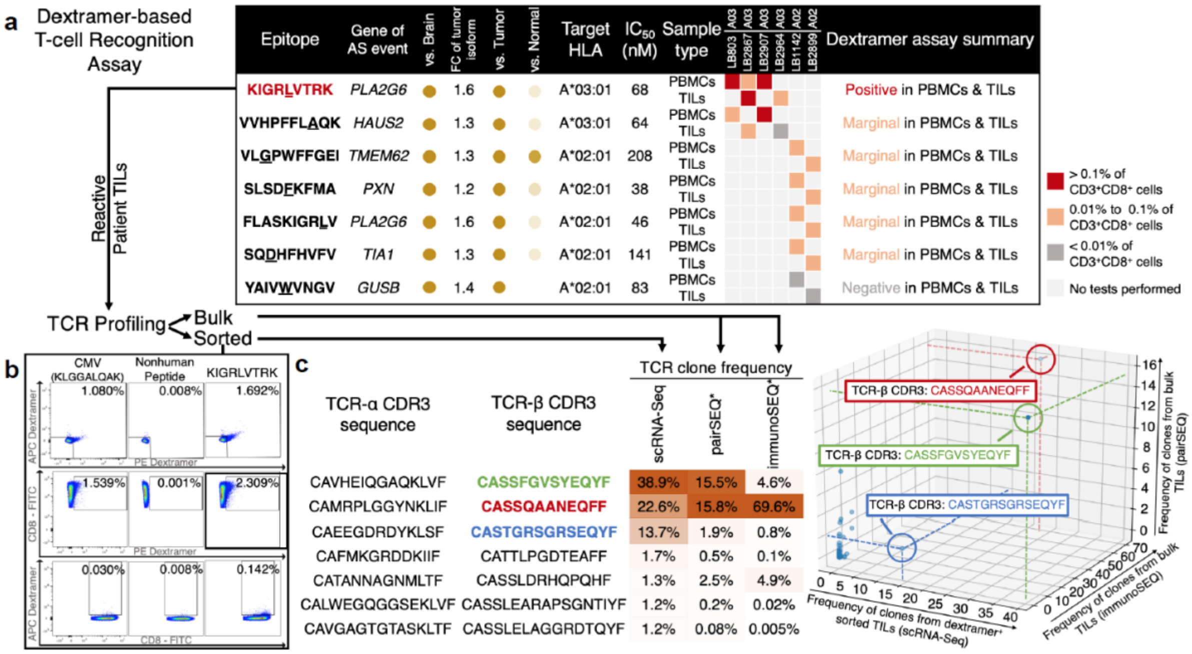

a IRIS: A Big-data Immunotherapy Target Discovery Framework
Processing of RNA-Seq Data
In Silico Screening of Splice Isoforms
Integrated Prediction of TCR/CAR-T Targets
INPUT
Raw (fastq)
rMATS (txt)
Discovery of mRNA Splice Junctions
Quantification of mRNA Splice Isoforms
S Dedicated statistical methods for IRIS
P Dedicated programs built for IRIS
Screening of Tumor Splice Isoforms S
Normal Transcriptomes (GTEx RNA-Seq)
Inference of Recurrent Tumor Splice Isoforms S
Tumor Transcriptomes (TCGA RNA-Seq)
Reduction of Sequencing Technical Variance P
HLA typing from RNA-Seq
Construction of Tumor Junction Peptides P
MS Data (Whole cell/ Surfaceome/HLAp)
INPUT
Prediction of HLA-binding Tumor Epitopes (TCR)
Building Customized MS Library with Splice Isoform Sequences P
Annotation of Extracellular Tumor Antigens (CAR-T) P
Performing MS Library Search
OUTPUT
TCR/CAR-T Candidates
b TCR/CAR-T Target Discovery for 22 Patients with GBM via IRIS
Summary
RNA-Seq Processing
22 GBM RNA-Seq samples
SE events
190,232
IRIS Screening & Selection
Normal Brain
Tumor: GBM, LGG
Normal: 11 more
Primary
6,276
Prioritized
1,738
IRIS Translation
Known ORFs in UniProtKB
Junction peptides
5,423; 1,492
IRIS Prediction
TCR and CAR-T target prediction
TCR
4,153; 1,127
CAR-T
416; 87
Examples of Prioritized Candidate TCR Targets
PSI by tissue or tumor
Summary
Predicted HLA-epitope binding
Gene of AS event
GBM-input
Brain
GBM
LGG
Heart
Skin
Blood
Lung
Liver
Nerve
Muscle
Spleen
Thyroid
Kidney
Stomach
vs. Brain
FC of tumor isoform
vs. Tumor
vs. Normal
Junction epitopes
Best HLA
# Pt. w/HLA
IC50 (nM)
1.0
0.5
0
TRIM11 ● 65.6 ● ● DETCVLWQD B*18:01 3 118
LWQDIKDAL C*14:02 2 367
RCOR3 ● 4.0 ● ● LAVQGTDPT C*03:03 1 449
FAM76B ● 1.7 ● ● KPSNGDSSSI B*07:05 4 71
PSNGDSSSI C*15:02 1 230
SLMAP ● 1.4 ● ● NNPSILQPV A*68:02 1 188
REKGNNPSIL B*40:01 2 264
TMEM62 ● 1.3 ● ● LYTVLGPWFF A*23:01 6 14
VLGPWFFGEI A*02:01 5 209
PLA2G6 ● 1.6 ● ● FLASKIGRLV A*02:01 5 46
KIGRLVTRK A*03:01 6 68

Figure 3.1 IRIS: A big data-powered platform for discovering AS-derived cancer immunotherapy targets

(a) Workflow for IRIS, integrating computational modules, large-scale reference RNA-Seq panels, and dedicated statistical testing programs. IRIS has three main modules: RNA-Seq data processing (top), *in silico* screening (middle), and TCR/CAR-T target prediction (bottom). The prediction module includes an option for proteo-transcriptomics integration of RNA-Seq and MS data. **(b)** Stepwise results of IRIS to identify AS-derived cancer immunotherapy targets from 22 GBM samples (top). Identified skipped-exon (SE) events from the IRIS data-processing module were screened against tissue-matched normal panel ('Normal Brain') to identify tumor-associated events ('Primary' set), followed by tumor panel and normal panel to identify tumor-recurrent and tumor-specific events, respectively ('Prioritized' set). After constructing splice-junction peptides of tumor isoforms, TCR/CAR-T targets were predicted. As an illustrative example, IRIS readouts for prioritized candidate TCR targets are shown (bottom). Violin plots (left) show PSI values of individual AS events across GBM ('GBM-input') versus three reference panels. Dots (middle) summarize screening results. Darker-colored dots indicate stronger tumor features (association/recurrence/specificity) versus each reference panel. FC is estimated fold change of tumor isoform's proportion in GBM ('GBM-input') versus tissue-matched normal panel ('Brain'). Predicted HLA-epitope binding (right) is output of prediction module. Preferred features for immunotherapy targets in this study are shown in blue. Amino acids at splice junctions in epitopes are underlined. 'Best HLA' is HLA type with best predicted affinity (median IC50) for given splice-junction epitope. '#Pt. w/HLA' is number of patients with HLA type(s) predicted to bind to a given epitope. Three epitopes in *TMEM62* and

PLA2G6 (blue) were predicted to bind to common HLA types (HLA-A02:01 and HLA-A03:01) and were selected for experimental validation.



