Detection and Validation of Aberrant Splicing in Cancer

By Julianne K. David

A DISSERTATION

Presented to the Department of Biomedical Engineering
and the Oregon Health & Science University
School of Medicine

in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

April 2022

Department of Biomedical Engineering
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Julianne K. David

has been approved

_____
Mentor: Abhinav Nellore, Ph.D
Assistant Professor of Biomedical Engineering

_____
Mentor: Reid Thompson, M.D., Ph.D
Assistant Professor of Biomedical Engineering

_____
Chair: Jeremy Goecks, Ph.D
Associate Professor of Biomedical Engineering

_____
Member: Stuart Ibsen, Ph.D
Assistant Professor of Biomedical Engineering

_____
Member: Paul Spellman, Ph.D
Professor of Molecular and Medical Genetics

_____
Member: Zheng Xia, Ph.D
Assistant Professor of Biomedical Engineering

TABLE OF CONTENTS

Table of Contents ..... **i**
Acknowledgements ..... **iii**
List of Figures ..... **iv**
List of Tables ..... **vii**
List of Abbreviations ..... **viii**
Abstract ..... **x**
Chapter 1: Introduction **1**
1.1 Introduction 1
1.2 Utility of identifying cancer-specific variants 1
1.2.1 Cancer immunotherapy 2
1.2.2 Biomarkers for early detection and other applications 3
1.2.3 Cancer neoepitope identification 4
1.3 RNA splicing and splice variants 5
1.3.1 RNA and RNA splicing 5
1.3.2 Intron retention 7
1.3.3. Regulation of alternative splicing 8
1.3.3 Splicing dysregulation in cancer 10
1.4 Identification of aberrant splicing from RNA-seq data 12
1.4.1 Challenges of cancer-specific splicing identification 13
1.4.2 RNA sequencing, splice-aware alignment, and splice site identification 14
1.4.3 Additional challenges in intron retention detection 16
1.4.4 Defining “normal” splicing 18
1.5 Summary 20
1.5.1 Challenges and opportunities 20
1.5.2 Contributions 21
Chapter 2: Exploring the cancer-specificity of tumor junctions **22**
2.1 Abstract 22
2.2 Background 22
2.3 Methods 24
2.4 Results 31
2.4.1 Cancers harbor many novel shared exon-exon junctions not present in adult non-cancer tissues or cells 31
2.4.2 Shared novel junctions in cancer distinguish cancer identity and subtype 35

2.4.3 Novel junctions in cancer are found among developmental and known cancer-related pathways....37
2.5 Discussion....41
Chapter 3: Exploring the validity of cancer-specificity of junctions....44
3.1 Abstract....44
3.2 Background....44
3.3 Results....46
3.3.1 Overview of filtering experiments and two pipelines....46
3.3.2 Junctions and junction peptides found in ovarian and breast cancer samples....48
3.3.3 Summary of filter experiment results....50
3.3.4 Mass spectrometry queries of BRCA peptides across experiments....52
3.4 Discussion....54
3.5 Methods....57
Chapter 4: Testing the reliability of retained intron detection....68
4.1 Abstract....68
4.2 Background....68
4.3 Results....70
4.3.1 Sample-paired short- and deep long-read RNA-seq data can robustly test RI detection....70
4.3.2 Intron properties explain similarities across short-read RI detection tool outputs....72
4.3.3 Precision and recall are poor across short-read RI detection tools....75
4.3.4 Short introns and introns that do not overlap exons are more reliably called....77
4.3.5 Persistent introns or called RIs occur in genes with experimentally validated IR....79
4.4 Discussion....80
4.5 Methods....82
Chapter 5: Conclusion....94
5.1 Summary....94
5.2 Future directions....94
5.3 Concluding remarks....95
References....97
Appendix A: Supplementary Figures and Tables....117

Acknowledgements

I would like to thank the following people for their help and support during my degree:

- My advisors Abhi Nellore and Reid Thompson, who took a chance on me and taught me much about genomics, programming, and biology, and have been supportive mentors.
- My dissertation advisory committee chair Jeremy Goecks, members Stuart Ibsen and Zheng Xia and former member Joe Gray, and defense committee member Paul Spellman, for their useful feedback, ideas, and encouragement.
- My collaborators at the Pacific Northwest National Lab, Andy Lin, and ETH Zürich, Gunnar Rätsch, André Kahles, and especially Laurie Prélot, for embarking on a joint project with me (presented in chapter 3 here) and for their rich intellectual and scientific contributions to it.
- Current and former PDXgx group members, who have been good colleagues and collaborators: Chris Anderson, Chris Loo, Sean Maden, Ryan Melson, Austin Nguyen, Janice Patterson, Max Schreyer, Ben Weeder, Mary Wood, and interns Racheal Atanga, Matthew Chang, Zachary Fried, Ellysia Li, and Savitha Srinivasan.
- The Computational Biology program and the Biomedical Engineering department at OHSU, especially JoAnn Takabayashi, Holly Chung, Nermina Radaslic, Erica Hankins Regalo, Alex Breiding, Monica Hinds, and Sandra Rugonyi, and the rest of the staff, students, and faculty for providing a supportive environment.
- The employees and volunteers of the OHSU Office of Visitors and Volunteers, especially Josh Beebe, Desza Dominguez, Rob Wedlake, Andi McNeil, Carol Markt, and Sallie Snyder, who facilitated my participation with Maxwell in the Animal Assisted Therapy volunteer program, which provided context and meaning to my cancer research work.
- My friends and family members, especially Sharyn Campbell, Gordon Campbell, Becky Lambert, Kris Snow, Pat Snow, Geoff Schau, Mary Wood, Crystal Van Dyken, Jason Van Dyken, Ginger Wisseman, Nick Wisseman, Fable Wisseman, Indi Wisseman, Bryan Lung, Jennifer Van Os, Carol David, Rob David, Sharon Kuck, David Kuck, Rachel Luo, Jonathan Kuck, Ryan Kuck, John David, and Maxwell, without whose logistical, emotional, and technical support I would not have completed this work.

List of Figures

Chapter 1:

Figure 1.1: Presentation of neoantigens on the cell surface by MHC I molecules 3

Figure 1.2: Potential types of AS and resulting junctions 7

Figure 1.3: Junction-spanning reads 15

Figure 1.4: Potential confounders of RI detection 17

Chapter 2:

Figure 2.1: Distribution of exon-exon junctions across and within TCGA cancer cohorts 32

Figure 2.2: Clustering by cohort prevalence of shared novel junctions not found in core normal samples 36

Figure 2.3: Junction set assignments and antisense junction prevalence in additional normal tissue and cell type categories from the Sequence Read Archive, across cancers 40

Supplementary Figure S2.1: Distribution and prevalences of TCGA cancer sample junctions 117

Supplementary Figure S2.2: Similarity of TCGA junctions and non-cancer SRA junctions 122

Supplementary Figure S2.3: Distribution of junctions not found in core normal samples and unexplained junctions 124

Chapter 3:

Figure 3.1: Decision points in an ASN detection experiment 48

Figure 3.2: Distributions of 9-mer splice motifs and annotation states 49

Figure 3.3: Effect of filters on remaining 9-mer counts 51

Figure 3.4: Effect of JP filters on final 9-mer count for BRCA sample TCGA-C8-A12P 52

Figure 3.5: Effect of JP filters on validated 9-mer count and validation ratios for BRCA sample TCGA-C8-A12P 54

Supplementary Figure 3.1: Effect of filter steps on remaining 9-mer counts in $\log_{10}$ scale 138

Supplementary Figure 3.2: Effect of JP filters on final 9-mer count for BRCA samples 139

Supplementary Figure 3.3: Aggregate effect of JP filter parameters on filtered 9-mer count across BRCA samples 141

Supplementary Figure 3.4: Proportion and overlap of TCGA-AO-A0JM 9-mers predicted by two pipelines 142

Supplementary Figure 3.5: Aggregate effect of JP filter parameters on validated 9-mer counts and validation ratios across BRCA samples 143

Supplementary Figure 3.6: Effect of JP filter parameters on validated and filtered junction peptide counts across BRCA samples 145

Supplementary Figure 3.7: Effect of JP filters on validated 9-mer count and validation ratios for BRCA samples.....147
Supplementary Figure 3.8: Tryptic peptide counts per sample resulting from the JP and the GP 148
Supplementary Figure 3.9: Aggregate effect of GP filter parameters on validated junction peptide counts and validation ratios across BRCA samples.....149
Supplementary Figure 3.10: Effect of GP filter parameters on validated and filtered junction peptide counts across BRCA samples.....150

Chapter 4:

Figure 4.1: Overview of experimental plan 72
Figure 4.2: Intron persistence and other properties 73
Figure 4.3: Performance of short read tools.....76
Figure 4.4: Properties of introns across detection truth categories.....78
Figure 4.5: Short-read tool performance across nine genes with experimentally validated RIs ...80
Supplementary Figure S4.1: Progression of transcription and splicing.....155
Supplementary Figure S4.2: Short- and long-read coverage of genes by sample 156
Supplementary Figure S4.3: Distribution of intron persistence values for introns in HX1 and iPSC samples 157
Supplementary Figure S4.4: Splicing similarity between samples.....158
Supplementary Figure S4.5 Correlations between intron expression values output by short-read tools.....159
Supplementary Figure S4.6: cRNA overlap of called RIs.....161
Supplementary Figure S4.7: Correlation of short-read expression with intron properties.....161
Supplementary Figure S4.8: Association of intron persistence with transcript position.....162
Supplementary Figure S4.9: Set overlaps of persistent introns and called RIs 162
Supplementary Figure S4.10: Potential vs. called RI set sizes across short-read detection tools 163
Supplementary Figure S4.11: Performance summaries across persistence cutoffs.....163
Supplementary Figure S4.12: Intron properties by truth category for potential called RIs.....164
Supplementary Figure S4.13: Distribution of intron properties for shared false positive introns 164
Supplementary Figure S4.14: Distributions of binned intron properties by truth category.....165
Supplementary Figure S4.15: Intron abundance by truth category across genes with validated RIs 166

Supplementary Figure S4.16: Example length-weighted median expression (LWM) at intron chr1:29053313-29064981 167
Supplementary Figure S4.17: Processing and alignment quality control 168

List of Tables

Chapter 2:

Table 2.1: Junction novelty specification 27

Supplementary Table S2.1: Sources and counts of tumor and tissue-matched normal samples.128

Supplementary Table S2.2: Percent of junctions not found in core normal samples, averaged across cancer types.....131

Supplementary Table S2.3: Selection of additional normal tissue and cell types analyzed.....132

Supplementary Table S2.4: Unexplained junctions occurring in >10% of samples in multiple cancer types.....134

Supplementary Table S2.5: Junction counts and ratio of antisense junctions for TCGA cancer types 135

Supplementary Table S2.6: Genes not currently cancer-associated with high novel junction burdens.....137

Chapter 3:

Table 3.1: Proportion of filtered 9-mers generated by both pipelines 52

Table 3.2: Filter parameter values across pipelines 57

Supplementary Table 3.1: JP annotation and summary across cancer types, for junctions and 9-mers.....152

Supplementary Table 3.2: Summary of JP junction translation across cancer types.....152

Supplementary Table 3.3: Mutual GP and JP junctions and translation across BRCA samples.153

Supplementary Table 3.4: Annotation and motif summary for JP filtered junctions and 9-mers 153

Supplementary Table 3.5: Annotation and motif summary for JP junctions and 9-mers validated in BRCA 154

Chapter 4:

Table 4.1. Short read tools studied.....74

Supplementary Table S4.1: Sample & run availability by platform.....169

Supplementary Table S4.2: Performance metrics for called RIs across persistence thresholds..170

Supplementary Tables S4.3: Counts and performance metrics of called iPSC RIs, called HX1 RIs, potential iPSC RIs, and potential HX1 RIs 171

Supplementary Table S4.4: Properties and sources of experimentally validated RIs studied.....175

List of Abbreviations

AS	alternative splicing
ASN	AS neoepitope
BAM	binary Sequence Alignment/Map
bp	base pair
BRCA	breast cancer
CDS	coding DNA sequence region
CPTAC	Clinical Proteomics Tumor Analysis Consortium
cRNA	circular RNA
DNA	deoxyribonucleic acid
FDR	false discovery rate
FLNC	full-length non-concatemer LR
FN	false negative
FP	false positive
FPKM	fragments per kilobase of exon per million mapped fragments
GBM	glioblastoma multiforme
GP	graph-based pipeline
GTEx	Genotype Tissue Expression Project
GTF	gene transfer format file
hnRNP	heterogeneous nuclear RNP
HX1	human whole blood biological specimen
ID	intron detention
iPSC	human induced pluripotent stem cell line biological specimen
IR	intron retention
JP	junction-based pipeline
LOESS	locally estimated scatterplot smoothing
LR	long-read RNA-seq (in Chapter 4, PacBio Iso-Seq specifically)
LWM	length-weighted median
MHC	major histocompatibility complex
mRNA	messenger RNA
MS	mass spectrometry

NMD	nonsense mediated decay
OV	ovarian cancer
pASN	potential AS neoepitope
PSI	percent spliced in
PTC	premature termination codon
RI	retained intron
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RNP	ribonucleoprotein
snRNA	small nuclear RNA
snRNP	small nuclear RNP
SNS	subset-neighbor search
SNV	single nucleotide variant
SR	short-read RNA-seq
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TMB	tumor mutational burden
TN	true negative
TP	true positive
TPM	transcripts per million
UTR	untranslated region

Abstract

Cancers are among the most deadly and intractable of human diseases, for which the standard of care may involve aggressive yet only partially effective therapies. Immunotherapy, which harnesses the body's immune system to target cancer cells and can lead to long-term remission in some of the most advanced treatment-refractory cases, and liquid biopsy assays, which can potentially detect cancer at an earlier and more easily treatable stage, require the identification of cancer-specific genetic variants. For immunotherapy, genetic mutations leading to non-self proteins can be identified through genomic sequencing, and the immune system primed to target these tumor-associated peptides. RNA-specific variants such as aberrant splicing have the potential to yield peptides that differ significantly from those arising from the germline genome, and accurate identification of cancer-specific splicing in RNA sequencing (RNA-seq) data is an important but complex problem. Identifying tumor-specific mutations requires comparison against a normal background for a patient, but RNA splicing is dynamic and can differ across tissues and time points, so that normal background splicing cannot be identified from a single normal RNA-seq experiment. My work focuses on the identification of aberrant cancer-specific splice variants in tumor samples, leveraging publicly available large-scale RNA-seq data from the Genotype Tissue Expression project, The Cancer Genome Atlas and the Sequence Read Archive. I first focus on calling aberrant exon-exon junctions, and the difficulties of correctly identifying true positive calls that are both true junctions and cancer-specific, while minimizing false positives. I explore the specificity of such calls, finding that many putatively cancer-specific junctions are found in normal samples, though in some cases rarely. I then explore the validity of these calls, querying peptides arising from cancer-specific junctions called by a variety of methods against the sample's intracellular

proteome via mass spectrometry data from the Clinical Proteomics Tumor Analysis Consortium. Novel peptides can also arise from intron retention, where splicing that would normally occur does not, and an intron remains in the resulting processed transcript to be potentially translated. This presents different challenges from identifying exon-exon junctions, as the absence of evidence of splicing such as retained intronic sequence is not the same as the evidence of absence, i.e. that splicing across the junction does not actually occur. I develop a sample-matched test dataset with which I compare the performance of current short read retained intron detection methods against introns with evidence of retention in deep long read sequencing of the same sample, finding that short read detection performance is generally poor. Overall in this work, I show the potential scale of sample counts and types required for a comprehensive normal background of splicing against which truly cancer-specific RNA splicing can be identified; the instability of such identifications and their sensitivity to specific methods and filtering parameter values used; and the poor reliability of intron retention detection from short-read RNA-seq data.