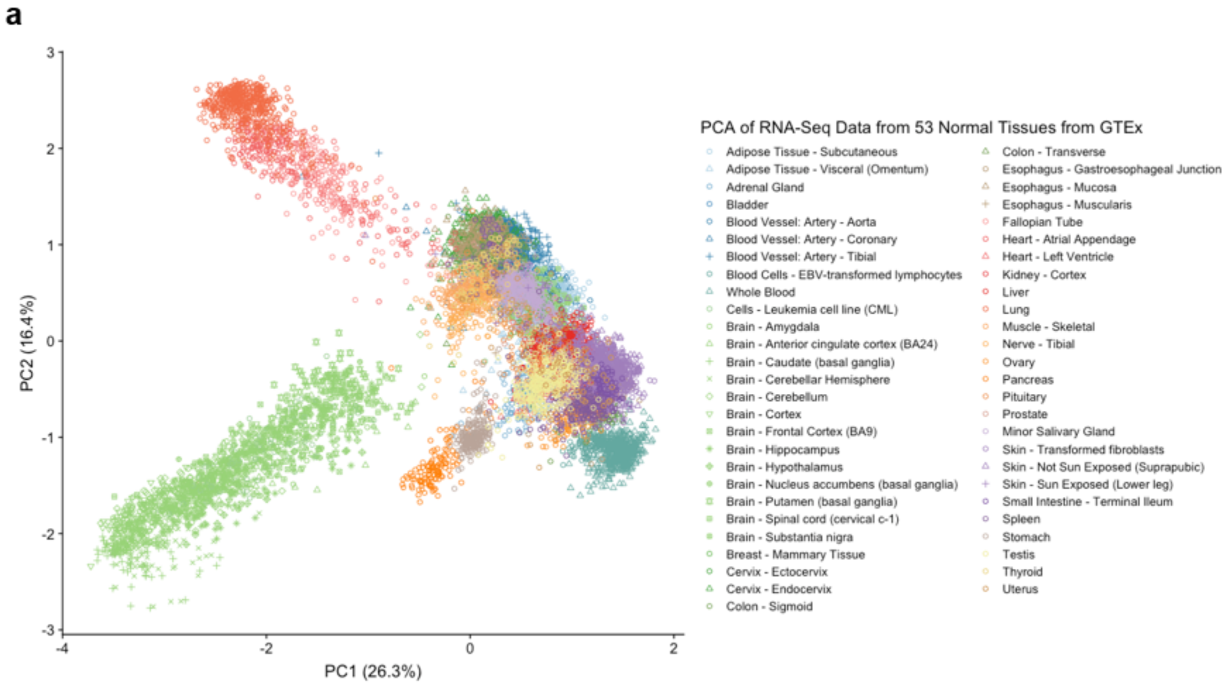
Epitope	Gene of AS event	vs. Brain	FC of tumor isoform	vs. Tumor	vs. Normal	Target HLA	IC50 (nM)	Sample type	LB803 A03	LB2867 A03	LB2907 A03	LB2964 A03	LB1142 A02	LB2899 A02	Dextramer assay summary
KIGRLVTRK	*PLA2G6*	●	1.6	●	○	A*03:01	68	PBMCs TILs	🟥 ⬜	🟧 🟥	🟥 ⬜	⬜ 🟧	⬜ ⬜	⬜ ⬜	Positive in PBMCs & TILs
VVHPFFLAQK	*HAUS2*	●	1.3	●	○	A*03:01	64	PBMCs TILs	🟧 ⬜	⬜ 🟧	🟥 ⬜	⬜ ▨	⬜ ⬜	⬜ ⬜	Marginal in PBMCs & TILs
VLGPWFFGEI	*TMEM62*	●	1.3	●	●	A*02:01	208	PBMCs TILs	⬜ ⬜	⬜ ⬜	⬜ ⬜	⬜ ⬜	🟧 ⬜	⬜ 🟧	Marginal in PBMCs & TILs
SLSDFKFMA	*PXN*	●	1.2	●	○	A*02:01	38	PBMCs TILs	⬜ ⬜	⬜ ⬜	⬜ ⬜	⬜ ⬜	🟧 ⬜	⬜ 🟧	Marginal in PBMCs & TILs
FLASKIGRLV	*PLA2G6*	●	1.6	●	○	A*02:01	46	PBMCs TILs	⬜ ⬜	⬜ ⬜	⬜ ⬜	⬜ ⬜	🟧 ⬜	⬜ 🟧	Marginal in PBMCs & TILs
SQDHFHVFV	*TIA1*	●	1.3	●	○	A*02:01	141	PBMCs TILs	⬜ ⬜	⬜ ⬜	⬜ ⬜	⬜ ⬜	🟧 ⬜	⬜ 🟧	Marginal in PBMCs & TILs
YAIVWVNGV	*GUSB*	●	1.4	●		A*02:01	83	PBMCs TILs	⬜ ⬜	⬜ ⬜	⬜ ⬜	⬜ ⬜	▨ ⬜	⬜ ▨	Negative in PBMCs & TILs

TCR-α CDR3 sequence	TCR-β CDR3 sequence	TCR clone frequency scRNA-Seq	 pairSEQ*	 immunoSEQ*
CAVHEIQGAQKLVF	CASSFGVSYEQYF	38.9%	15.5%	4.6%
CAMRPLGGYNKLIF	CASSQAANEQFF	22.6%	15.8%	69.6%
CAEEGDRDYKLSF	CASTGRSGRSEQYF	13.7%	1.9%	0.8%
CAFMKGRDDKIIF	CATTLPGDTEAFF	1.7%	0.5%	0.1%
CATANNAGNMLTF	CASSLDRHQPQHF	1.3%	2.5%	4.9%
CALWEGQGGSEKLVF	CASSLEARAPSGNTIYF	1.2%	0.2%	0.02%
CAVGAGTGTASKLTF	CASSLELAGGRDTQYF	1.2%	0.08%	0.005%

Figure 3.2 IRIS-predicted AS-derived TCR targets recognized by CD3+CD8+ T cells in tumors and peripheral blood from patients

(a) Summary of dextramer-based validation of IRIS-predicted AS-derived epitopes. PBMCs and/or TILs from four HLA-A03 and two HLA-A02 patients were tested for recognition of IRIS-predicted epitopes. Within each HLA type, epitopes are listed by order of tumor specificity (high to low) versus normal panel (11 normal nonbrain tissues). Reactivity (‘Positive’, ‘Marginal’, or ‘Negative’) in assay was evaluated as percentage of dextramer-labeled cells among PBMCs/TILs (>0.1%, 0.01%-0.1%, or <0.01% of CD3+CD8+ cells, respectively) after subtracting negative control (nonhuman peptide). 'Dextramer assay summary' was determined by the mean percent reactivity of CD3+CD8+ cells across individual tests. **(b)** Flow cytometric analysis showing that *ex vivo*-expanded TILs from one HLA-A03 patient (LB2867) contained T cells that recognized epitope KIGRLVTRK. Rows correspond to cells that recognize APC- and PE-labeled dextramers (top), only PE-labeled dextramers (middle), or only APC-labeled dextramers (bottom).

Percentages of epitope-specific cells are shown. **(c)** Immune profiling results revealing immune repertoire composition of KIGRLVTRK-specific T cells from one patient (LB2867). The scRNA-Seq assay was performed on sorted KIGRLVTRK-specific T cells, whereas pairSEQ and immunoSEQ assays captured TCR clones from bulk TIL RNAs of same patient. Table (left) lists seven most abundant T-cell clones from scRNA-Seq, with percentages of matching CDR3 sequences from TCR β chains. *For pairSEQ and immunoSEQ, percentages are the best frequencies of matching TCR pair or β-chain clones. The 3D scatterplot (right) shows that these approaches converged on three dominant TCR clones. For comparison, the same epitope in the table and 3D scatterplot are identified by use of the same color for the sequence (table) and text box (plot).



b

Summary of 53 Normal Tissues from GTEx

Tissue	Sample Size	Events Detected	Events Selected	Tissue	Sample Size	Events Detected	Events Selected
Adipose	620	293,489	177,621	**Prostate**	119	145,285	77,386
Adrenal	159	154,489	103,394	**Salivary**	70	129,283	68,459
Blood	595	305,861	185,067	**Skin**	974	342,777	212,661
Blood Vessel	750	303,856	190,594	**Small Intestine**	104	152,146	84,231
Bone Marrow	102	177,163	126,033	**Spleen**	118	141,698	93,070
Brain	1,409	399,295	221,221	**Stomach**	204	176,310	114,884
Breast	218	197,683	133,548	**Testis**	203	263,987	179,735
Cervix	11	76,175	46,535	**Thyroid**	361	224,850	149,243
Colon	376	224,758	147,544	**Uterus**	90	145,394	101,249
Esophagus	790	281,759	177,670	**Vagina**	97	148,627	101,789
Fallopian	7	69,716	41,180	**Muscle**	475	221,632	132,075
Heart	489	223,803	133,012	**Nerve**	335	222,251	146,857
Kidney	36	102,385	68,138	**Ovary**	108	151,202	103,115
Liver	136	135,295	79,865	**Pancreas**	197	140,163	83,941
Lung	374	273,703	180,005	**Pituitary**	124	158,497	105,267