Chapters 2, 3, and 4 are, at least in part, adapted from the following manuscripts:

1. Julianne K. David, Sean K. Maden, Benjamin R. Weeder, Reid F. Thompson, and Abhinav Nellore, "Putatively cancer-specific exon-exon junctions are shared across patients and present in developmental and other non-cancer cells", published in *NAR Cancer* (2020).[1]

2. Julianne K. David*, Laurie Prélot*, Andy Lin, André Kahles, Gunnar Rätsch, Reid F. Thompson, and Abhinav Nellore, "Methods for detection and validation of peptides from cancer-specific splicing", in preparation (2022).

* co-first authors

3. Julianne K. David*, Sean K. Maden*, Mary A. Wood, Reid F. Thompson, and Abhinav Nellore, "Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads", under review (2022).[2]

* co-first authors

Chapter 1: Introduction

1.1 Introduction

In the last decade, much computational genomics research has aimed to analyze cancer genomes and build tools to perform cancer-related immunological tasks, motivated by the growing success of immunotherapeutic treatment of cancers and fueled by the public release of large-scale normal and cancer datasets. These tasks include the search for neoantigens arising from cancer-specific mutations,[3,4] the identification of peptide cleavage sites,[5,6] and the prediction of peptide binding to major histocompatibility complex (MHC) molecules[7–9] and of T-cell receptor recognition of such potential targets.[10–13] Within the large field of computational cancer genomics, this work focuses specifically on the analysis of RNA sequencing (RNA-seq) data, and specifically, the detection and identification of cancer- and sample-specific splicing variants.

Here, I introduce background concepts that provide a framework for this research and a survey of the challenges that motivate it. Section 1.2 gives a brief overview of cancer as a genetic disease and the uses of detecting cancer variants for immunotherapeutic treatment and early detection. Section 1.3 introduces RNA splicing and splice variants, and how splicing can be dysregulated in cancer. Section 1.4 delves into the challenges in identification of aberrant and cancer-specific splicing. Finally, Section 1.5 reviews the challenges and opportunities that my research addresses, and the contributions made to the field in this dissertation.

1.2 Utility of identifying cancer-specific variants

Cancers are a group of genetic diseases, in which genetic mutations lead to phenotype modifications that provide the cancerous cells with advantage over surrounding normal cells, including disproportionate growth, resistance to cell death, and the abilities to stimulate

angiogenesis and to invade healthy tissue.[14,15] While identifying the effects of specific mutations on phenotypes or pathways is complex,[15,16] some of the mutations may make useful targets for fighting the cancer in various ways, including via early detection and immunotherapy.

1.2.1 *Cancer immunotherapy*

The promise of cancer immunotherapy lies in the natural ability of the body to rid itself of harmful cells, either from external sources including viruses and bacteria, or internally produced via somatic genomic mutations.[17] The immune system's mechanisms for recognizing "non-self" cells and targeting them for destruction can in fact recognize and eliminate many tumor cells in early stage cancers.[17] This occurs when antigens are presented on a cell's surface by MHC class I molecules and recognized as foreign by antibodies or mature T-cells (Figure 1.1[18]). Although this mechanism works for some early stage tumor cells, as a cancer tumor progresses it acquires immune tolerance.[19] This takes different forms, including selective removal of immunogenic cells and changes to the tumor microenvironment such as downregulation of immune activation pathways and upregulation of immunosuppressive pathways.[20] A mature cancer tumor, therefore, has very low immunogenicity, allowing for unregulated growth.[18]

The term immunotherapy is broadly applied to multiple types of treatments that reactivate the immune system against cancerous tumor cells, including immune checkpoint inhibitor therapy, in which blockers of immune checkpoints such as cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) or programmed cell death protein 1 (PD-1) antibodies are given to a patient therapeutically to reduce immunosuppression.[21] Toxicity is a risk, especially when CTLA-4 and PD-1 are combined, but finding predictors of efficacy and toxicity is an area of active research.[22–24] T-cells can also be "primed" against the tumor by introducing neoantigens to be

targeted within the newly immune-permissive microenvironment.[25,26] ("Neoepitope" is frequently used as an equivalent term to neoantigen, although technically refers to the segment of a neoantigen to which an antibody binds.) Correct identification of robust target neoepitopes is therefore critical to successful immunotherapeutic response. These must 1) be present in the tumor cells; 2) have a high binding affinity with MHC class I molecules; and 3) differ significantly from epitopes presented by normal tissue cells for non-self T-cell recognition. Epitopes presented on a cancer cell's surface fall into several categories, including tumor-associated antigens (which are still present in lower amounts in healthy tissue), viral antigens for those cancers originating from viral infections, and neoepitopes, or antigens unique to the cancerous tumor arising from mutations of the patient's genome.[27] Of these, neoepitopes are the most immunogenic, giving the strongest anti-tumor immune response[28] and lowest chance of normal tissue toxicity.[29]

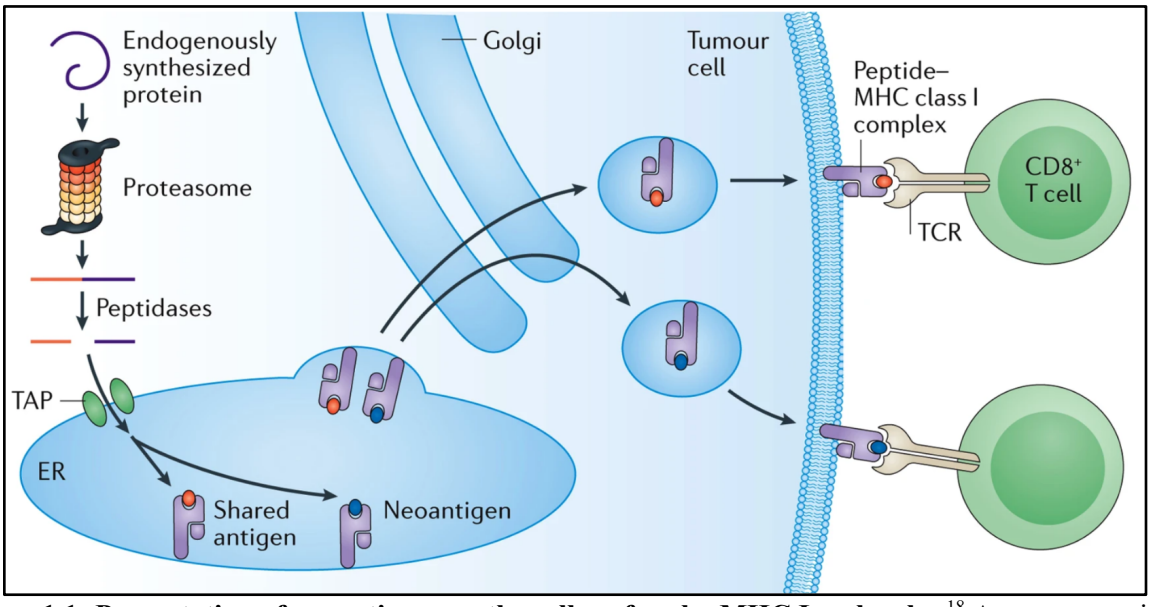


Figure 1.1: Presentation of neoantigens on the cell surface by MHC I molecules.[18] A cancer-specific mutation in an endogenously synthesized protein (upper left, blue) may yield both shared antigens (epitopes, red) and cancer-specific neoantigens (neoepitopes, blue) for potential MHC presentation on the cell surface and recognition by $CD8^+$ T-cells (right). (© 2017 Nature Publishing Group)

1.2.2 *Biomarkers for early detection and other applications*

Identified cancer-specific mutations can also be used as biomarkers for liquid biopsy early detection of cancer, in which an easily obtainable patient sample is examined for potential tumor byproducts. Cancer is generally more easily treatable, with higher life expectancies and likelihood of long term remission, when discovered early, and many cancers do not give rise to concerning symptoms until they have advanced past a more easily treatable stage.[30] Target early detection biomarkers can include circulating tumor DNA, circulating tumor cells, cell-free DNA or RNA, peptides shed by tumor cells, exosomes, and more.[31] Some requirements for such biomarkers are similar to those of immunotherapy targets: they must be highly expressed in cancer, and lowly or not at all expressed in normal tissue. Ideally, they would also be expressed specifically in early stage tumors, and with enough frequency across the larger cancer-type cohort that one or a panel of such biomarkers can be expected to cover a substantial subset of patients.[32] Finally, they must be present in the target sample type (such as blood or urine) at a detectable concentration by being directly released into the blood, or being shed via exosomes or other mechanisms such as apoptosis.[32,33] Genomic biomarkers of cancer may also be used for applications such as prognosis after a cancer has been diagnosed, or to monitor a patient for recurrence after a course of treatment has finished.[32]

1.2.3 *Cancer neoepitope identification*

The majority of cancer-specific neoepitopes arise from missense mutations – single nucleotide variants (SNVs) that cause a three-nucleotide DNA sequence to code for a different amino acid – but around 15% of neoepitopes may result from other types of mutations.[16] Either *in silico* prediction or direct identification can be used to discover targetable, cancer-specific neoepitopes for use in immunotherapy. Direct identification, through mass spectrometry (MS)[34]

or screening of a patient's tumor-infiltrating lymphocytes,[35,36] is time consuming, expensive, and difficult.[37] MS is the easier of the two, but has low sensitivity[38] and still requires detection of differences between cancer and germline DNA sequence, purification and MS analysis of MHC-bound neoepitopes from a sufficiently large tumor sample, and comparison between the MS and genomic results.[39] Given these challenges, several *in silico* prediction tools have been developed, which search the cancer genome for mutated sequence that differs from the normal genome and may, when translated, produce neoepitopes. Many prediction tools, however, fail to search comprehensively across all possible somatic mutation types, searching for neoepitopes resulting from missense SNVs only, or possibly also insertions and deletions (indels)[3,4,40] or gene fusions.[40] SNVs and other DNA-specific variants can be easily identified by comparing DNA sequencing of a cancer sample against DNA sequence from a neutral, non-cancerous site in the body such as blood.[41] However, neoepitopes arising from nonsynonymous missense SNVs differ from a normal self peptide by only one amino acid and may not result in a strong immune response, as immunogenicity is related to dissimilarity from the normal proteome.[42] Neoepitopes arising from indels (especially those causing a shift in reading frame[43]) and gene fusions[44] may be more immunogenic, but overall, searching for genome-level variants in DNA sequencing yields only a subset of potential immunogenic cancer variants.[45]

1.3 RNA splicing and splice variants

1.3.1 *RNA and RNA splicing*

The historically described "central dogma" of molecular biology holds that DNA is the storage mechanism for genetic information; that DNA is transcribed into RNA, which determines what genetic information will be used and how; and that RNA is translated into protein products which perform biological functions.[46] Although this somewhat limited view of

how genetic information is stored and used has been updated with the advent of new technologies,[47,48] and it is now known that RNA has a variety of cellular functions and less than 1/3 of human genes are protein-coding (19,954 out of 60,656 genes in GENCODE v.35[49]), this basic process still describes one of RNA's most common roles.[50]

Of primary interest here is one of the ways in which RNA is used to determine how genes are used, namely the creation from a single gene of different isoforms or transcripts (often interchangeably referring to different forms that a gene can take in mature RNA). These can differ in various ways, including the start or stop codons that indicate the region that will be translated to protein, but of greatest interest here is alternative splicing of the gene to yield different isoforms.[51–53] A precursor, or immature, RNA molecule will contain the entire sequence from its gene's DNA, comprising three categories: 1) untranslated regions (UTRs); 2) exons, or expressed regions that will remain in the mature RNA; and 3) introns, or intragenic regions that are spliced out (physically removed) from the sequence during the co-[54–59] or post-[60–62] transcriptional RNA processing. The inclusion of different exons and splicing out of different introns within a given gene in varying combinations is called alternative splicing (AS), and leads to variable functionality of the resulting isoforms. In the case of protein-coding genes, alternative isoforms' mature messenger RNA (mRNA) will have different functionality when translated into protein.[63,64]

AS and the generation and use of different isoforms from a single gene regularly occurs in normal human biology.[66] Over 90% of human genes undergo AS[67,68] and 86% of genes have a minor isoform that is significantly expressed (>=15% frequency).[67] AS and the related expression of certain isoforms over others is frequently associated with tissue type[67], sex[69–71] or developmental or life stage,[72,73] and other biological states such as disease,[72,74] whereas splicing

variation between individuals is much less common.[67] The specific type of AS described above is that of differing exon-exon junction (hereafter, "junction") usage, where a junction refers to a set of splice sites at the edges of two exons from between which an intron is removed (Figure 1.2[65]).

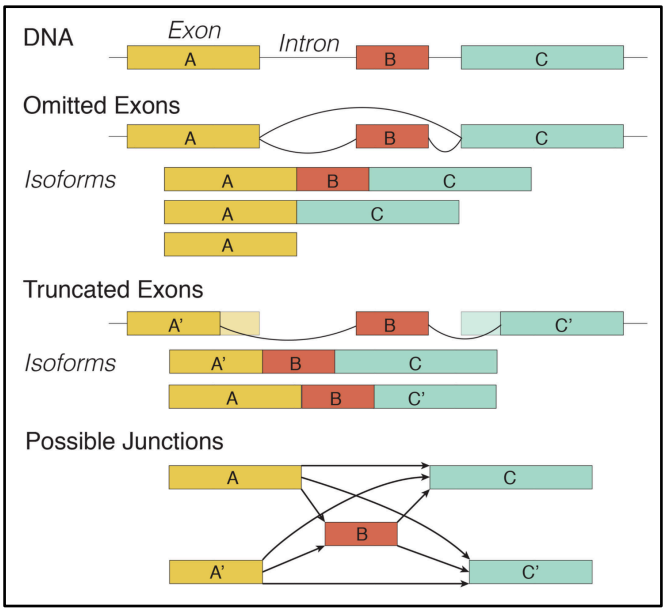


Figure 1.2: Potential types of AS and resulting junctions.[65] Exon A' has an alternate 3' splice site vs. exon A, and C' an alternate 3' splice site vs. exon C. Exon B is shown included or skipped. (© 2016 IEEE)

1.3.2 *Intron retention*

In addition to varying junction use, another form of AS is intron retention (IR), in which intronic sequence that is normally spliced out instead remains in the RNA sequence as a retained intron (RI).[75] IR sometimes occurs in normal tissue; about 10% of almost 230,000 total transcripts covering over 60 thousand genes in GENCODE v.35 annotation have an IR flag, and IR is found to affect 80% of protein coding genes in humans[76] and 51-77% of genes in other vertebrates.[75] IR was originally understood as a transcriptional processing mistake[77] with no functional use, since it will often lead to nonsense transcripts that quickly undergo nonsense mediated decay (NMD).[78] However, the current understanding is that IR coupled with NMD can be used in a cell as a tool for mRNA localization[79] and for regulating gene[80–87] or transcript[80,88]

expression, splicing,[89] and response to stress such as hypoxia [90] or other external stimuli.[91,92] Furthermore, although many transcripts containing retained introns contain premature termination codons (PTCs) and undergo NMD or other degradation, some escape NMD and yield alternative protein isoforms.[79] (Up to 10% of potential NMD candidates that contain PTCs may escape NMD and be translated.[93]) Transcripts containing retained introns may also be detained in the nucleus ("intron detention", or ID); some stable ID transcripts may be used to control gene expression over time.[94]

1.3.3 *Regulation of alternative splicing*

Splicing is a multi-step process involving the breaking of chemical bonds between the two ends of the intron and their neighboring exons and the forming of a new chemical bond between the two exons. This process is catalyzed by a ribonucleoprotein (RNP) complex called the spliceosome, comprising five small nuclear RNPs (snRNPs).[95] Each snRNP contains a small nuclear RNA (snRNA) (U1, U2, U4, U5, and U6, after which their corresponding snRNPs are named)[95] in complex with a number of associated protein factors.[96] Spliceosomes may be recruited to a nascent unprocessed RNA as soon as transcription has been initiated.[97] Splicing of a given intron may begin as soon as its 3' splice site has been transcribed, and often proceeds rapidly.[59,98] Most splicing in humans has been found to occur cotranscriptionally,[54–58] although up to 20% of splicing may occur post-transcriptionally.[54,58,60–62] Since splicing only begins after the full intronic sequence is available, an early hypothesis about the order in which introns are removed was the "first come first served" pattern of splicing, where introns at the 5' end of the transcript are more likely to be spliced out first, having been transcribed first.[99] Additional evidence for the "in-order" splicing effect is that splicing of a given intron is more likely to occur quickly after a neighboring intron is spliced.[100] However, splicing has been shown to occur "out

of order"[101–103] and is now understood to be a stochastic process.[104,105] Splicing may follow the same order of introns in certain transcripts,[106] with the caveats that variation from this order occurs more frequently in humans than other animals, and that the predicted orders have been validated only with synthetic data.[107]

Splicing regulation falls under two categories, cis- and trans-acting, i.e. performed by regulators located physically in the same region as the splicing targets (cis), and by those physically remote from the target region (trans). Cis-regulatory elements include splicing silencers and enhancers located in the intron to be removed or in the exons flanking it, and comprise specific short base sequences that affect splice site selection. Exonic splicing enhancers act to include the exon in which they are located,[108] whereas exonic and intronic splicing silencers lead to exclusion of an exon from the final spliced transcript.[109] Intronic splicing enhancers promote the removal of the intron in which they are located and are commonly correlated with presence of nearby exonic splicing enhancers.[110] When located in a region that may either be removed as an intron or remain in the processed transcript as an exon with splicing occurring at an alternate 5' or 3' splice site, these can act as splicing suppressors, inducing the removal of the potentially-exonic sequence in which they are located[110] rather than inclusion of that exonic region in the transcript.

Cis-regulatory elements can then act as binding sites for trans-acting regulatory elements, or splicing factors, such as heterogeneous nuclear RNP (hnRNP) complexes[111] and serine/arginine-rich proteins[112] which act in concert to regulate splicing.[113] These proteins generally bind to splicing enhancer sequences and promote the inclusion of associated exons by stimulating the initial steps of the spliceosome assembly,[114,115] while hnRNPs generally bind to splicing silencer sequences, physically blocking splice sites, leading to exon skipping.[116] The

noncoding spliceosome component U1 (an snRNA) is responsible for recognizing and binding to 5' splice sites.

A number of genomic features or metagenomic modifications can lead to IR. It can occur as the result of a histone modification[117] that destabilizes the spliceosome and represses RNA Pol-II elongation.[75,118,119] It can also be associated with genomic regions with high intronic CG content that can form secondary structures that may lead to the spliceosome pausing over introns, or to reduced binding of splicing enhancers.[77,120] Regions with dense CG content may also have reduced methylation and reduced recruitment of splicing factors to the unprocessed RNA.[121–123] In other cases, weak splice sites with less conserved splicing motifs and weak polypyrimidine tracts, which promote spliceosome assembly, may not be as readily recognized by the spliceosome and may lead to increased IR.[120,124] Short introns are also likelier to be retained, as they have less availability of splicing factor binding sites.[120,124] An enrichment of RNA protein binding sites in an intronic region can also lead to increased IR, via an increase in binding of splicing repressors.[125] Finally, reduced splicing factor expression can lead to higher IR.[126,127]

1.3.4 *Splicing dysregulation in cancer*

Splicing is known to be "noisier" in cancer than in normal tissue,[128,129] and often comprises alternative splicing that is cancer-specific or different from expected normal alternative splicing, generally referred to as "aberrant."[130] Aberrant cancer-specific splicing has been associated with many cancer types, with functional effects on disease progression and response to treatment. In colorectal carcinoma, aberrant splicing related to protein kinase activity and signaling pathways has been observed.[131,132] Alternative or aberrant splicing signatures may improve prognosis predictions in pancreatic,[133] ovarian,[134–136] and cervical[137,138] cancer, as well as in hepatocellular carcinoma,[139] in which dysregulated splicing factor expression is also

implicated in cancer development.[74] AS in *ERBB2*,[140] *CD44*, and *TP53*[141] has been shown to promote breast cancer tumor growth[142,143] and metastasis.[140] In several cancer types, specific AS isoforms are associated with both oncogene activation and loss of function of tumor suppressors; for instance, a skipped exon in *TNR6* can inhibit Fas-mediated cell death in breast[144] and uterine[145] cancers and in leukemia.[146] Castration-resistant prostate cancer may take advantage of androgen receptor splice variants to promote growth in low-androgen conditions,[147] mediating resistance to treatment.[148,149]

High levels of IR have also been found across a wide range of cancer types[150–152] and may contribute to the inactivation of tumor suppression.[153] IR has known effects in cancer, such as regulation of gene expression,[125] and probable ones, such as nonsense-induced transcriptional compensation and improved cell survival in nutrient-poor environments.[125] Accurate detection of retained introns is also important for cancer immunotherapy as they may be a source of tumor neoepitopes,[154–156] and RI neoepitope burden is correlated with prognosis in multiple myeloma,[157] although current methods for calling RI neoepitopes may be inconsistent: RI neoepitope prediction on the same cell lines by two independent studies yielded different neoepitopes.[154,155]

The causes of splicing dysregulation in cancer also can be categorized under cis- and trans-acting effects. The former comprise mutations in splice site motifs or in local splice regulatory elements physically close to the dysregulated splice site in the genome. A new 5’ or 3’ cryptic splice site may be directly generated,[158] or a normal splice site eliminated,[153,159] by a mutation. A polypyrimidine tract may be weakened,[160] or a mutation may occur in an exonic or intronic splicing enhancer or silencer sequence[161] changing observed splicing patterns.[162]

Trans-acting dysregulation can be more complex, with remote mutations or cancer-specific pathway changes leading to new splicing patterns at locations distant from the original mutation. Spliceosome components are frequently mutated in cancer,[163] including *SF3B1* in chronic lymphocytic leukemia,[164,165] *U2AF1* and *SRSF2* in myelomonocytic leukemia and all three in myelodysplastic syndromes.[166–169] Mutations in *SF3B1* can induce the use of cryptic 3′ splice sites[170] in breast cancer,[171] uveal melanoma,[172] and chronic lymphocytic leukemia,[173] while mutations in the snRNA spliceosome component U1[174] can lead to cryptic 5' splice site recognition.[175] Even without direct mutations to spliceosome components, they may undergo changes in expression in cancer cells, leading to changes in splicing.[176] The presence of binding motifs for the splicing factor RBM9 has been shown to cause breast cancer subtype-specific AS.[177] Overexpression of polypyrimidine tract binding proteins can lead to increased AS in ovarian,[178] colorectal,[179] and bladder cancers,[180] possibly by impairing autoregulation of SRSF3.[181] Finally, genes other than spliceosome components can also affect splicing in cancer, such as the cytoplasmic adaptor *CRKL*, a participant in signaling pathways that regulate alternative splicing in cervical cancer.[182]

1.4 Identification of aberrant splicing from RNA-seq data

The identification of novel, sample-specific splicing in cancer can be useful for elucidating the biology of disease[183–185] and identifying potential treatment targets.[154–156,186] The search for neoepitope targets arising from splicing variants for immunotherapy treatment has been an area of active recent research.[154–156,186] Novel junctions may create downstream frame shifts that yield codon usage not seen in normal tissue, or may include a 3' splice site in what would normally be an intron or a UTR. Likewise, novel IR can also lead to transcribed sequence not translated under normal circumstances.[125] In all cases, if the novel transcript escapes NMD

and is translated into a protein product, the resulting sequence may differ significantly from sequence represented in the normal human proteome. (Up to 10% of transcripts with PTCs may be translated instead of undergoing NMD.[93]) As noted above, peptides significantly different from normal human peptides are more likely to be immunogenic and to stimulate an immune response against the novel peptide generating cell.[42]

1.4.1 *Challenges of cancer-specific splicing identification*

Potential cancer-specific splicing variants have only in recent years begun to be studied because they are much more difficult to identify than variants in DNA. The sequence of a person's DNA is largely static across a person's lifetime; it may undergo mutations due to viral infections[187], exposure to radiation[188] or some chemicals,[189] or duplication errors[190] although the cell has mechanisms to repair most duplication errors before they become permanent, through proofreading or mismatch repair.[191] Unrepaired DNA mutations may accumulate in the body over time, but only very slowly and only in the direct lineage of the cell in which the mutation originated.[192] If an accumulation of these mutations leads to cancer,[193] the DNA mutations occur only in the tumor cells, and DNA from other tissues and cells remains an appropriate baseline for identifying cancer-specific DNA mutations: the sequenced tumor DNA can be compared against a non-cancerous sample from the same patient to identify tumor-specific variants.[3,4]

However, because RNA commonly undergoes AS across tissues and at different times in one organism, the "normal" background cannot be defined by sequencing a single sample of the patient's blood or other normal tissue, as it can for the static DNA. Even using several samples or GENCODE annotation as "normal" will not fully represent the splicing breadth and diversity that can occur in normal tissues.[194] What has allowed these cancer-specific splicing studies to move forward is the development, in the last decade, of several large-scale RNA-seq data

collection projects. These include the Genotype-Tissue Expression Project (GTEx)[195] with nearly 10,000 samples covering 29 normal tissues from hundreds of donors, and The Cancer Genome Atlas (TCGA),[196] with over 11,000 tumor samples of 33 cancer types. The public release and uniform processing[197,198] of these data have allowed for various large-scale analyses of "normal" and cancer splicing.[154–156,186]

However, despite the advent of large sets of normal-tissue RNA-seq data, it remains difficult to define what truly comprises normal splicing, and by extension, what should be regarded as cancer specific. For novel junction identification, two questions are outstanding: 1) what can be positively identified as a real splice site? And 2), out of the set of real splice sites, what should be categorized as "normal" splicing and not included as cancer-specific?