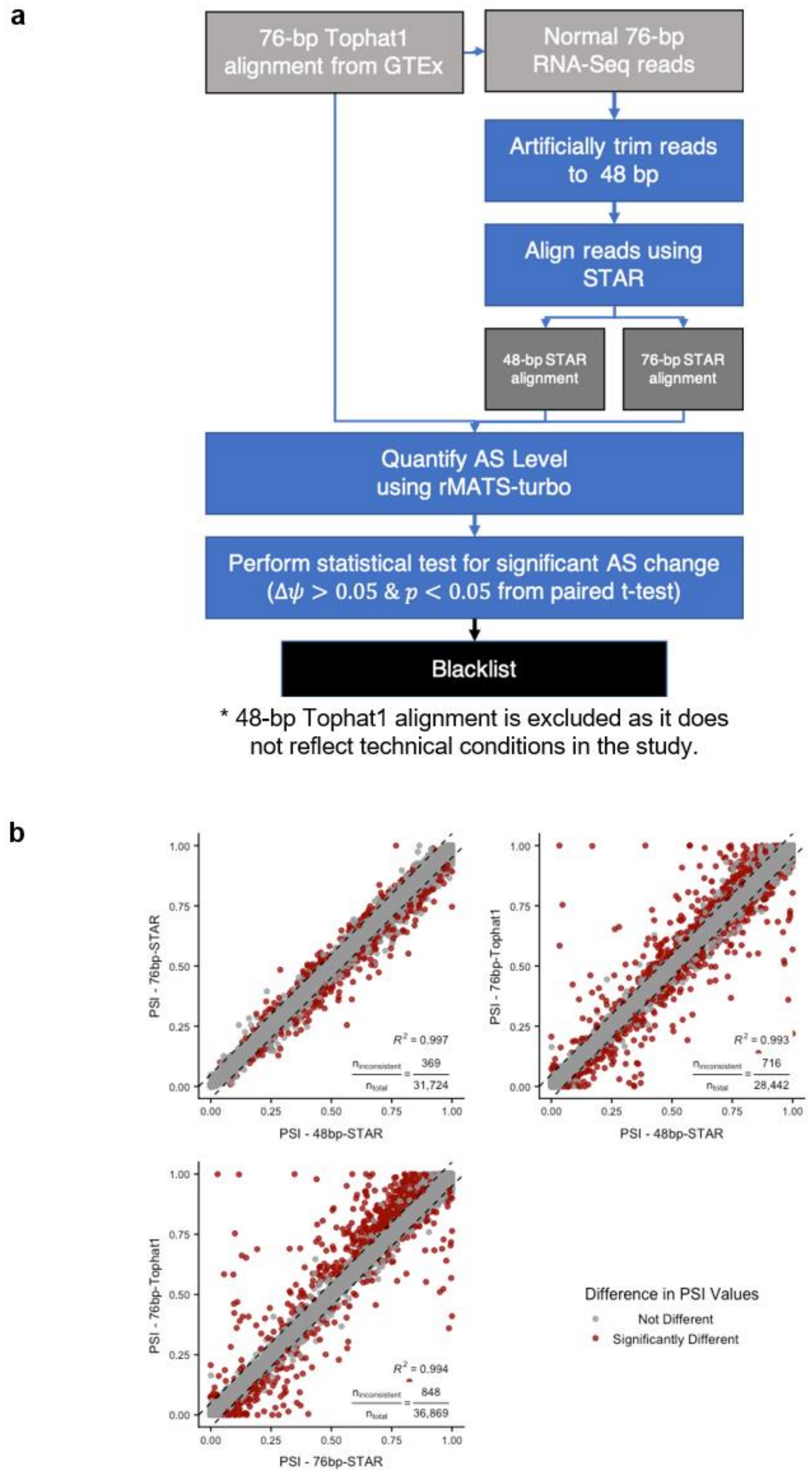
c

Summary of Tumor Reference Panel

Tumor	Sample Size	Events Detected	Events Selected
GBM	162	288,380	190,854
LGG	516	321,777	206,734

Supplementary Figure 3.3 RNA-Seq big-data reference panels in IRIS

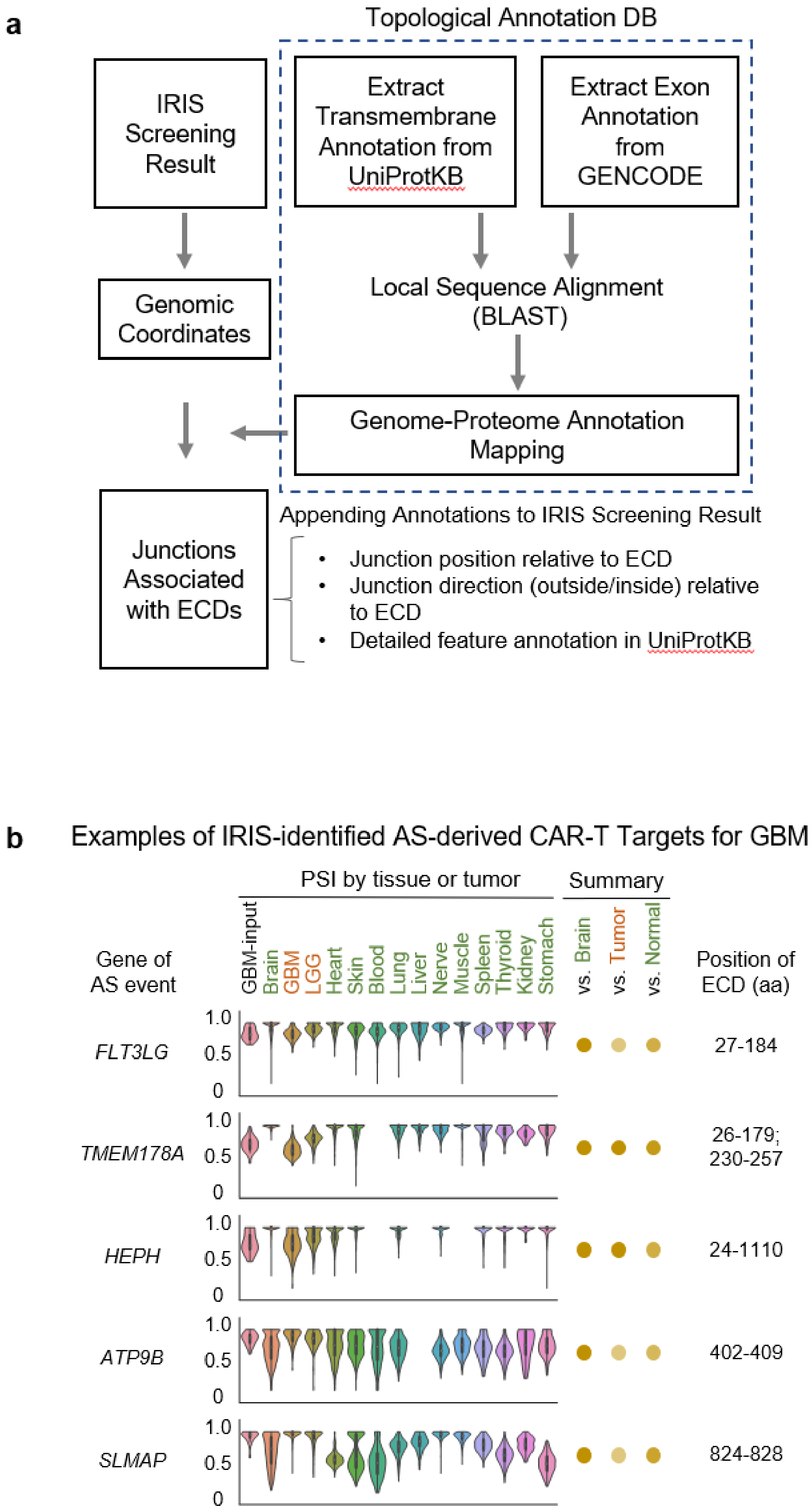
(a) Exon-based principal component analysis (PCA) of RNA-Seq data of 9,662 samples from 53 normal tissues from the GTEx consortium. Samples from the same histological site are grouped by color. Samples from different subregions of the same histological site are differentiated by different shapes. **(b)** Summary of 53 normal tissues from the GTEx consortium. Data for all 53 tissues are available to IRIS users as a reference panel of normal tissues. In the present study, 11 selected vital tissues (heart, skin, blood, lung, liver, nerve, muscle, spleen, thyroid, kidney, and stomach) were used for the ‘normal panel’. ‘Events Selected’ represent AS events with an average count ≥ 10 reads for the sum of all splice junctions across all samples in that tissue. **(c)** Summary of the tumor reference panel (TCGA tumor samples relevant to GBM). ‘Events Selected’ represent AS events with an average count ≥ 10 reads for the sum of all splice junctions across all samples in that tumor type.



* 48-bp Tophat1 alignment is excluded as it does not reflect technical conditions in the study.

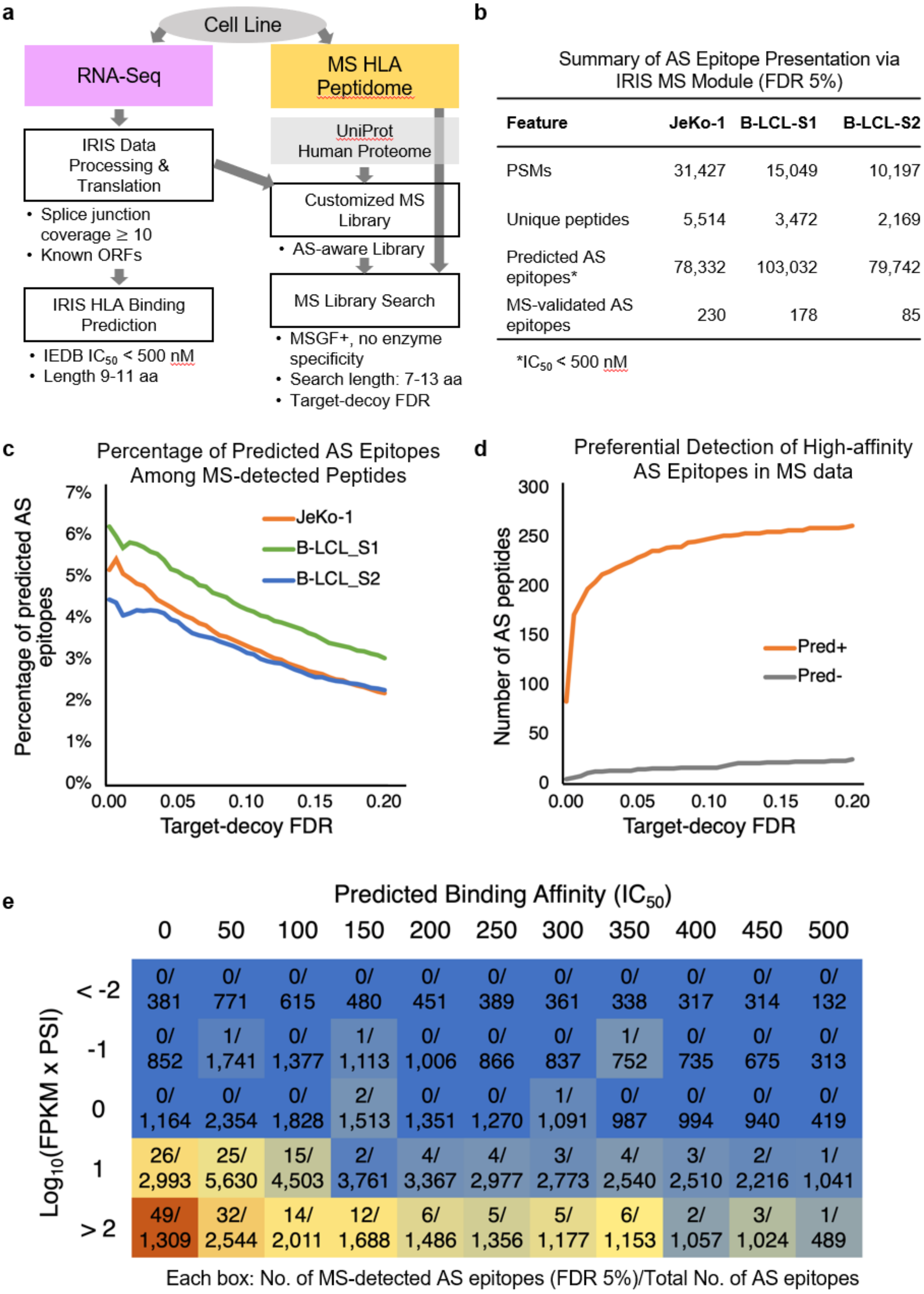
Supplementary Figure 3.4 Identification of AS events that are prone to measurement errors due to technical variances across big-data reference panels