1.4.2 *RNA sequencing, splice-aware alignment, and splice site identification*

Addressing the first question above requires an understanding of how splice sites are identified from short RNA-seq reads, which are generally in the range of 100-150 bases long. To determine splicing within a sample, the short reads are aligned to a reference genome via a splice-aware aligner, such as hisat2[199] or STAR.[200,201] Such an aligner will potentially accept large gaps in the alignment of a given read, where the read ends each map to separate exons and the gap corresponds to a spliced-out intron in between (Figure 1.3[65]).[200–202]

Splice-aware alignment can be done 1) with the help of a known set of annotated transcripts,[200] such as those curated by GENCODE[49] or ENSEMBL;[203] 2) via de novo junction identification, where the aligner attempts to find the best alignments without reference to known annotated isoforms; 3) in a two-pass system where an initial annotation-based alignment attempts to identify novel spliced alignments for reads that align neither to unspliced sequence nor to

annotated junctions, and then a second alignment uses as input both annotated junctions and those identified in the first-pass alignment;[200,201] or 4) via a hybrid method, where novel junctions discovered as the alignment is performed are used to inform alignment of the remaining reads.[202] (Many splice-aware aligners[200,202] are capable of running in more than one of these modes, allowing the user to choose one- or two-pass, or annotation informed or agnostic, alignment.)

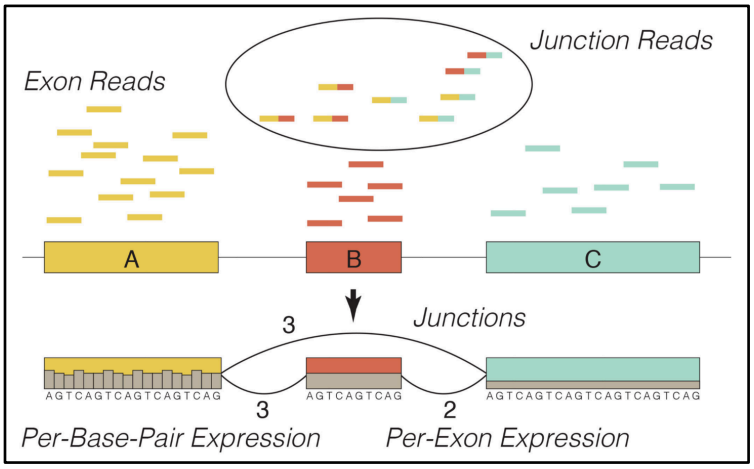


Figure 1.3: Junction-spanning reads.[65] A splice-aware aligner will map some reads to single exons ("exon reads", yellow, red, teal), and others ("junction reads," multicolored in the oval) with each end to separate exons, with spanning intronic sequence not included. (© 2016 IEEE)

One source of error in junction identification from short-read RNA-seq arises from mistakes and noise in sequencing, which can lead to lower alignment accuracy. While many improvements have been made in next generation short-read sequencing, errors still may occur, and public data from older platforms has a potentially higher rate of errors. Median error rates on Illumina platforms range from 0.1-0.6%,[204] with higher error rates occurring in certain sequence contexts.[204]

Mistakes may also occur in alignment, even in the absence of sequencing errors, especially for junction-overlapping reads. For instance, it can be hard to determine the correct alignment of a potentially very short split read segment within a highly repetitive genomic

region.[205] Some alignment decisions may affect error rates, such as the choice of final alignment for multimapping reads[206] and the allowed anchor length. Anchor length is the minimum length of a segment into which a full read is split in splice-aware alignment. An anchor length set to 5 may allow more reads to be aligned, since a junction may well occur close to the end of the read, but the 5-base segment could be difficult to align accurately; conversely an anchor length of 15 would allow for more accurate alignment of the short end of the read, but fewer reads aligned overall. Confidence in a short-read identified junction can be increased by observing multiple reads aligned to the junction either within the same sample, or across samples generally or within a specific sample-type cohort. This does not guarantee that the junction is biologically meaningful, but does reduce the likelihood that it is an artifact of a sequencing or alignment mistake. Junctions can also be identified from long-read RNA-seq data, which provides higher confidence in alignment and deeper transcriptional context, although the current significantly higher cost of long-read sequencing reduces the frequency of long-read use.[207]

1.4.3 *Additional challenges in intron retention detection*

As noted above, in addition to identifying changes in observed splice sites, there has been interest in detecting intron retentions, where splicing that should normally occur does not and the intron remains to be translated in the resulting processed transcript. In theory, RI detection from RNA-seq data may seem straightforward: if intronic sequence that should be spliced out is instead found in data aligned with a splice-aware aligner, especially sequence running from an exon into an adjacent intron, this is evidence of an RI. However, in practice it is not so simple, and RIs are significantly more difficult to identify than junctions, where (with exceptions for uncertainty with multimapping or low confidence alignments) a single read clearly maps to two separate exons with a large intronic gap in between. In RI detection, seeing no evidence of

splicing is not the same as definitively proving that splicing would not actually occur, so simply identifying reads aligned to an intronic region is not sufficient for a confident RI call.

There are two major challenges that lead to significant uncertainty in the quality of called intron retentions in any given sample. The first is that the coordinates of any given intron often overlap other non-intronic features (Figure 1.4[208]), such as exonic sequence from another transcript in the same gene, exons from an antisense transcript in the same genomic region, other overlapping features such as snoRNAs, and novel, unannotated exons that may be part of the true transcript from which the RNA-seq read arose.[208] Short RNA-seq reads in the range of 100-150 bases long are generally not long enough to give a true understanding of their transcript context when aligned to a reference genome; if a short RNA-seq read maps to the middle of an intronic sequence that overlaps one or more of the features mentioned above, it is hard to determine its overall splicing context or the original transcript molecule from which it was sequenced.

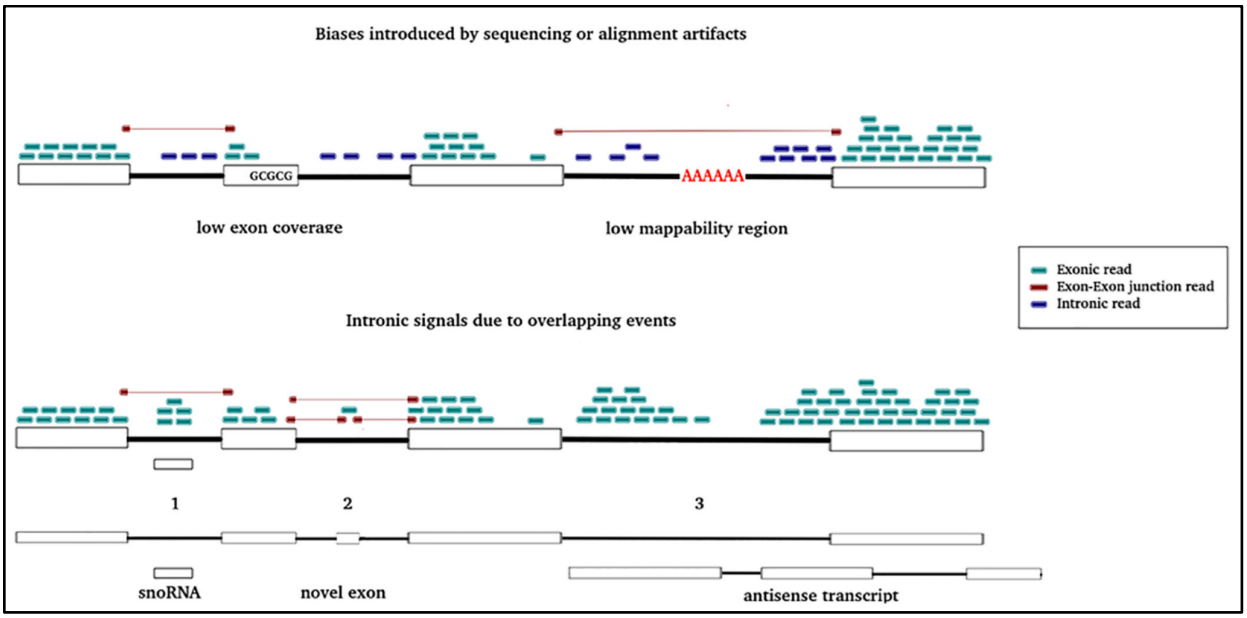


Figure 1.4: Potential confounders of RI detection.[208] These include overlapping or antisense features, the presence of novel unannotated exons, and low mappability regions. (© 2020 Lucile Broseus and William Ritchie)

The second challenge is that RI detection can be hampered by pre-mRNA or DNA contamination of the physical sample in an RNA-seq experiment, or in the case of public data use, the potential mislabeling of whole RNA-seq (unspliced) data as mRNA-seq (spliced). Any unspliced molecules will have a high rate of supposed IR, without giving a true representation of the sample's IR. Some spliced introns may also form circular RNAs, which may be detected in RNA-seq data.[209] (Additional uncertainty in RI detection occurs in cancer samples, in which cancer-specific transcript processing such as cancer-specific polyadenylation may occur.[210,211])

Several tools have been developed specifically to detect IR, including keep me around (KMA),[212] IRCall,[213] IRFinder,[76] IntEREst,[214] iREAD,[215] superintronic,[216] and IRFinder-S,[217] and which attempt to address these challenges. To handle potential confounding by overlapping features, many tools detect introns only across "measurable" intronic regions that do not overlap other annotated features, instead of across full annotated introns. Furthermore, many tools rely on poly(A)-selected library preparation to ensure that the sequenced RNA was spliced.[76,215,216] In many cases, the assumption that poly(A)-selected data represents fully processed, mature RNA may be acceptable, although there is strong evidence that intronic sequence is commonly found in poly(A)-selected RNA-sequencing experiment data.[54,75] Overall, these solutions have significant limitations: importantly, 1) the first leaves many potentially retained introns out of the scope of detection entirely, and 2) the stochastic nature of the splicing process and the significant proportion of transcripts that are post-transcriptionally spliced[54,58] after polyadenylation increases the chance of false positive RI detection from poly(A)-selected data.

1.4.4 *Defining "normal" splicing*

Altogether, the question of what splicing is aberrant or cancer-specific hinges on what is conversely defined as "normal." Even within the context of neoepitope identification, there is no

consensus on what should be included in normal splicing. This can span from the relatively small number of junctions annotated in GENCODE,[218] or a handful of paired normal samples,[154] to the full breadth of all normal samples attainable, for example across all of GTEx and the Sequence Read Archive (SRA),[194] but is often defined somewhere in between. Many studies of retained introns in cancer samples only take into account the background retention common in normal tissues in a limited way, with only one[150,151] or a handful[154] of non-tumor samples as the normal comparator, or do not compare against a normal IR background at all.[153,157] Most available tools for detecting retained introns allow only for comparison against a small number of samples,[76,212–216] with SplAdder[219] as a notable exception, having been applied across all tumor RNA-seq samples from TCGA[196] with ~10,000 normal RNA-seq samples from GTEx[195] as the normal comparator.[156] One other study[155] has identified putatively cancer-specific neoepitopes arising from IR using GTEx, in the form of CHESS annotation,[220] as the normal comparator.

For cancer-specific junction identification, the current standard seems to converge on using a well-curated and large set of normals, namely GTEx,[155,156] as the normal comparator. However, even with a defined set of samples assigned as normal, several questions remain, such as the acceptable leniency of the normal filter. If a junction or retained intron is detected in a single read in a single normal sample, is that sufficient to label it a product of normal splicing? If not, with what frequency is it allowable in the normal sample set? There likely will not be a single answer, with instead the ultimate goal of a given aberrant splicing detection experiment driving the leniency of its definition of normal. An application such as neoepitope identification for immunotherapy treatment, where cross-reactivity may cause negative side effects or in rare cases patient death,[221–223] may need a stricter definition of normal than early cancer detection, where significant enrichment in a patient sample may be sufficient for an accurate result; a false

positive detection would produce some hardship for the patient in the form of mental distress and additional testing, but not risk of death. The question of transcript abundance in RNA compared to peptide abundance in the translated product is also important; if the end goal is the peptide product, such as in neoepitope identification, a low level of RNA expression of an isoform will not necessarily correspond to low expression of the resulting protein[64,224] or to low presentation by MHC on the cell surface.[225]

1.5 Summary

1.5.1 *Challenges and opportunities*

The accurate identification of aberrant splicing in cancer samples is of increasing importance, especially for applications to early detection and neoepitope identification for cancer immunotherapy. Steps have been made to improve the accuracy and precision of this identification, but much remains challenging. Notably, detecting legitimate sample- or cancer-specific junctions in noisy splicing data with high confidence requires a delicate balance between demanding evidence for the junction's existence while also ensuring that it does not fall in the realm of normal splicing. Establishing a definition of what is "normal" is also difficult, and likely will practically depend on the final application of detected junctions. Retained intron detection presents unique challenges in extracting signal from data of uncertain context and origin; leveraging information about the processing & splicing progression of transcripts and patterns of splicing within genes and transcripts may help to inform and improve RI detection. One of the primary problems of RI detection, and of cancer-specific junction identification, is the lack of known ground truth. The generation of large RNA-seq data sets; multi-omics experiments with paired long- and short-read RNA-seq or proteomics and RNA-seq assays on the same

physical sample; and the increasing availability of powerful computational resources has opened opportunities to address these challenges.

1.5.2 *Contributions*

In this work, I address the questions introduced above: In the identification of cancer-specific splicing, how can we be confident both in the true biological reality of an aberrantly spliced transcript, and in its cancer-specificity? I use large public sets of short-read RNA-seq data to probe the cancer-specificity of junction identification, paired proteomics data to address the existence of peptide products from called novel junctions, and sample-paired short- and long-read RNA-seq dataset to show the full transcript and splicing context of potential detected IR from current short-read detection tools. I show the limitations and potential scale required for a true normal background of splicing against which novel or aberrant RNA splicing can be accurately identified; the sensitivity of such identifications to specific filtering method and parameter values chosen; and the quality of current methods for detecting novel intron retentions from short-read RNA-seq data as well as issues that must be overcome to accurately detect these going forward.