(a) Computational workflow to create a ‘blacklist’ of error-prone AS events. Normal 76-bp RNA-Seq reads were artificially trimmed to 48 bp. RNA-Seq files (76- and 48-bp) were aligned by using two different aligners (Tophat and STAR). AS events were quantified by rMATS-turbo. AS events with statistically significant differences in PSI values among RNA-Seq datasets with distinct technical conditions were identified and included in a blacklist. **(b)** Scatter plots comparing PSI values of GTEx normal brain RNA-Seq data estimated under distinct technical conditions (read lengths: 48- and 76-bp, aligners: STAR and Tophat). ‘Significantly different’ AS events were defined as those with significantly different PSI values ($p < 0.05$, $abs(\Delta\psi) > 0.05$ from paired *t*-test).


a
Topological Annotation DB
IRIS Screening Result
Extract Transmembrane Annotation from UniProtKB
Extract Exon Annotation from GENCODE
Genomic Coordinates
Local Sequence Alignment (BLAST)
Genome-Proteome Annotation Mapping
Junctions Associated with ECDs
Appending Annotations to IRIS Screening Result
• Junction position relative to ECD
• Junction direction (outside/inside) relative to ECD
• Detailed feature annotation in UniProtKB
b
Examples of IRIS-identified AS-derived CAR-T Targets for GBM
PSI by tissue or tumor
Summary
Gene of AS event
GBM-input
Brain
GBM
LGG
Heart
Skin
Blood
Lung
Liver
Nerve
Muscle
Spleen
Thyroid
Kidney
Stomach
vs. Brain
vs. Tumor
vs. Normal
Position of ECD (aa)
FLT3LG
TMEM178A
HEPH
ATP9B
SLMAP
1.0
0.5
0
27-184
26-179;
230-257
24-1110
402-409
824-828

Supplementary Figure 3.5 CAR-T target prediction by IRIS

(a) Computational workflow to annotate protein extracellular domain (ECD)-associated AS events for CAR-T target discovery. **(b)** Five examples of IRIS-identified AS-derived CAR-T targets for 22 GBM samples. Position of the ECD in amino acid (aa) sequence was obtained from UniProtKB.



Summary of AS Epitope Presentation via IRIS MS Module (FDR 5%)

Feature	JeKo-1	B-LCL-S1	B-LCL-S2
PSMs	31,427	15,049	10,197
Unique peptides	5,514	3,472	2,169
Predicted AS epitopes*	78,332	103,032	79,742
MS-validated AS epitopes	230	178	85

*IC_{50} < 500 nM

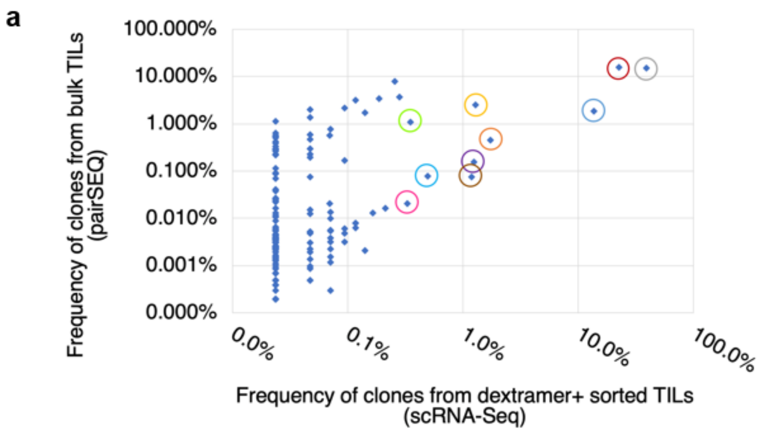
Predicted Binding Affinity (IC_{50})

Log_{10}(FPKM x PSI)	0	50	100	150	200	250	300	350	400	450	500
< -2	0/ 381	0/ 771	0/ 615	0/ 480	0/ 451	0/ 389	0/ 361	0/ 338	0/ 317	0/ 314	0/ 132
-1	0/ 852	1/ 1,741	0/ 1,377	1/ 1,113	0/ 1,006	0/ 866	0/ 837	1/ 752	0/ 735	0/ 675	0/ 313
0	0/ 1,164	0/ 2,354	0/ 1,828	2/ 1,513	0/ 1,351	0/ 1,270	1/ 1,091	0/ 987	0/ 994	0/ 940	0/ 419
1	26/ 2,993	25/ 5,630	15/ 4,503	2/ 3,761	4/ 3,367	4/ 2,977	3/ 2,773	4/ 2,540	3/ 2,510	2/ 2,216	1/ 1,041
> 2	49/ 1,309	32/ 2,544	14/ 2,011	12/ 1,688	6/ 1,486	5/ 1,356	5/ 1,177	6/ 1,153	2/ 1,057	3/ 1,024	1/ 489

Each box: No. of MS-detected AS epitopes (FDR 5%)/Total No. of AS epitopes

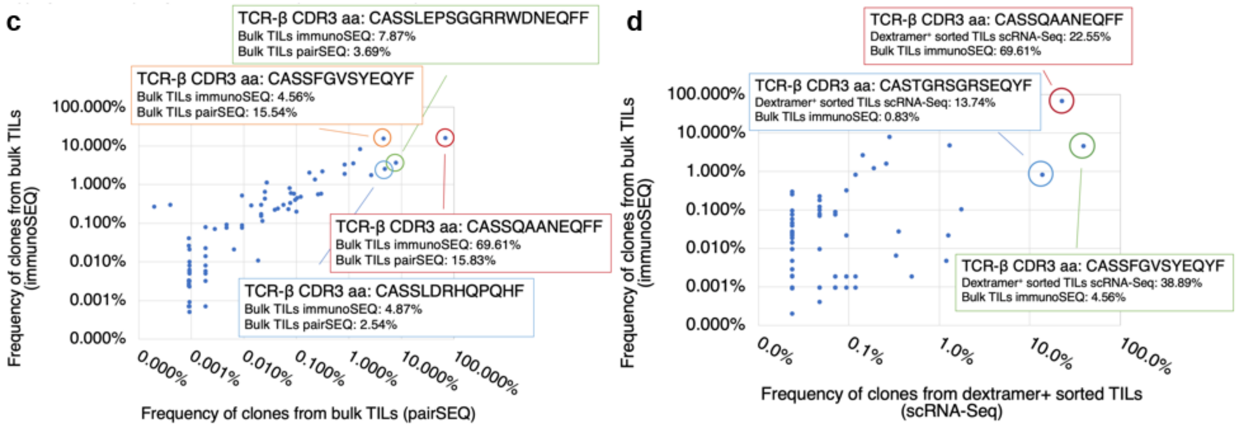
Supplementary Figure 3.6 Proteo-transcriptomic analysis of HLA presentation of AS-derived epitopes in normal and tumor cell lines

(a) Proteo-transcriptomics workflow adopted by IRIS to discover splice-junction peptides in MS datasets. IRIS inputs MS data (right), such as whole-cell proteomics, surfaceomics, or immunopeptidomics (HLA peptidomics) data. RNA-Seq-based custom proteome library is constructed and searched using MSGF+. **(b)** Summary of HLA presentation of AS-derived epitopes in JeKo-1 (lymphoma) and B-LCL (normal) cell lines. Peptide-spectrum matches (‘PSMs’) and ‘Unique peptides’ are provided by MSGF+ with a target-decoy FDR of 5%. ‘Predicted AS epitopes’ are generated by the IRIS prediction module, which utilizes IEDB predictors. AS epitopes that are predicted by IRIS and detected in the MS data are considered ‘MS-validated AS epitopes’. **(c)** Percentage of IRIS-predicted AS-derived epitopes among all MS-detected peptides. Graph shows the percentage of all MS-detected peptides that are IRIS-predicted AS-derived epitopes (y-axis) as a function of the MSGF+ target-decoy FDR (x-axis). **(d)** Preferential detection of high-affinity AS-derived peptides in MS data. Graph shows the number of AS-derived peptides detected in JeKo-1 MS data (y-axis) as a function of the MSGF+ target-decoy FDR (x-axis). Peptides with high (IC_{50} < 500 nM; Pred+, orange) and low (IC_{50} > 500 nM; Pred-, grey) predicted HLA binding affinities are shown. **(e)** Heatmap depiction of distribution of AS-derived epitopes in JeKo-1 MS immunopeptidome, as a function of predicted HLA binding affinity and transcript expression level. AS-derived peptides are binned by the corresponding transcripts’ expression levels and IEDB-predicted binding affinity scores. Heatmap is colored from red (high) to yellow (90_{th} percentile) to blue (low), reflecting the proportion of IRIS-predicted AS-derived epitopes that are MS-detected in each bin.



b

Amino acid sequence		TCR clone frequency	
TCR-α CDR3	TCR-β CDR3	Dextramer^{+} sorted TILs (scRNA-Seq)	Bulk TILs (pairSEQ)
CAVHEIQGAQKLVF	CASSFGVSYEQYF	38.8876%	15.5410%
CAMRPLGGYNKLIF	CASSQAANEQFF	22.5520%	15.8253%
CAEEGDRDYKLSF	CASTGRSGRSEQYF	13.7415%	1.8803%
CAFMKGRDDKIIF	CATTLPGDTEAFF	1.7294%	0.4544%
CATANNAGNMLTF	CASSLDRHQPQHF	1.2853%	2.5376%
CALWEGQGGSEKLVF	CASSLEARAPSGNTIYF	1.2386%	0.1565%
CAVGAGTGTASKLTF	CASSLELAGGRDTQYF	1.1919%	0.0778%
CAVFQGGSEKLVF; CAAADFSGTYKYIF	CASSPEPQGANFYEQYF	0.4908%	0.0802%
CALSEGSNFGNEKLTF	CASSEGTVLDEQYF	0.3505%	1.1144%
CAAGGNYGGSQGNLIF	CASSLGSSTQYF	0.3272%	0.0214%



TCR-β CDR3 aa: CASSLEPSGGRRWDNEQFF
Bulk TILs immunoSEQ: 7.87%
Bulk TILs pairSEQ: 3.69%

TCR-β CDR3 aa: CASSFGVSYEQYF
Bulk TILs immunoSEQ: 4.56%
Bulk TILs pairSEQ: 15.54%

TCR-β CDR3 aa: CASSQAANEQFF
Bulk TILs immunoSEQ: 69.61%
Bulk TILs pairSEQ: 15.83%

TCR-β CDR3 aa: CASSLDRHQPQHF
Bulk TILs immunoSEQ: 4.87%
Bulk TILs pairSEQ: 2.54%

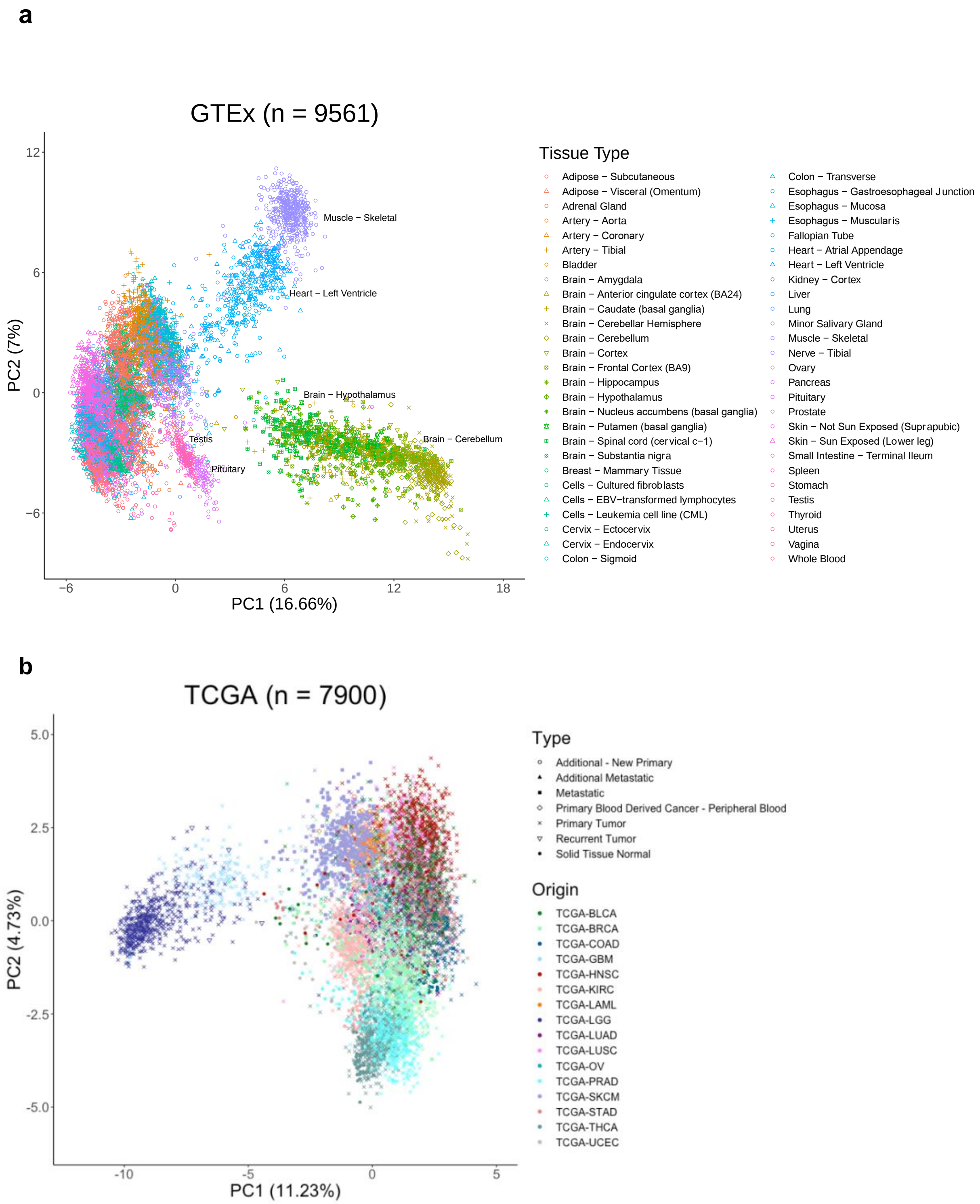
TCR-β CDR3 aa: CASSQAANEQFF
Dextramer^{+} sorted TILs scRNA-Seq: 22.55%
Bulk TILs immunoSEQ: 69.61%

TCR-β CDR3 aa: CASTGRSGRSEQYF
Dextramer^{+} sorted TILs scRNA-Seq: 13.74%
Bulk TILs immunoSEQ: 0.83%

TCR-β CDR3 aa: CASSFGVSYEQYF
Dextramer^{+} sorted TILs scRNA-Seq: 38.89%
Bulk TILs immunoSEQ: 4.56%

Supplementary Figure 3.7 Consistent distributions of high-frequency TCR clones in one patient's TIL population revealed by multiple TCR sequencing approaches

(a) Scatter plot comparing scRNA-Seq and bulk TIL pairSEQ for detection of high-frequency TCR clones. Graph shows frequency detected from bulk TIL samples using pairSEQ (y-axis) and scRNA-Seq on dextramer-positive sorted TIL samples (x-axis). As a complementary validation of scRNA-Seq, clonotypes from pairSEQ were matched to scRNA-Seq results by either CDR3 pairs or β chains, whichever matched best. The 10 most abundant TCR clones by scRNA-Seq that overlapped with clones detected by bulk TIL pairSEQ are circled. **(b)** Table showing CDR3 amino acid sequences of the 10 most abundant TCR clones detected by scRNA-Seq and their corresponding detection frequencies by bulk TIL pairSEQ. As a complementary validation of scRNA-Seq, clonotypes from pairSEQ were matched to scRNA-Seq results by either CDR3 pairs or β chains, whichever matched best. **(c)** Scatter plot comparing bulk TIL immunoSEQ and bulk TIL pairSEQ for detection of high-frequency TCR clones. Graph shows frequency detected from bulk TIL samples using immunoSEQ (y-axis) and pairSEQ (x-axis). Clonotypes from immunoSEQ were matched to pairSEQ results by the best CDR3 β chains. Four high-frequency overlapping clones from both methods are circled and color-coded, with β-chain CDR3 amino acid sequences and frequencies by each method shown in boxes. **(d)** Scatter plot comparing scRNA-Seq and bulk TIL immunoSEQ for detection of high-frequency TCR clones. Graph shows frequency detected from bulk TIL samples using immunoSEQ (y-axis) and scRNA-Seq on dextramer-positive sorted TIL samples (x-axis). As a complementary validation of scRNA-Seq, clonotypes from immunoSEQ were matched to scRNA-Seq results by the best CDR3 β chains. Three high-frequency overlapping clones from both methods are circled and color-coded, with β-chain CDR3 amino acid sequences and frequencies by each method shown in boxes.