Chapter 2: Exploring the cancer-specificity of tumor junctions

This work has been formatted for inclusion in this dissertation from the manuscript "Putatively cancer-specific exon-exon junctions are shared across patients and present in developmental and other non-cancer cells" by Julianne K. David, Sean K. Maden, Benjamin R. Weeder, Reid F. Thompson, and Abhinav Nellore, published in *NAR Cancer* (2020).[1] The author of this dissertation is the primary author of the manuscript.

2.1 Abstract

This study probes the distribution of putatively cancer-specific junctions across a broad set of publicly available non-cancer human RNA-seq datasets. We compared cancer and non-cancer RNA sequencing (RNA-seq) data from The Cancer Genome Atlas (TCGA), the Genotype-Tissue Expression (GTEx) Project, and the Sequence Read Archive (SRA). We found that: 1) averaging across cancer types, 80.6% of exon-exon junctions thought to be cancer-specific based on comparison with tissue-matched samples (σ = 13.0%) are in fact present in other adult non-cancer tissues throughout the body; 2) 30.8% of junctions not present in any GTEx or TCGA normal tissues are shared by multiple samples within at least one cancer type cohort, and 87.4% of these distinguish between different cancer types; and 3) many of these junctions not found in GTEx or TCGA normal tissues (15.4% on average, σ = 2.4%) are also found in embryological and other developmentally associated cells. These findings refine the meaning of RNA splicing event novelty, particularly with respect to the human neoepitope repertoire. Ultimately, cancer-specific exon-exon junctions may have a substantial causal relationship with the biology of disease.

2.2 Background

Aberrant RNA splicing is increasingly recognized as a feature of malignancy,[176,183,226–228] potentially driving cancer progression[227] and with potential prognostic significance across many

cancer types including non-small cell lung cancer, ovarian cancer, breast cancer, colorectal cancer, uveal melanoma, and glioblastoma.[134,229–233] Due to its potential for generating novel peptide sequences, aberrant RNA splicing is also interesting as a potential source of neoantigens for cancer immunotherapy targeting.[158] For instance, retained intronic sequences can give rise to numerous potential antigens among patients with melanoma, although they are not a significant predictor of cancer immunotherapy response,[154] and a patient-specific neoantigen arising from a gene fusion has been shown to lead to complete response from immune checkpoint blockade.[44] Novel cancer-specific exon-exon junctions have also been shown to be a source of peptide antigens,[156,186] and represent compelling potential targets for personalized anti-cancer vaccines.[45] However, the ability of the adaptive immune system to target a given antigen as "foreign" depends on a complex prior tolerogenic education, and in particular on whether or not a given antigen has been previously "seen" by the immune system in a healthy context.[234] Therefore, prediction of cancer-specific antigens depends explicitly on their sequence novelty, and thus requires a comparison with non-cancer cells.

Choosing a "normal" tissue standard for comparison is difficult in the context of RNA-seq data analysis, given the presence of alternative splicing throughout normal and cancerous biological processes.[176,235,236] Previously, cancer-specific aberrant splicing has been detected by comparing tumor RNA-seq data against a single reference annotation[218] or a limited "panel of normals".[154] A TCGA network paper[156] used the large publicly available datasets of TCGA[196] and GTEx[195,196] to identify and validate thousands of novel splicing events including exon-exon junctions present in a specific TCGA cancer type but not in the corresponding normal adult tissue in GTEx. This study also predicted alternative splicing neoepitopes via this comparison, and validated several of these neoepitopes shared between multiple patients with the intracellular

proteomics data available for select ovarian and breast cancer TCGA donors in the CPTAC dataset.[156] More recently, another study has leveraged TCGA and GTEx, as well as cell line data, to discover and validate neoepitopes derived from alternative splicing.[186]

Here, we propose that the comparison of cancer junctions with only matched-normal GTEx tissue data allows a significant number of junctions to be erroneously identified as cancer-specific, and that GTEx provides neither an appropriately specific nor a fully comprehensive standard for normal splicing comparison. We investigate the sharedness of cancer junctions within and across cancer-type cohorts, and their presence across multiple normal cell and tissue types, including cohorts representing diverse developmental stages and potential cell types of cancer origin.

2.3 Methods

Data Download

Previously called exon-exon junction data including phenotype table, BED and coverage files for both TCGA and GTEx v6 were downloaded from the recount2 service at https://jhubiostatistics.shinyapps.io/recount.[197] These data were previously extracted[198] from RNA-seq experiments encompassing 10,549 tumor samples across 33 TCGA cancer types, 788 paired normal samples across 25 TCGA cancer types, 9,555 normal samples across 30 GTEx tissue types (Supplementary Table S2.1). recount2 used Rail-RNA[201] to align RNA-seq samples, and all command-line parameters affecting alignment are referenced in supplementary information from the recount2 paper.[198] The metaSRA[237] web query form at http://metasra.biostat.wisc.edu/ (a tool for identifying SRA samples of interest) was queried for experiment accession numbers for 1) non-cancer cell and tissue type samples (see Supplementary Table S1 for cancer-matched samples and Supplementary Table S2.3 for non-cancer samples,

and "Comparison with SRA tissue and cell types" for a description of how these samples were chosen) and 2) TCGA-matched cancer types (see Supplementary Table S2.1). For the non-cancer samples, the term "cancer" was explicitly added as an excluded ontology term in the query, and the resulting files were filtered to remove any samples with "tumor" in the sample_name field. The resulting accession numbers represent 12,231 human samples from the SRA, specifically 10,827 samples from 33 normal tissue and cell types and 1,404 samples from 14 cancer types (Supplementary Tables S2.1 and S2.3). These accession numbers were queried against the Snaptron junction database using the query snaptron tool (for interfacing with uniformly-extracted recount2 junctions).[198,238] This query yielded junctions also previously extracted by recount2 with the same pipeline used for the GTEx and TCGA samples for the tissue and cell types of interest,[198] which were subsequently downloaded. TCGA tumor mutational burden (TMB) data (file mutation-load-updated.txt) were downloaded from https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin.[239] Patient somatic mutation calls were downloaded from the GDAC firehose,[240] while a list of human splicing-associated gene mutations (keyword search "mRNA splicing [KW-0508]") was downloaded from the UniProt database.[241] Two lists of cancer-associated genes were downloaded: the COSMIC cancer gene census cancer gene list from https://cancer.sanger.ac.uk/census,[242] and the OncoKB cancer gene list from https://oncokb.org/cancerGenes.[243]

Indexing of GTEx and TCGA junctions

The GENCODE gene transfer format (GTF file was parsed to collect full coordinates and left and right splice sites of junctions from annotated transcripts and a searchable tree of protein-coding gene boundaries. The GTEx phenotype file was parsed to collect tissue of origin information and donor gender; bone marrow samples derived from leukemia cell line cells were

eliminated. The TCGA phenotype file was parsed to collect information on cancer type, cancer stage at diagnosis, patient gender, vital status, and sample type (primary tumor, matched normal sample, recurrent tumor, or metastatic tumor). Cancer subtype classifications were collected for five cancer types beyond their TCGA designations (Figure 2.2B, Supplementary Table S2.1): cervical squamous cell carcinoma and endocervical adenocarcinoma was separated into cervical squamous cell carcinoma, endocervical adenocarcinoma, and cervical adenosquamous; esophageal carcinoma was separated into esophagus adenocarcinoma and esophagus squamous cell carcinoma; brain lower grade glioma was separated into astrocytoma, oligoastrocytoma, and oligodendroglioma; sarcoma was separated into leiomyosarcoma, myxofibrosarcoma, malignant peripheral nerve sheath tumors, desmoid tumors, dedifferentiated liposarcoma, synovial sarcoma, and undifferentiated pleomorphic sarcoma; and pheochromocytoma and paraganglioma were separated. A new SQLite3 database was created to index all GTEx and TCGA junctions, with linked tables containing 1) sample ids and associated junction ids; 2) sample ids and phenotype information for each sample; and 3) junction ids and junction information including 0-based closed junction coordinates, GENCODE annotation status, and location within protein coding gene boundaries. SQL indexes were created on junction ID and sample ID columns for fast and flexible querying.

Selection of cancer-specific junction filters

For all analyses we apply a light filter, requiring a junction to have at least a two-read coverage across GTEx, TCGA, and the selected cancer and non-cancer SRA samples, to exclude false positive junctions but allow for the existence of splicing noise; we do not require a minimum read count per sample. To characterize junction novelty in cancer with respect to normal cells, we defined a hierarchical filter that specifies inclusion and exclusion of junctions in

different RNA-seq datasets (Table 2.1). In order from most to least permissive, these filters are: 1) junctions not found in tissue-matched GTEx or TCGA normal samples, 2) junctions not found in any GTEx or TCGA normal ("core normal") samples, and 3) junctions not found in any core normal samples or in selected SRA tissue and cell type non-cancer samples. For our analyses, we do not explicitly filter on whether a junction is annotated in GENCODE. We do not set a limit on presence in the core normal sample cohorts: any junction present at any coverage level in only one sample is counted as "in" these cohorts. This yields a more stringent filter on normality than that used by the TCGA splicing paper, which uses the term "neojunctions" to refer to junctions not found in tissue-matched GTEx or TCGA normal samples, with a 10-read coverage requirement in TCGA, and allowing through the filter lowly expressed junctions in GTEx tissue-matched samples.[156]

Table 2.1: Junction novelty specification

Junction Novelty Stage	Definition
0	All junctions
1+	Junctions not found in tissue-matched GTEx or TCGA normal samples
2+	Junctions not found in any GTEx or TCGA normal ("core normal") samples
3+	Junctions not found in any core normal samples or in selected SRA tissue and cell type non-cancer samples

Extraction and analysis of cancer-specific junctions

We queried the junction database to extract junctions of interest, specifically 1) all junctions for all tumor samples of each cancer type and 2) all junctions not present in any core normal samples for each cancer type cohort, with their cohort prevalence levels. All junctions are presented in a 0-based closed coordinate system. We also identified a set of "shared junctions"

for every cancer type, defined as up to 200 most highly recurring junctions that occur in at least 1% of the cancer type samples and are not found in any core normal samples. Protein coding region presence was determined for all junctions, with location assessment as follows: the junction is categorized as protein-coding if it is present in a protein-coding gene region (with at least one junction splice site within the gene boundaries) and antisense if it is present on the reverse strand of a protein-coding gene region, based on gene regions described in GENCODE v.28.[49] Cancer-associated genes were collected from the OncoKB and the COSMIC cancer gene census; any gene listed in one or both lists was categorized as a cancer-associated gene. Any junction assigned to a protein-coding gene region corresponding to one of these genes was categorized as associated with cancer-relevant loci.

For comparison between cancer-sample junctions found vs. not found in core normal samples, we performed a Kruskal-Wallis H-test to determine the significance of the decreased sharedness levels, since the junction prevalence data is not normally distributed and there are many fewer cancer-specific junctions than junctions found in core normal samples.

Comparison with SRA tissue and cell types

Non-cancer sample types from the SRA were chosen via manual curation informed by a clustering of junctions according to ontology term prevalence, with commonly occurring terms that do not meaningfully distinguish junctions eliminated. The selected sample types in Supplementary Table S2.3 comprise all non-cancer data from the SRA analyzed. All junctions for samples associated with these cell and tissue types but not with "cancer" were downloaded via Snaptron, translated to a 0-based closed coordinate system, and compared with those found in TCGA cancer samples. Junctions present in a TCGA cancer-type cohort and SRA samples from a specific assigned category determined set assignments, which were used for subsequent data

analysis. To exclude false positive junctions but allow for the existence of splicing noise, only junctions with at least two reads across GTEx, TCGA, and the selected cancer and non-cancer SRA samples are considered true junctions. All SRA junctions not found in TCGA cancer samples were ignored. For the supplementary 2-sample minimum filter analysis, we retained all junctions that are present in only 1 SRA sample, but required at least two samples across the broad SRA category (adult, developmental, or stem cell) for inclusion in that set. (For developmental subsets, only one sample within a subset category was required, as long as the 2-sample criterion across the full developmental category was met.)

For comparison between TCGA cancer-sample junctions not found in core normal samples with SRA junctions from matched cancer type samples, we performed a Kruskal-Wallis H-test to determine the significance of the increased sharedness levels, since the junction prevalence data is not normally distributed and the difference in junction counts between the two cohorts (TCGA junctions in or not-in the SRA matched cohort) is large.

Comparison of junction burden and TMB

Silent and non-silent mutations per Mb per patient were added to give a total TMB per patient. Junctions considered for the "junction burden" calculation were all tumor sample junctions not found in core normal samples. The total junction count per patient was divided by the mapped read count of the sample divided by 10,000 (scaling to "per Mb" with the assumption of 100-base pair reads) to give the final junction burden. A linear regression was performed on the junction burden vs. TMB across all TCGA tumor samples.

Splicing Factor Mutation Analysis

Patient somatic mutation call files were downloaded from the GDAC firehose (http://gdac.broadinstitute.org/). While we note the potential importance of mutations in non-coding sequences, we confined our attention exclusively to non-synonymous mutations. Patients were classified based on two different separation criteria: 1) a *de novo* analysis of whether or not they had at least one mutation in a gene that codes for a protein annotated as involved in mRNA splicing, based on the UniProt protein annotation database, and 2) whether or not they had at least 1 mutation in a gene previously identified as sQTL associated (*U2AF1*, *SF3B1*, *TADA1*, *PPP2R1A*, and/or *IDH1*) in the TCGA cohort by the TCGA splicing paper.[156] For each cancer type, and each stratification method, the number of cancer-specific junctions per patient was compared for patients with and without at least one mutation in the defined set (Supplementary Figures S2.1F and S2.1G). Differences in the number of novel junctions across cancer types and stratification groups was assessed via two-way analysis of variance (ANOVA) with a Benjamini-Hochberg p-value correction.

In addition to comparing the levels of cancer specific junctions between patients with and without splicing associated mutations, we also compared junction sharedness based on the same two stratification criteria used above. For each cancer type, all junctions identified in two or more patients were selected. For each, the number of junction occurrences in patients with mutations in splicing associated genes was calculated and compared to the overall number of occurrences in the corresponding cancer cohort, using a Fisher's exact test (Supplementary Figures S2.1H and S2.1I).