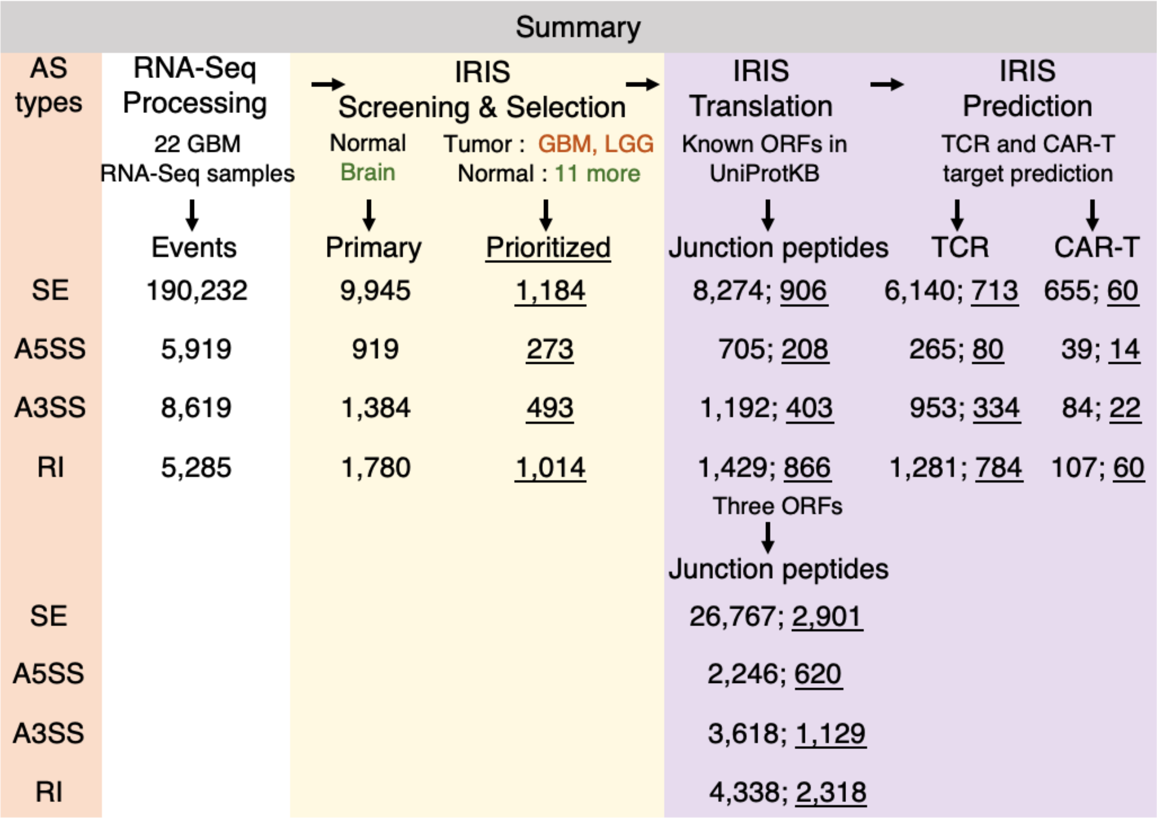


Supplementary Figure 3.8 Updated RNA-seq big-data AS reference panels of IRIS

(a) Exon-based principal component analysis (PCA) of RNA-Seq data of 9,561 samples from 30 normal tissues and one cell lines from the GTEx consortium. Samples from the same histological site are grouped by color. Samples from different subregions of the same histological site are differentiated by different shapes.

(b) Exon-based principal component analysis (PCA) of RNA-Seq data of 7,900 samples from 16 tumor types from the TCGA consortium. Samples from the same tumor type are grouped by color. Samples representing different disease stage of the same tumor type are differentiated by different shapes.

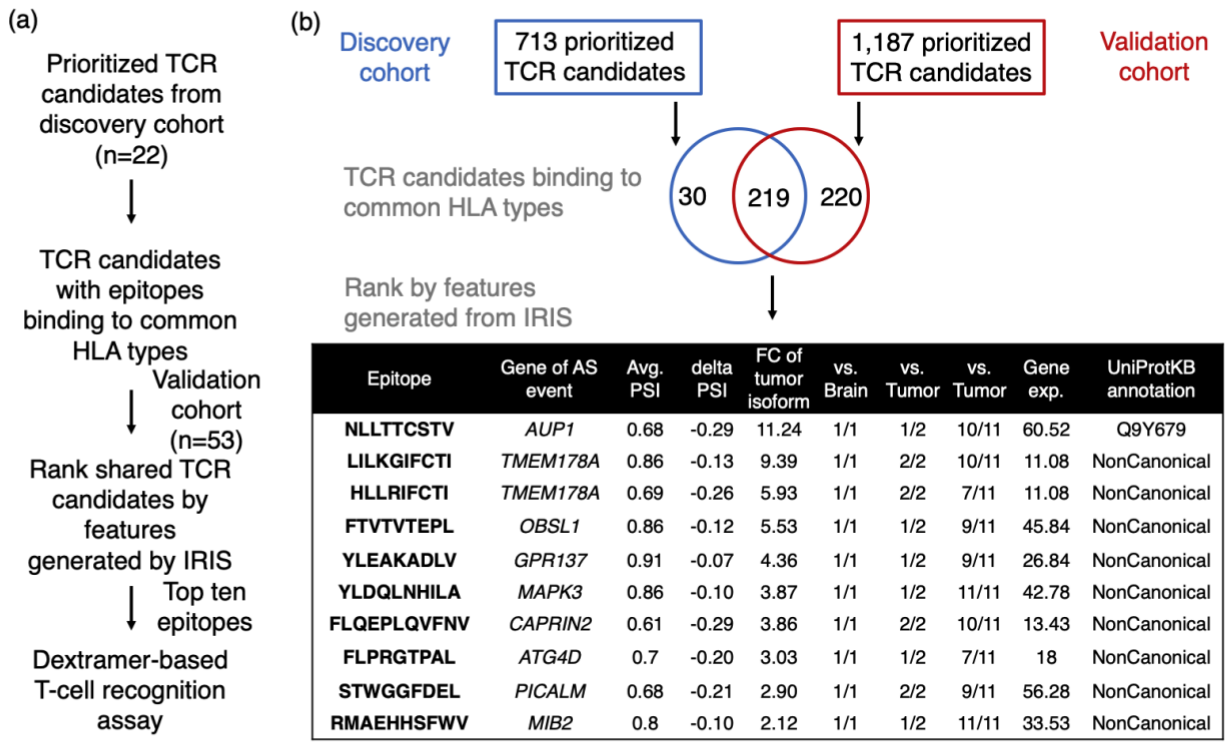
a



Summary						
AS types	RNA-Seq Processing →	IRIS Screening & Selection →		IRIS Translation →	IRIS Prediction	
	22 GBM RNA-Seq samples	Normal Brain	Tumor : GBM, LGG Normal : 11 more	Known ORFs in UniProtKB	TCR and CAR-T target prediction	
	Events	Primary	Prioritized	Junction peptides	TCR	CAR-T
SE	190,232	9,945	1,184	8,274; 906	6,140; 713	655; 60
A5SS	5,919	919	273	705; 208	265; 80	39; 14
A3SS	8,619	1,384	493	1,192; 403	953; 334	84; 22
RI	5,285	1,780	1,014	1,429; 866	1,281; 784	107; 60
				Three ORFs		
				Junction peptides		
SE				26,767; 2,901		
A5SS				2,246; 620		
A3SS				3,618; 1,129		
RI				4,338; 2,318		

Supplementary Figure 3.9 Discover diverse forms of AS-derived tumor antigens from 22 GBM samples using upgraded IRIS

(a) Stepwise results of updated IRIS to identify AS-derived cancer immunotherapy targets from 22 GBM samples (top). Identified skipped-exon (SE), alternative 5’ splice site (A5SS), alternative 3’ splice site (A3SS), and retained intron (RI) events from the IRIS data-processing module were screened against tissue-matched normal panel (‘Normal Brain’) to identify tumor-associated events (‘Primary’ set), followed by tumor panel and normal panel to identify tumor-recurrent and tumor-specific events, respectively (‘Prioritized’ set). Followed by inference based on normal and tumor reference, two translation strategies were used: translating based on known ORFs in proteome database or using all three ORFs. After constructing splice-junction peptides of tumor isoforms, TCR/CAR-T targets were predicted.


(a)
Prioritized TCR candidates from discovery cohort (n=22)
TCR candidates with epitopes binding to common HLA types
Validation cohort (n=53)
Rank shared TCR candidates by features generated by IRIS
Top ten epitopes
Dextramer-based T-cell recognition assay
(b)
Discovery cohort
713 prioritized TCR candidates
1,187 prioritized TCR candidates
Validation cohort
TCR candidates binding to common HLA types
30
219
220
Rank by features generated from IRIS

Epitope	Gene of AS event	Avg. PSI	delta PSI	FC of tumor isoform	vs. Brain	vs. Tumor	vs. Tumor	Gene exp.	UniProtKB annotation
NLLTTCSTV	*AUP1*	0.68	-0.29	11.24	1/1	1/2	10/11	60.52	Q9Y679
LILKGIFCTI	*TMEM178A*	0.86	-0.13	9.39	1/1	2/2	10/11	11.08	NonCanonical
HLLRIFCTI	*TMEM178A*	0.69	-0.26	5.93	1/1	2/2	7/11	11.08	NonCanonical
FTVTVTEPL	*OBSL1*	0.86	-0.12	5.53	1/1	1/2	9/11	45.84	NonCanonical
YLEAKADLV	*GPR137*	0.91	-0.07	4.36	1/1	1/2	9/11	26.84	NonCanonical
YLDQLNHILA	*MAPK3*	0.86	-0.10	3.87	1/1	1/2	11/11	42.78	NonCanonical
FLQEPLQVFNV	*CAPRIN2*	0.61	-0.29	3.86	1/1	2/2	10/11	13.43	NonCanonical
FLPRGTPAL	*ATG4D*	0.7	-0.20	3.03	1/1	1/2	7/11	18	NonCanonical
STWGGFDEL	*PICALM*	0.68	-0.21	2.90	1/1	2/2	9/11	56.28	NonCanonical
RMAEHHSFWV	*MIB2*	0.8	-0.10	2.12	1/1	1/2	11/11	33.53	NonCanonical

Supplementary Figure 3.10 A validation cohort of 53 GBM samples to prioritize IRIS-predicted TCR targets

(a) A flowchart of TCR targets replication and prioritization for T-cell based assay. From top to bottom are steps from targets predicted from the discovery cohort (22 GBM samples) to shared targets from the validation cohort (53 GBM samples) and eventually ranked and nominated for dextramer-based T-cell recognition assay. ‘Prioritized TCR candidates’ in the flowchart refers to SE-derived epitope-producing splice junctions.