Survival analysis for ovarian cancer patients with target antisense MSLN junction

All TCGA ovarian patients with data in columns "xml_days_to_last_followup" or "gdc_cases.diagnoses.days_to_death" in our TCGA phenotype file were included in the survival analysis. The survival curve was plotted for the second column, with dropout patients with no days-to-death data censored at days to last followup.

Data and Software Availability

All data is publicly available and accessible online as described in the Data Download section above. Python code and corresponding descriptors for the implementation of methods as described is publicly available on GitHub at https://github.com/JulianneDavid/ shared-cancer-splicing.

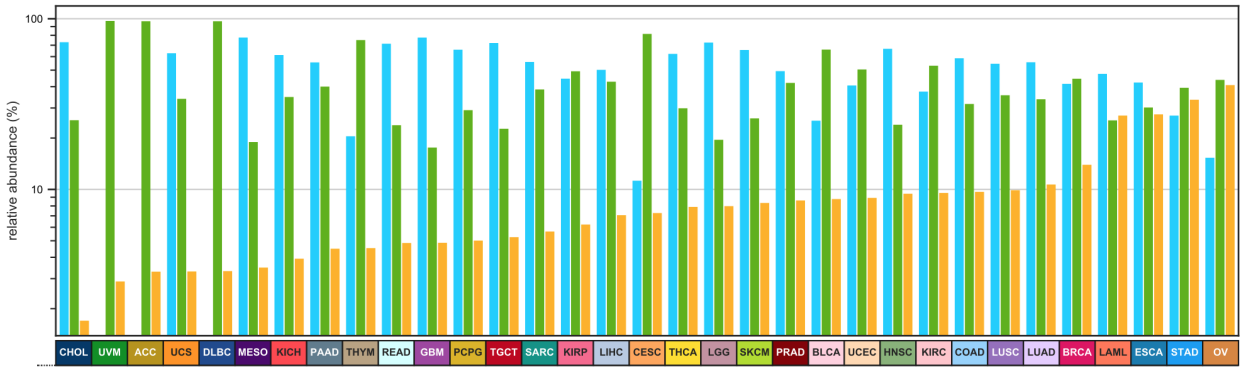
2.4 Results

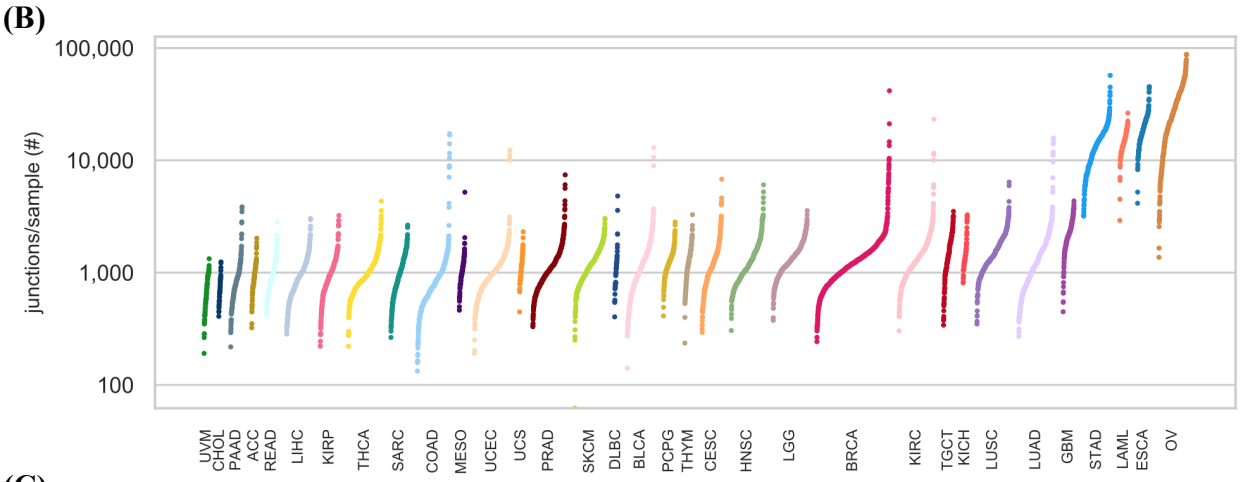
2.4.1 *Cancers harbor many novel shared exon-exon junctions not present in adult non-cancer tissues or cells*

While cancer-specific exon-exon junctions identified using tissue-matched normal samples have the potential to give rise to neoantigens,[156] we reasoned that they could be expressed in other normal tissues due to variability in patterns of transcription and alternative splicing among different tissues.[244] In such cases, these junctions might not yield bona fide neoantigens due to the prior tolerogenic education of the immune system. We therefore re-evaluated the incidence of cancer-specific junctions using RNA-seq data from TCGA and the large compendium of adult tissues from GTEx. We found that on average, across cancer types, 80.6% of junctions potentially thought to be cancer-specific based on comparison only with tissue-matched samples (σ = 13.0%) are in fact present in other adult non-cancer tissues and cell types throughout the body. Across cancer types, an average of 90.2% of all junctions found in

cancer samples (σ = 9.1%) are also present in one or more adult normal samples from GTEx or TCGA ["core normals"] (Figure 2.1A). The overall number of these novel junctions varies both within and across different cancer types, with ovarian carcinoma and uveal melanoma having the highest and lowest average number of junctions per sample, respectively (Figure 2.1B, Supplementary Table S2.1), and is independent of TMB (Supplementary Figure S2.1A). The set of junctions defined as "novel" is highly sensitive to the filtering criteria used (see Supplementary Figure S2.1B, Supplementary Table S2.2, and "Selection of cancer-specific junction filters" in Methods). We are interested in junctions that are widely expressed across samples, and for this analysis we sought to optimize sensitivity and specificity to detect shared cancer-specific junctions. High prevalence across a cancer-type cohort provides strong support for the existence of junctions, despite low coverage of these junctions within any individual sample (we require a minimum of 2 reads across all studies, but do not set a lower bound on sample coverage). Going forward, we use strict lack of occurrence of a junction in core normals as our baseline definition of cancer specificity, where even a single read in the target "normal" set eliminates a junction from the cancer-specific designation (see "Selection of cancer-specific junction filters" in Methods).

(A)





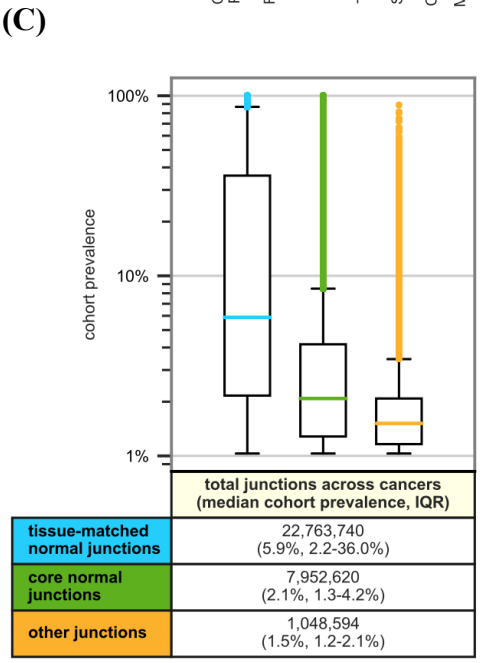


Figure 2.1: Distribution of exon-exon junctions across and within TCGA cancer cohorts.
(A) Log-scale bar charts describing the percentage of all junctions of a given cancer-type cohort present in three sub-cohorts. Blue (left) bars give the percentage of cohort junctions found in GTEx or TCGA tissue-matched normal samples (Supplementary Table S2.1); green (center) bars give the percentage of the remaining junctions that are found in other core normals; and yellow (right) bars give the percentage of cohort junctions found in no core normals; cancer types are ordered by relative abundance of junctions in this last set. Cancer types with no blue (left) bar have no tissue-matched normal samples (Supplementary Table S2.1). **(B)** Log-scale sorted strip plots representing the number of non-core normals junctions per sample for each of 33 TCGA cancer types. Each point represents a single TCGA tumor sample and the width of each strip is proportional to the size of the cancer type cohort.[156] Supplementary Figure S2.1B shows analogous data with additional filters applied. **(C)** Log-scale box plots representing the prevalences within each cancer-type cohort of junctions occurring in at least 1% of cancer-type samples, summarized across all TCGA cancer types. Junction prevalences are shown in blue (left) for those found in GTEx or TCGA tissue-matched normal samples (Supplementary Table S2.1); junctions not present in tissue-matched normals but found in other core normals are shown in green (center); and junctions found in no core normals are shown in yellow (right). Note that any junction found in multiple cancer types is represented by multiple data points, one for each cancer type in which it is found. A detailed breakdown by TCGA cancer type is available in Supplementary Figure S2.1E.

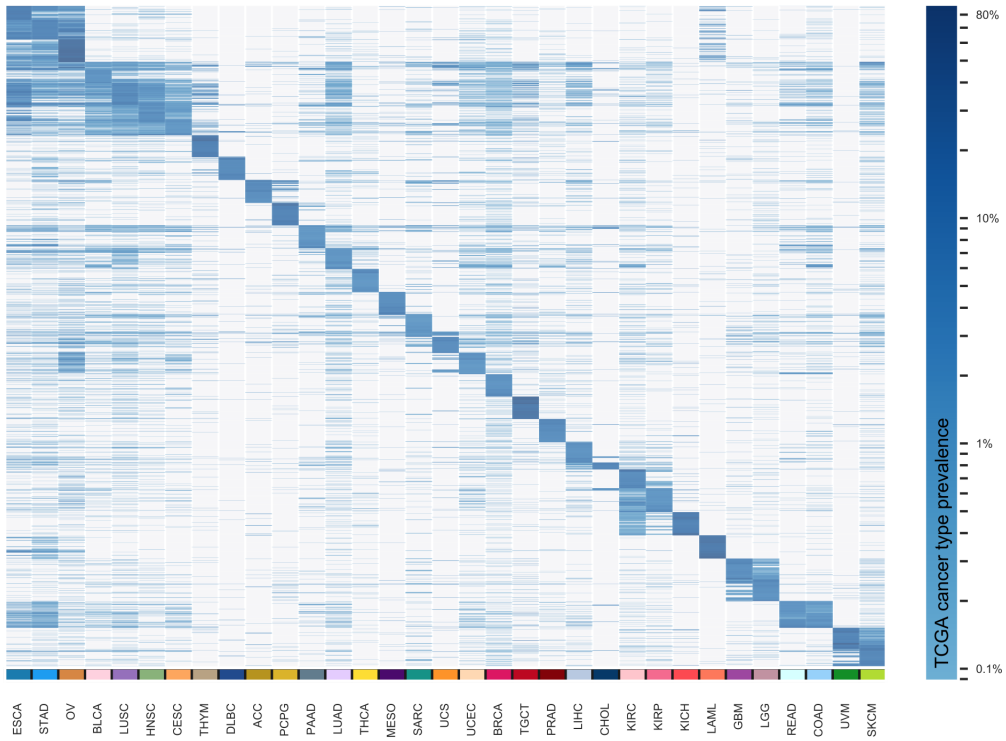
We next assessed the extent to which a given junction not found in core normals might be shared among multiple samples of the same cancer type. We observed that over half (52.8%) of these junctions are confined to individual samples, although a small but significant subset (0.41%) is shared across at least 5% of samples in at least one cancer-type cohort (Supplementary Figure S2.1C). We also noted that 40.6% of novel junctions are shared between multiple cancer types, with a total of 1,609 junctions present in at least 5% of samples each across two or more TCGA cancer cohorts (Supplementary Figure S2.1D). Sharedness was significantly higher among junctions that were also present in normal tissues (Figure 2.1C and Supplementary Figure S2.1E). We observed that the number of junctions not found in core normals per patient was comparable for patients with and without splicing factor-associated mutations across all cancer types, with the exception of breast adenocarcinoma (Supplementary Figures S2.1F and S2.1G). We also observed that splicing-associated mutations had minimal effect on the sharedness within a cancer-type cohort of junctions not found in core normals (Supplementary Figures S2.1H and S2.1I).

We finally assessed whether these junctions were also shared among independent cancer cohorts, using publicly available RNA-seq data in the SRA.[245] Many TCGA cancer junctions not found in core normals were found to occur in cancer-type matched SRA samples: 11 of 14 cancer types had more than 50 junctions in common between the matched cohorts. Moreover, we found that junctions also present in matched SRA cancer cohorts were associated with significantly higher levels of sharedness in the TCGA cohort (H statistic = 3.85-2,803 and p = <0.0001-0.0495; Supplementary Figure S2.1J).