(b) Venn diagram (top) showing numbers of Prioritized TCR candidates from both discovery and validation cohorts. (Bottom) A table of selected tumor antigen candidates shared by both discovery and validation cohorts and prioritized for FC of tumor isoform, common HLA type and gene expression of AS events.

3.6 Tables

Supplementary Table 3.1 Summary of IRIS reference panel of normal tissues

Tissue	**Sample Size**	**SE**	**A5**	**A3**	**RI**	**AS Events***
Adipose	609	611,200	13,136	18,776	5,746	648,858
Adrenal	160	589,492	12,639	18,304	5,654	626,089
Bladder	10	491,483	9,927	14,288	4,897	520,595
Blood	440	482,479	10,828	15,775	5,112	514,194
Blood Vessel	743	603,839	12,767	18,169	5,596	640,371
Brain	1,392	621,813	13,164	19,219	5,893	660,089
Breast	218	627,785	13,663	19,430	5,891	666,769
Cells	537	656,764	14,420	20,214	5,832	697,230
Cervix	11	486,228	10,061	14,557	4,968	515,814
Colon	379	617,566	13,344	19,180	5,900	655,990
Esophagus	797	618,476	13,285	18,946	5,817	656,524
Fallopian	7	439,840	9,086	13,256	4,733	466,915
Heart	476	577,088	11,969	17,316	5,460	611,833
Kidney	36	552,862	11,954	17,347	5,661	587,824
Liver	138	527,080	11,498	16,866	5,531	560,975
Lung	351	637,267	13,936	19,844	5,978	677,025
Muscle	464	548,270	11,203	15,974	5,106	580,553
Nerve	338	613,817	13,326	19,080	5,771	651,994
Ovary	112	604,841	13,238	18,909	5,748	642,736
Pancreas	192	541,108	11,797	16,904	5,530	575,339
Pituitary	130	622,966	13,587	19,551	5,988	662,092
Prostate	120	611,928	13,370	19,220	5,917	650,435
Salivary	70	600,999	12,969	18,269	5,764	638,001
Skin	661	599,501	12,915	18,482	5,756	636,654
Small Intestine	103	622,114	13,640	19,629	5,967	661,350
Spleen	118	572,067	12,919	18,818	5,772	609,576
Stomach	207	602,230	12,954	18,521	5,828	639,533
Testis	201	683,591	14,642	21,102	6,230	725,565
Thyroid	345	620,664	13,650	19,579	5,882	659,775
Uterus	95	607,750	13,360	19,050	5,770	645,930
Vagina	101	619,740	13,513	19,204	5,869	658,326

*AS Events are rMATS detected events with junction reads coverage ≥ 10

Supplementary Table 3.2 Summary of IRIS reference panel of tumors

Tissue	Type	Sample Size	SE	A5	A3	RI	AS Events*
BLCA	Normal	19	94,331	6,589	10,286	6,079	117,285
	Tumor	414	304,431	10,427	15,501	6,468	336,827
BRCA	Normal	111	186,607	8,265	12,654	6,266	213,792
	Tumor	1,105	446,799	12,134	17,748	6,562	483,243
COAD	Normal	41	114,906	7,024	10,922	6,139	138,991
	Tumor	300	252,997	9,662	14,253	6,414	283,326
GBM	Normal	5	70,604	5,939	9,400	5,970	91,913
	Tumor	170	311,583	10,162	15,064	6,425	343,234
HNSC	Normal	44	134,131	7,366	11,331	6,183	159,011
	Tumor	522	353,471	10,854	16,117	6,476	386,918
KIRC	Normal	72	165,549	7,942	12,139	6,237	191,867
	Tumor	542	323,704	10,578	15,790	6,465	356,537
LAML	Tumor	179	231,141	9,033	13,339	6,340	259,853
LGG	Tumor	534	352,756	10,708	16,013	6,471	385,948
LUAD	Normal	59	131,136	7,356	11,308	6,172	155,972
	Tumor	541	326,893	10,623	15,904	6,472	359,892
LUSC	Normal	51	152,524	7,798	11,911	6,231	178,464
	Tumor	501	374,540	11,259	16,649	6,528	408,976
OV	Tumor	430	479,478	12,657	18,532	6,607	517,274
PRAD	Normal	52	135,478	7,509	11,504	6,196	160,687
	Tumor	502	268,025	9,900	14,771	6,418	299,114
SKCM	Normal	1	49,122	5,475	8,665	5,893	69,155
	Tumor	472	322,572	10,520	15,666	6,462	355,220
STAD	Normal	37	151,891	7,694	11,876	6,204	177,665
	Tumor	416	444,670	12,035	17,811	6,545	481,061
THCA	Normal	59	154,081	7,871	11,954	6,229	180,135
	Tumor	513	287,287	10,047	15,129	6,442	318,905
UCEC	Normal	24	95,209	6,679	10,376	6,113	118,377
	Tumor	184	214,604	9,064	13,627	6,377	243,672

*AS Events are rMATS detected events with junction reads coverage ≥ 10

3.7 Supplementary materials

Supplementary Data 3.1

IRIS MS analysis of AS-derived epitopes in cell line immunopeptidomics datasets.

a. JeKo-1 cancer cell line with FDR = 5%.

b. B-LCL-S1 normal cell line with FDR = 5%.

c. B-LCL-S2 normal cell line with FDR = 5%.

Supplementary Data 3.2

IRIS screening results of tumor AS events in 22 GBM samples.

a. IRIS identified 6,276 tumor-associated AS events (Primary set)

b. IRIS identified 1,738 tumor-recurrent AS events with high tumor-specificity in GBM samples (Prioritized set)

Supplementary Data 3.3

IRIS prediction of AS-derived TCR and CAR-T targets for 22 GBM samples.

a. Prioritized TCR targets listed by unique splice junction.

b. Prioritized CAR-T targets listed by unique splice junction.

Supplementary Data 3.4

Summary results for seven selected AS-derived tumor-associated epitopes for dextramer-based T-cell recognition testing.

Supplementary Data 3.5

Summary results of dextramer-based T-cell recognition testing of seven AS-derived tumor epitopes using PBMCs and TILs from six patients and six healthy donors with two different HLA types. Data are shown normalized to results with nonhuman epitope NI3233 as a negative control. A common virus found in 50-80% of the population, cytomegalovirus (CMV) was included as a control (Macguire et al., Methods, 2017). n/a, results not available. Green, 'positive' reactivity; yellow, 'marginal' reactivity; red, 'negative' reactivity.

Supplementary Data 3.6

Summary results for TCR clonotypes of KIGRLVTRK-positive T cells from one patient, profiled by single-cell and bulk RNA-seq based approaches. Amino acid sequences of TCR CDR3 and frequencies of clonotypes by scRNA-seq, pairSEQ, and immunoSEQ are shown.

3.8 References

1. Sun, C., Mezzadra, R. and Schumacher, T.N. (2018) Regulation and Function of the PD-L1 Checkpoint. *Immunity*, **48**, 434-452.
2. Rosenberg, S.A. and Restifo, N.P. (2015) Adoptive cell transfer as personalized immunotherapy for human cancer. *Science*, **348**, 62-68.
3. Yarchoan, M., Johnson, B.A., 3rd, Lutz, E.R., Laheru, D.A. and Jaffee, E.M. (2017) Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer*, **17**, 569.
4. Schumacher, T.N. and Schreiber, R.D. (2015) Neoantigens in cancer immunotherapy. *Science*, **348**, 69-74.
5. Lee, C.H., Yelensky, R., Jooss, K. and Chan, T.A. (2018) Update on Tumor Neoantigens and Their Utility: Why It Is Good to Be Different. *Trends Immunol*, **39**, 536-548.
6. Coulie, P.G., Van den Eynde, B.J., van der Bruggen, P. and Boon, T. (2014) Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat Rev Cancer*, **14**, 135-146.
7. (2017) The problem with neoantigen prediction. *Nat Biotechnol*, **35**, 97.
8. Vitiello, A. and Zanetti, M. (2017) Neoantigen prediction and the need for validation. *Nat Biotechnol*, **35**, 815-817.
9. Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M.J., van de Haar, J., Engin, H.B., de Prisco, N., Ideker, T., Hildebrand, W.H., Font-Burgada, J. *et al.* (2017) MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*, **171**, 1272-1283 e1215.

10. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L. *et al.* (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, **547**, 217-221.

11. Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Lower, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrors, B. *et al.* (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, **547**, 222-226.

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

13. Laumont, C.M., Vincent, K., Hesnard, L., Audemard, E., Bonneil, E., Laverdure, J.P., Gendron, P., Courcelles, M., Hardy, M.P., Cote, C. *et al.* (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med*, **10**.

14. Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.K. and Van Allen, E.M. (2018) Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*, **36**, 1056-1058.

15. Zhang, M., Fritsche, J., Roszik, J., Williams, L.J., Peng, X., Chiu, Y., Tsou, C.C., Hoffgaard, F., Goldfinger, V., Schoor, O. *et al.* (2018) RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat Commun*, **9**, 3919.

16. Kahles, A., Lehmann, K.V., Toussaint, N.C., Huser, M., Stark, S.G.,

Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research, N. *et al.* (2018) Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, **34**, 211-224 e216.

17. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.

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

19. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q. and Xing, Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, **111**, E5593-5601.

20. Xie, Z. and Xing, Y. (2019) rMATS-turbo. *http://rnaseq-rmats.sourceforge.net/rmats4.0.2*.

21. Bonifant, C.L., Jackson, H.J., Brentjens, R.J. and Curran, K.J. (2016) Toxicity and management in CAR T-cell therapy. *Mol Ther Oncolytics*, **3**, 16011.

22. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M. *et al.* (2017) Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*, **46**, 315-326.

23. 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.

24. Hadrup, S.R. and Schumacher, T.N. (2010) MHC-based detection of antigen-specific CD8+ T cell responses. *Cancer Immunol Immunother*, **59**, 1425-1433.

25. 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.

26. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562-578.

27. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, **22**, 1760-1774.

28. Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.

29. Tukey, J.W. (1977) *Exploratory data analysis*. Reading, Mass. : Addison-Wesley Pub. Co.

30. Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A. and Grant, G.R. (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods*, **14**, 135-139.

31. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, **46**, 2699.

32. Boegel, S., Lower, M., Schafer, M., Bukur, T., de Graaf, J., Boisguerin, V., Tureci, O., Diken, M., Castle, J.C. and Sahin, U. (2012) HLA typing from RNA-Seq sequence reads. *Genome Med*, **4**, 102.

33. Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A. *et al.* (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*, **43**, D405-412.

34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

35. Kim, S. and Pevzner, P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*, **5**, 5277.

36. Laumont, C.M., Daouda, T., Laverdure, J.P., Bonneil, E., Caron-Lizotte, O., Hardy, M.P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P. *et al.* (2016) Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*, **7**, 10238.

37. Khodadoust, M.S., Olsson, N., Wagar, L.E., Haabeth, O.A., Chen, B., Swaminathan, K., Rawson, K., Liu, C.L., Steiner, D., Lund, P. *et al.* (2017) Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. *Nature*, **543**, 723-727.

Chapter 4 Concluding Remarks

The rapid development of high throughput sequencing technologies in the past decade has profoundly changed the face of biomedical research. Fueled by advanced sequencing technologies, cancer research has also witnessed tremendous progress in the past ten years. Large numbers of cancer drivers, signatures and predictive biomarkers were identified through large-scale deep sequencing efforts or functional genomics studies. Advances in cancer immunotherapy has led to breakthroughs in cancer treatment. Collectively, huge opportunities await in data-driven knowledge discovery in cancer research.

The exponential growth of large-scale sequencing data of various types of cancer greatly empowers research for elucidating cancer mechanisms and developing therapies. However, the challenge remains in how to effectively utilize and mine gigantic sequencing datasets for knowledge discovery. This dissertation has been focused on leveraging novel big data frameworks to study tumor transcriptomes, in particular, on the roles of RNA alternative splicing in cancers and its therapeutic applications. Alternative splicing is a prevalent source of transcriptomic and proteomic diversity in cancer cells and exerts important functions during oncogenesis. Aberrant alternative splicing events have been extensively reported in cancers. Large-scale transcriptome analyses often find large amounts of splicing alterations in tumors with limited understanding of their roles in cancer development and therapeutic potentials. Seeking to translate sequencing big data to in-depth understanding of alternative splicing in tumors, computational tools were developed to leverage the large-scale transcriptomic or multi-omics data to address fundamental or translational questions in cancer research, with the following two components.

In chapter 2, we aimed to study the roles of alternative splicing dysregulation during cancer progression. Large-scale RNA-seq analyses are able to detect thousands of altered splicing events in cancer, while the ability of distinguishing essential changes associated with oncogenesis from the large amount of detected passenger events is lacking. Therefore, we developed a novel analytical method, called PEGASAS, bridging the gap between alternative splicing to oncogenic signals using a pathway-guided correlation analysis to mine large-scale RNA-seq data. From the same sample, PEGASAS computes orthogonal and robust measurements of splicing level and oncogenic pathway activity, maximizing the use of RNA-seq data. Applying PEGASAS to study a comprehensive meta-dataset of prostate cancer samples, we revealed a group of splicing events tied to oncogenic signals alterations and established a regulatory role of Myc in RNA processing, which was validated experimentally. This study demonstrated a system biology approach coupling big data analysis with experimental perturbations to gain insights in RNA regulations in cancers. Expanding the PEGASAS analysis to two other epithelial tumors, we observed similar mechanisms. Our findings highlighted the capability of PEGASAS in mining massive RNA-seq data to uncover intrinsic mechanisms in cancers, which could lead to mechanistic discoveries as well as therapeutic targets.

A natural extension of this study is to carry out a pan-cancer analysis of alternative splicing using this pathway-driven method. Pan-cancer analysis of alternative splicing and its genetic basis are extensive, while their interplay with oncogenic activations is less well characterized. In our published work, the successful application of PEGASAS to investigate prostate, breast, and lung malignancies suggests this framework can be applied to elucidate conserved or tumor-specific interplays between splicing and oncogenic pathways. Moreover, PEGASAS presents a generic framework that can be used to robustly assess associations between gene signatures and other RNA-level dysregulations (e.g. RNA editing) in cancer through large-scale RNA-seq data integration.

In chapter 3, the widespread dysregulated alternative splicing events in tumors were systematically examined for their potential as tumor antigens for cancer immunotherapy. One major limitation of the current paradigm of tumor antigen discovery is that they are mainly genomic variation-based, resulting in a limited number of targets for tumors with moderate or low mutation. Instead, alternative splicing dysregulations generate tumor isoforms, as suggested in chapter 2. To discover novel tumor antigens from this underexploited source, we developed a big data-powered platform, named IRIS, which harnesses large-scale cancer and normal transcriptomics data to infer potential targets that are common and specific to tumor cells. Aiming to provide an integrated solution, IRIS is built to provide a systematically identification and inference of tumor splicing isoforms along with a simultaneous prediction of both T-cell receptor (TCR) and chimeric antigen receptor T-cell (CAR-T cell) therapies. We subjected IRIS to study RNA-Seq data from 22 patients with glioblastoma, we discovered thousands of candidate targets for TCR and CAR-T cell therapies, which would have been overlooked by existing immunotherapy target discovery strategies. We experimentally confirmed that predicted AS-derived tumor antigens were recognized by patient T cells, validating the utility of IRIS for discovery tumor antigens arising from alternative splicing.

IRIS presented additional conceptual and technical advances. A recent review (Frankiw, L., et al. (2019). Nat Rev Immunol) specified one major issue in the existing works attempting to utilize splicing-derived targets, which is the absence of methods for robust screening for tumor events. This underlined the need of building a uniformly processed, standardized and widely accessible reference database of alternative splicing patterns in normal and disease cells. The reference database established by IRIS contains splicing profiles from thousands normal and tumor transcriptomes, reflecting the dynamic of splicing events across individuals and the specificity between tissues. The IRIS reference is made public available along with the IRIS standalone program. Moreover, the proteomics MS validation module of IRIS adopted a proteo-transcriptomic approach (Nesvizhskii, A. I. (2014). Nat Methods), allowing identifying aberrantly expressed novel isoforms that are undetectable using a default known protein database for MS search. Altogether, IRIS provides a standardized multi-omics framework that not only performs robust inference of tumor specificity for splicing events, but also greatly promotes the reproducibility of the antigen discovery analysis.

Furthermore, the big data-informed tumor antigen discovery framework proposed by IRIS is not limited to alternative splicing derived targets. With minor modification, the IRIS platform will be able to incorporate other types of tumor-specific or tumor-associated events resulting from dysregulated RNA processing in cancers. As reviewed in chapter 1 and other articles (Smith, C. C., et al. (2019). Nat Rev Cancer), chimeric RNAs, circular RNAs, RNA editing events, Alu exons and many other RNA events that showed aberrant expression and can be translated in cancer cells could be leveraged as tumor targets. Some of these aberrant RNA molecules may introduce strong immunogenicity as their peptide sequences may have high foreignness to the immune system. Altogether, these novel categories of aberrant isoforms can further expand and diversify the repertoire of tumor antigens, benefiting more cancer patients with transcriptome dysregulations.

Moving forward, cancer target discovery and therapy development will significantly benefit from the fast evolving of detection technologies and big-data integration platforms. These new detection technologies will include but not limited to the third generation long-reads sequencing technologies, single-cell manipulation and profiling technologies, etc. Undoubtedly, such emerging sequencing technologies will produce huge volumes of data requiring novel computational solutions. This opens doors for massive discovery of tumor specific targets. With the development of more enabling detection technologies in the next decade, the picture of dynamic interactions between cells will become more complete, which will help to better elucidate intriguing mechanisms of tumor-immune interaction or other involved biological systems. A system-level understanding enabled by high dimensional data integration will fundamentally improve the therapeutic targets search and treatment development in the future.