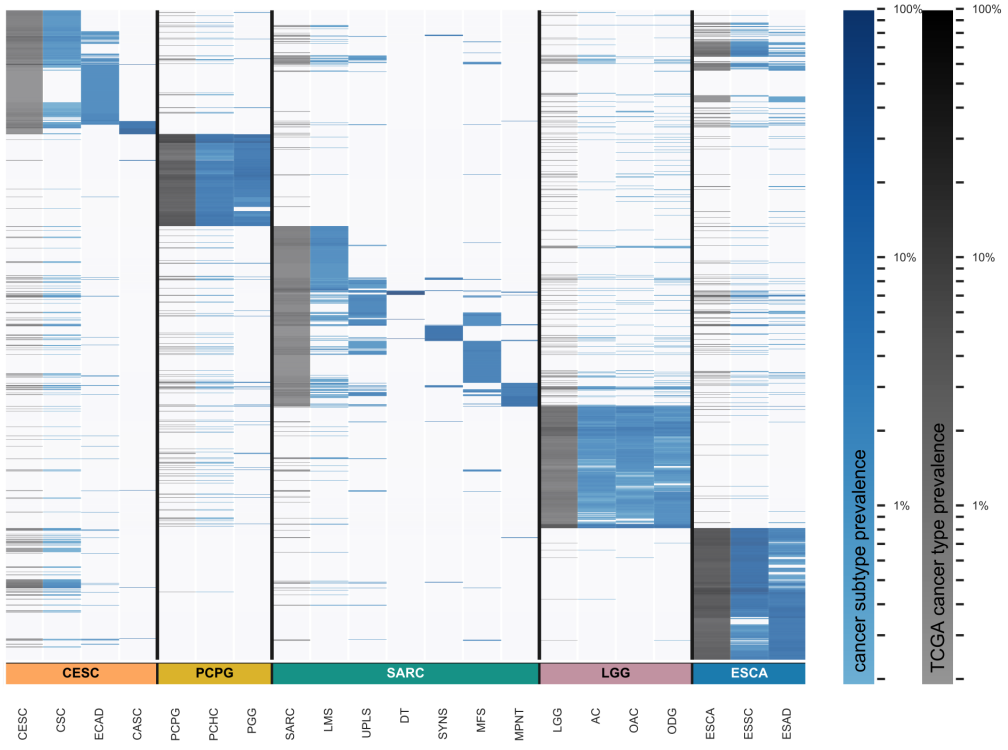
2.4.2 *Shared novel junctions in cancer distinguish cancer identity and subtype*

We hypothesized that a high level of exon-exon junction sharedness across samples is likely to be reflective of underlying conserved biological processes (e.g. among normal tissues). We therefore investigated the sharedness of novel junctions present in different cancer types. Interestingly, these novel junctions can readily distinguish disparate cancer types and show similarities among cancer types with shared biology, such as cutaneous and uveal melanomas (Figure 2.2A). These novel junctions also reflect shared biology among additional cancer types with similar anatomic origins: colon and rectal adenocarcinoma, clear cell, chromophobe, and papillary renal cell carcinomas, low and high grade gliomas, and stomach and esophageal adenocarcinomas (Figure 2.2A). Shared junctions from several cancer types also demonstrate similarities by histological subtype despite their differing anatomical origins, for instance squamous cell carcinomas of the lung, cervix, and head and neck (Figure 2.2A, Supplementary Figure S2.2A), consistent with previously published work.[246] Moreover, shared novel junctions are readily able to distinguish distinct histological subtypes of sarcoma and cervical cancer, among other diseases (Figure 2.2B). Using non-cancer cell types from the SRA we found that "novel" junctions from cancers arising from cell and tissue types poorly represented in GTEx normal tissue samples (e.g. melanocytes), or not present in GTEx at all (e.g. thymus tissue), can be found in many samples of the corresponding cell or tissue types of origin (Figure 2.2C, Supplementary Table S1). Sample-to-sample comparisons of all junctions from these rare-cell type cancers also show more similarity with cell type-matched normal samples from the SRA than with bulk tissue from GTEx (Supplementary Figure S2.2B).

(A)


80%
10%
1%
0.1%
TCGA cancer type prevalence
ESCA
STAD
OV
BLCA
LUSC
HNSC
CESC
THYM
DLBC
ACC
PCPG
PAAD
LUAD
THCA
MESO
SARC
UCS
UCEC
BRCA
TGCT
PRAD
LIHC
CHOL
KIRC
KIRP
KICH
LAML
GBM
LGG
READ
COAD
UVM
SKCM

(B)


100%
10%
1%
cancer subtype prevalence
100%
10%
1%
TCGA cancer type prevalence
CESC
PCPG
SARC
LGG
ESCA
CESC
CSC
ECAD
CASC
PCPG
PCHC
PGG
SARC
LMS
UPLS
DT
SYNS
MFS
MPNT
LGG
AC
OAC
ODG
ESCA
ESSC
ESAD

(C)

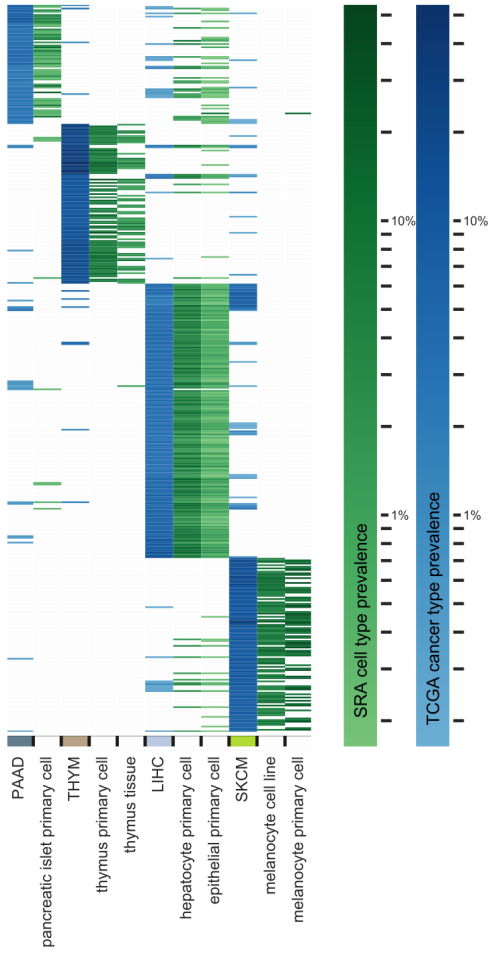


Figure 2.2: Clustering by cohort prevalence of shared novel junctions not found in core normal samples. **(A)** Heatmap showing junction prevalences across every TCGA cohort for each cancer type's top 200 shared junctions that are at least 1% prevalent in that cancer type and are not found in any core normal samples. **(B)** Heatmap showing shared junction prevalences across selected TCGA cancer types and their assigned histological subtypes for each subtype's top 200 shared junctions that are at least 1% prevalent in that subtype and are not found in any core normal samples. See Supplementary Table S2.1 for TCGA subtype abbreviations. **(C)** Heatmap showing shared junction prevalences across selected TCGA cancer types and a set of their matched SRA tissue and cell types of origin, for each cancer type's top 200 shared junctions that are at least 1% prevalent in that cancer cohort and are not found in any core normal samples. See Supplementary Table S2.3 for SRA sample type abbreviations.

2.4.3 *Novel junctions in cancer are found among developmental and known cancer-related pathways*

As many cancers are thought to recapitulate normal developmental pathways,[247–249] we further hypothesized that a subset of cancer-specific junctions may reflect embryological and developmental splicing patterns. We therefore compared cancer junctions not found in core normals with those from SRA samples pertaining to zygotic, placental, embryological, and fetal developmental processes: on average, per cancer type, 15.4% of these cancer junctions (σ = 2.4%) occur in SRA developmental cell or tissue samples. We also considered samples from

SRA normal stem cell samples and from selected SRA normal adult tissues and cell types: on average, per cancer type, 2.7% (σ = 1.4%) and 26.5% (σ = 3.3%) of cancer junctions not found in core normals occur in stem cell and selected adult tissues, respectively (Figure 2.3A and Supplementary Figure S2.3A). Furthermore, many of the junctions found in SRA developmental, stem cell, and selected adult tissues are highly prevalent shared junctions (Supplementary Figure S2.2A). The remaining significant majority of these cancer junctions not found in core normals were also not found in any non-cancer SRA tissue or cell type studied (64.9% on average per cancer type cohort (σ = 4.0%), Figure 2.3A and Supplementary Figure S2.3A). Many of these novel "unexplained" junctions still exhibit high levels of sharedness both within (Supplementary Figures S2.3B and S2.3C) and between (Supplementary Figure S2.3D) different cancer types. At the upper end, 16 of these shared junctions were found in more than 10% of samples in each of two or more cancer types (Supplementary Table S2.4).

We note that the liberal set inclusion criterion we employed may reduce our ability to identify robust cancer-specific biology among unexplained junctions. For instance, the well-described deletion causing a splicing of exons 1 and 8 (*EGFRvIII*) occurs in 29.4% of TCGA patients with glioblastoma multiforme (GBM) and in no core normals, but is also present in a single read from a single human epithelial cell line sample on SRA, and therefore is classified not as an unexplained cancer-specific junction but as "adult non-cancer." However, this set inclusion condition does allow for the identification of some cancer-specific biology of interest. For instance, rarer alternative *EGFR* splicing events were detected in the unexplained set, such as *EGFRvIII* with an alternate exon 1 joined to exon 8 (chr7:55161631-55172981), detected in 2 patients with GBM and 1 patient with low grade glioma; the same alternate exon 1 joined with two alternate exon 16s (chr7:55161631-55168521 and chr7:55161631-55170305) (detected in 1

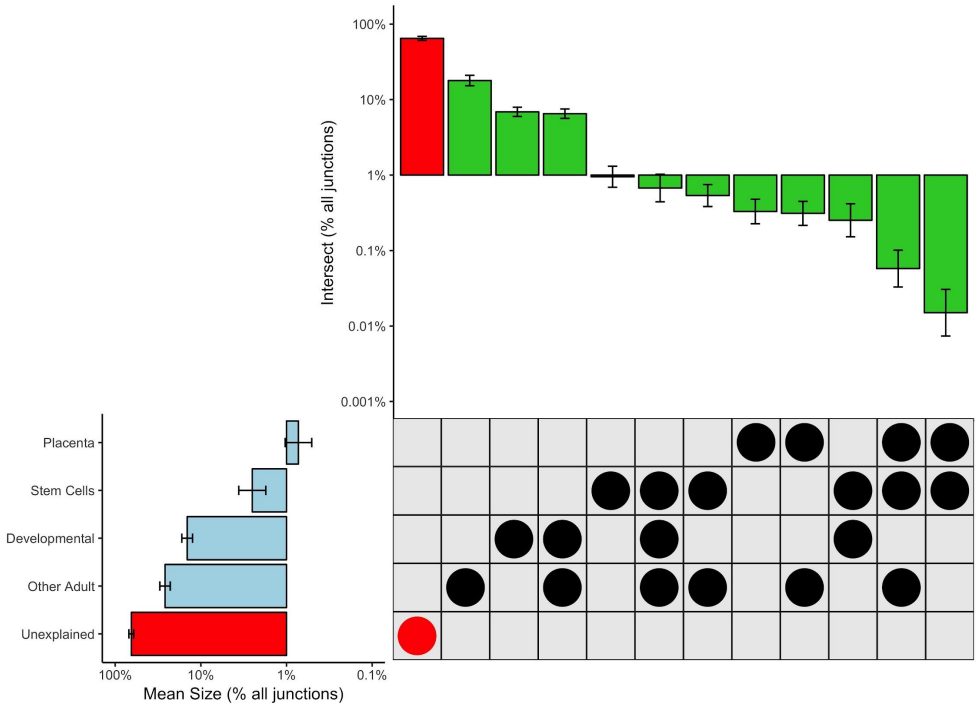
and 2 GBM patients, respectively); and the same alternate exon 1 joined with exon 20 (chr7:55161631-55191717) in 2 GBM patients. An alternative filtering approach that instead requires two samples per SRA category to define junction set membership yields a greater number of unexplained junctions (Supplementary Table S2.2 and Supplementary Figures S2.3E and S2.3F).

We observed a number of unexplained junctions shared by unusually large proportions of ovarian cancer (OV) samples in TCGA, including one cancer-specific junction (chr16:766903-768491 on the minus strand) present in the highest proportion of samples in any TCGA cohort (81.3%, or 350 of 430 samples in OV). This junction occurs in an antisense transcript of *MSLN*, which codes for a protein known to bind to the well-known ovarian cancer biomarker *MUC16* (CA125).[250,251] The functional consequences of this junction are unknown, but it does not appear to affect overall survival (Supplementary Figure S2.3G). Another unexplained junction (chr19:8865972-8876532 on the minus strand) is in the *MUC16* region itself and is present in 42.8%, or 184 of 430 samples in OV. In all, we identified 34 cancer-specific junctions present in >40% of OV samples. We further identified several novel pan-cancer splice variants (chr16:11851406 with chr16:11820297, chr16:11821755, and chr16:11828391, each present across up to 8 different cancers) in *RSL1D1* and its neighboring *BCAR4*, a long noncoding RNA known to promote breast cancer progression.[252,253]

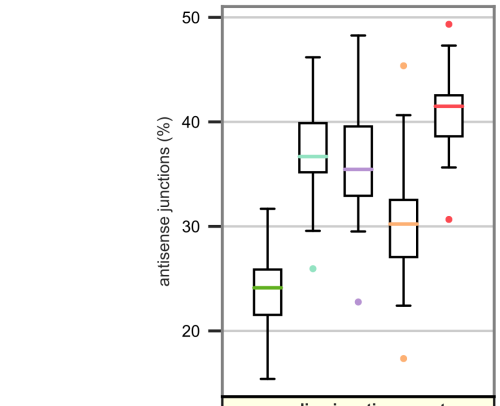
Among all otherwise unexplained junctions, an average of 4.78% (σ = 0.48%) across cancer types are associated with known cancer-predisposing or cancer-relevant loci. Further, an elevated proportion of otherwise unexplained junctions (on average, 40.9%, σ = 3.8%) occur in likely antisense transcripts and may therefore be of reduced interest as candidate neoantigens, but sustained interest in terms of cancer biology (Figure 3B, Supplementary Table S2.5). Finally, we

show that 20 genes not previously known to be cancer-associated each contain at least 25 novel, unexplained junctions present in at least 5% of samples of at least one cancer type (Supplementary Table S2.6).

(A)



(B)



	median junction count across cancer types
Core normals	2,148,683 (IQR: 1,381,622-2,571,867)
Other adult non-cancer	45,625 (IQR: 19,938-72,107)
Developmental	2,257 (IQR: 876-4,210)
Stem cell	12,435 (IQR: 5,087-20,596)
Unexplained	111,091 (IQR: 42,408-177,027)

Figure 2.3: Junction set assignments and antisense junction prevalence in additional normal tissue and cell type categories from the Sequence Read Archive, across cancers.
(A) Upset-style plot with bar plots showing junction abundances across major sets (left) and set overlaps (top) across 33 cancers (error bars). Shown junctions are absent from all core normals. Unexplained junctions (red highlights) comprise junctions not present in any set categories studied (see also expanded set assignments in Supplementary Figure S2.3A). The developmental set comprises human development-related junctions not present in the category placenta. Scale is log10 of percent of junctions not found in core normals, calculated for each cancer.
(B) Box plots showing, for each TCGA cancer type, the percent of junctions that are antisense for (green) junctions found in core normals; (aqua) junctions not found in core

normals but found in other selected non-cancer adult tissue and cell samples from the SRA; (lavender) junctions not found in core normals or SRA non-cancer adult samples but found in selected developmental samples on the SRA; (apricot) junctions not found in core normals or SRA non-cancer adult samples but found in selected stem cell samples on the SRA; and (red) junctions not found in core normals or selected non-cancer adult, developmental, or stem cell samples from the SRA. Each point represents the percent of junctions from one cancer type in the given category (e.g. developmental) that are antisense. The table shows the median and IQR of the number of junctions in that category across all TCGA cancer types.

2.5 Discussion

Previous studies have established the importance of alternative and aberrant splicing in cancer prognosis[134,229–233] and have begun to explore its potential relevance in cancer immunotherapy.[156,186,254] In this study, we explore "novel" exon-exon junction use among cancers with respect to a broad collection of normal tissues and cells. This is the largest such study to-date, integrating RNA-seq data from 10,549 tumor samples across 33 TCGA cancer types, 788 paired normal samples across 25 TCGA cancer types, 9,555 normal samples across 30 GTEx tissue types, and 12,231 human samples from the SRA (10,827 samples from 33 normal tissue and cell types and 1,404 samples from 14 cancer types) (Supplementary Tables S2.1 and S2.3). To the best of our knowledge, this is also the first study to examine the novelty of cancer junctions from the perspective of immune tolerance, considering all adult normal tissue types as potential sources of tolerogenic peptides rather than only the closest matched normal tissues. Moreover, this is the first study to quantitatively interrogate the sharedness of novel exon-exon junctions both within and across cancer types, demonstrating that these junctions can distinguish some cancers and their subtypes. We finally demonstrate that there is no one-size-fits-all definition of "novel" splicing, noting that purportedly cancer-specific junctions may in fact be present among, and perhaps biologically consistent with, a repertoire of embryological, developmentally-associated, and other cell types.

This study also has several limitations. We focus on the importance of exon-exon junctions as the predominant metric of alternative splicing, in particular on their presence or absence among different samples, but do not explore the potential for differences in gene dosage to drive differences in biology. Moreover, there are other sources of RNA variation (e.g. intron retention events[154] and RNA editing) that we do not explicitly study here, but which could be equally good sources of novel, cancer-specific protein sequence for immunotherapeutic and other applications. Importantly, there is substantial variability among analytical methods for identifying these exon-exon junctions. We note significant discordance between results of analyses of the same data using different junction filtering methods. While the same phenomena and general results appear to hold true independent of analytical technique, the identity and relative novelty of individual "cancer-specific" junctions vary between our results and those previously published.[156] We also acknowledge that GTEx and the SRA combined do not account for all sources of normal tissue(s) in the human body, and further acknowledge that the sample metadata used to search the SRA may be an imperfect surrogate for actual tissue/sample identities. Our assessment of embryological and developmentally-associated junctions is also limited by a relatively small number of relevant RNA-seq samples available on the SRA. Our splicing factor mutation analysis was also limited by sample size and was confined exclusively to non-synonymous mutations. Finally, due to the short-read nature of these RNA-seq data, we make no attempt to predict putative neoepitopes from cancer-specific junctions as we cannot confidently recapitulate reading frame or broader sequence context from isolated exon-exon junctions, particularly without access to the biological specimens to perform junction-level experimental validation.

While cancer-specific exon-exon junctions may indeed be a source of neoepitopes, their sharedness across individuals and occurrence in cancer-relevant loci (e.g. *EGFR*, *MUC16*) are suggestive of underlying but as-of-yet unexplored biology. This sharedness does not appear to be related to variants in splicing factor or splicing-associated proteins, and is not wholly explained by recapitulation of embryological/developmental transcriptional profiles. As such, we see this work as opening a broad area of future research into the role and relevance of these novel recurring exon-exon junctions.

Acknowledgements

We thank Paul Spellman for helpful discussions and Chris Wilks for facilitating Snaptron queries. We thank Mary Wood for her critical reading of the manuscript. The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Author Contributions

JKD, RFT, and AN formulated research goals and designed the experiments. JKD, BRW, SKM, and AN designed, implemented, and tested computer code. JKD and AN verified reproducibility of results. JKD, BRW, SKM, and AN performed the experiments including statistical and computational data analysis. JKD and BRW were responsible for data curation and maintenance. JKD, BRW, and SKM prepared figures for data visualization. RFT and AN provided computing resources, mentorship, and project oversight. JKD, BRW, RFT, and AN wrote the initial manuscript draft and all authors read, edited, and approved the final manuscript.

Chapter 3: Exploring the validity of cancer-specific junctions

This work has been formatted for inclusion in this dissertation from the manuscript "Methods for detection and validation of peptides from cancer-specific splicing" by Julianne K. David*, Laurie Prélot*, Andy Lin, André Kahles, Gunnar Rätsch, Reid F. Thompson, and Abhinav Nellore, in preparation (2022). The author of this dissertation is the co-primary author of the manuscript. (*co-first authors)

3.1 Abstract

Peptides arising from cancer-specific RNA splicing are of significant recent interest as neoepitopes for use in immunotherapy treatment, and identification of these potential targets implicitly relies on their biological reality and their cancer-specificity. Here, we explore methods for identifying cancer-specific exon-exon junctions and associated peptides on a set of 5 breast cancer and 5 ovarian cancer RNA-seq samples from The Cancer Genome Atlas (TCGA), against backgrounds of normal samples drawn from TCGA and the Genotype Tissue Expression Project. We find that using a method based on phasing short RNA reads into a splicing graph generates more potential cancer-specific junction peptides than filtering on junctions alone, and that splicing peptides are much more varied and novel across ovarian cancer samples compared with breast cancer. We query sample-matched mass spectrometry (MS) data from the Clinical Proteomic Tumor Analysis Consortium and find that the current MS detection method is highly variable between experiments and that under stringent filtering criteria, breast cancer samples produce too few junction peptides to detect via MS.

3.2 Background

Almost all genes undergo alternative splicing, a process of joining exons together in different combinations by which one gene can form various transcripts and, for protein-coding genes, translated proteins.[67,68] This process is frequently disrupted in cancer,[128] leading to noisy

alternative splicing and to aberrant splicing,[130] where novel exon combinations not seen in normal tissue are used and the production of functional proteins may be disrupted.[183] While many aberrantly spliced transcripts may undergo nonsense mediated decay due to the presence of premature stop codons,[78] aberrant, cancer-specific splicing may in some cases lead to a stable protein product.[255] This opens the possibility of using these potential junction-specific peptides as biomarkers[176,256,257] or as immunotherapy targets.[155,156,186,258] However, these uses rely on 1) confidence in the reality of the identified splicing, rather than it occurring as a technical artifact of sequencing or alignment, and 2) either true cancer-specificity, or aberrantly high expression in cancer, of the protein product, neither of which is straightforward to assess.

In-sample RNA expression is often used as a proxy for the former, but no single value clearly delineates an appropriate coverage level to guarantee "real" splicing. The experiment sequencing depth will affect confidence, and since a true but rare isoform would have low expression in RNA-seq, setting a conservative threshold may preclude identification of valid cancer-specific isoforms. Furthermore, these may have high protein expression despite low RNA abundance.[64,224]

The second problem, assessing cancer-specificity, is more complex and has been approached in different ways. The primary question, what is defined as "normal" splicing, has been addressed with varying set sizes and sample types comprising a normal cohort, and with varying levels of expression allowed within this cohort. The normal cohort can range from, leniently, a small number of tissue-matched samples[154] to a stringent set of thousands of samples from many tissue types.[155,156,186,258] However, setting even strict boundaries on lack of presence in large normal sample sets can allow "cancer specific" junctions to be identified that still occur in normal tissues.[1]

Other differences in approach include filtering at the RNA[1,155,186] or peptide[156] level (or both[258]), inclusion[156] or not[1,155,186] of peptides overlapping a normal junction that are novel due to non-splicing effects such as an upstream frame shift, and whether to attempt phasing of short RNA reads[156,219] to generate full transcript and reading frame context for each junction.

With no ground truth against which to assess the identification accuracy of cancer specific junctions or junction-associated peptides directly, a primary method of validation has been proteogenomics.[259] Typically, this involves generating a protein sequence database by translating sequences obtained from RNA-Seq data which is then used in a proteomics database search of sample-[156,258] or cell line-[155,186] matched mass spectrometry (MS) proteome data to detect peptides found in a sample. Traditionally, tandem MS spectra are searched against a database of all proteins that are reasonably expected to be in the sample. However, in cancer samples the number of possible proteins greatly increases due to processes such as aberrant splicing, leading to a loss of power. In addition, additional power is lost as a result of hypotheses that result from validating non-cancer specific junctions. Recently, a new method, subset-neighbor search (SNS) was developed allowing an MS data set to be queried specifically for a targeted subset of peptides directly relevant to a specific scientific question without sacrificing power and with proper false discovery rate estimation.[260] In this work, we perform a comprehensive examination of method and filter value choices in identification of cancer-specific splicing peptides, and use SNS on sample-paired proteomic data to determine the accuracy of identified peptide sets.

3.3 Results

3.3.1 *Overview of filtering experiments and two pipelines*

Here, we study the detection of potential “alternative splicing neoepitopes” (pASNs),

defined as 9-mers overlapping an exon-exon junction that have passed filtering against a normal proteome. We probe detection by two distinct methods as implemented in junction- and graph-based pipelines (JP and GP respectively). The JP identifies peptides translated directly from cancer-specific splice sites, and therefore performs most filtering on exon-exon junctions. The JP is explicitly annotation based, translating only junctions whose upstream splice site falls within the coding DNA sequence region (CDS) of an annotated protein-coding transcript, in the annotated reading frame. In contrast, the GP focuses on cancer-specific peptides arising from splice sites, which may not be novel or cancer-specific. A comprehensive graph[219] representing splicing across a set of cohort samples is built, and targets are generated by applying annotated reading frames across both annotated transcripts and across novel transcript paths.

We analyze the results of filter choices made during alternative splicing neoepitope (ASN) detection (Figure 3.1) by both methods, by exploring the parameter space of five independent filters that represent confidence in the biological reality of a given junction or peptide and its level of cancer-specificity. These filters are 1) thresholds for expression for the junction within the target sample's RNA-seq data; 2) requiring, or not, the junction's splice motif to be canonical (implemented by the JP only); 3) thresholds for expression across a sample-type cohort; 4) thresholds for expression across a normal tissue cohort; and 5) lack of pASN presence in the normal human proteome. Stringency in the first three filters increases confidence in the accuracy of the junction or peptide call, while stringency in the last two increases confidence in the peptide's cancer specificity.

We perform these combinatorial filter experiments on 10 tumor samples, five randomly selected from each of the TCGA breast cancer (BRCA) and ovarian cancer (OV) cohorts (Methods) for the JP, and on the five selected BRCA samples for the GP.

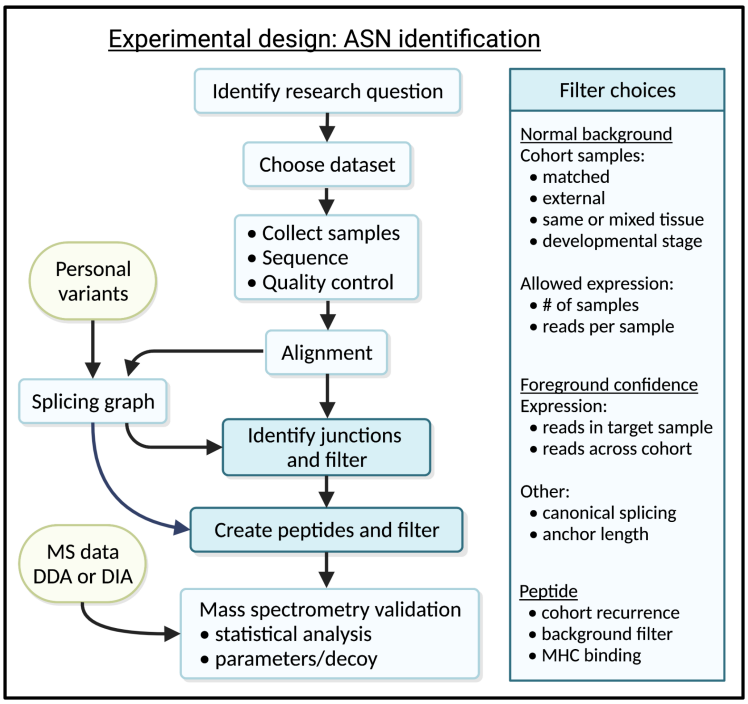


Figure 3.1: Decision points in an ASN detection experiment. Choices stem from the target research question, including the dataset in which the ASNs should be identified; the method of data collection, including sample sequencing, quality control, and alignment method and parameters; the method of identifying and filtering cancer-specific junctions; and the method of validation. We focus on the highlighted boxes in blue, for which detailed options are listed to the right. Created with BioRender.com.

3.3.2 *Junctions and junction peptides found in ovarian and breast cancer samples*

The unfiltered set of all junctions called from alignment was smaller for BRCA samples than OV, with averages of 265 and 444 thousand junctions per sample, respectively (Supplementary Table 3.1). More BRCA (63% on average) than OV (39%) junctions are fully annotated in GENCODE v.32[49] (Figure 3.2). BRCA junctions follow an expected distribution[261] across canonical splice motifs, while an average of 16% of called junctions in OV have non-canonical motifs (vs. 1.6% on average in BRCA) (Figure 3.2, Supplementary Table 3.1). Of the total sample junctions, 54% and 47% on average for BRCA and OV respectively are translated by the JP (Supplementary Table 3.2). OV contains a more 9-mers (40%) than BRCA (20%) that are not found in the normal human proteome (Supplementary Table 3.2).

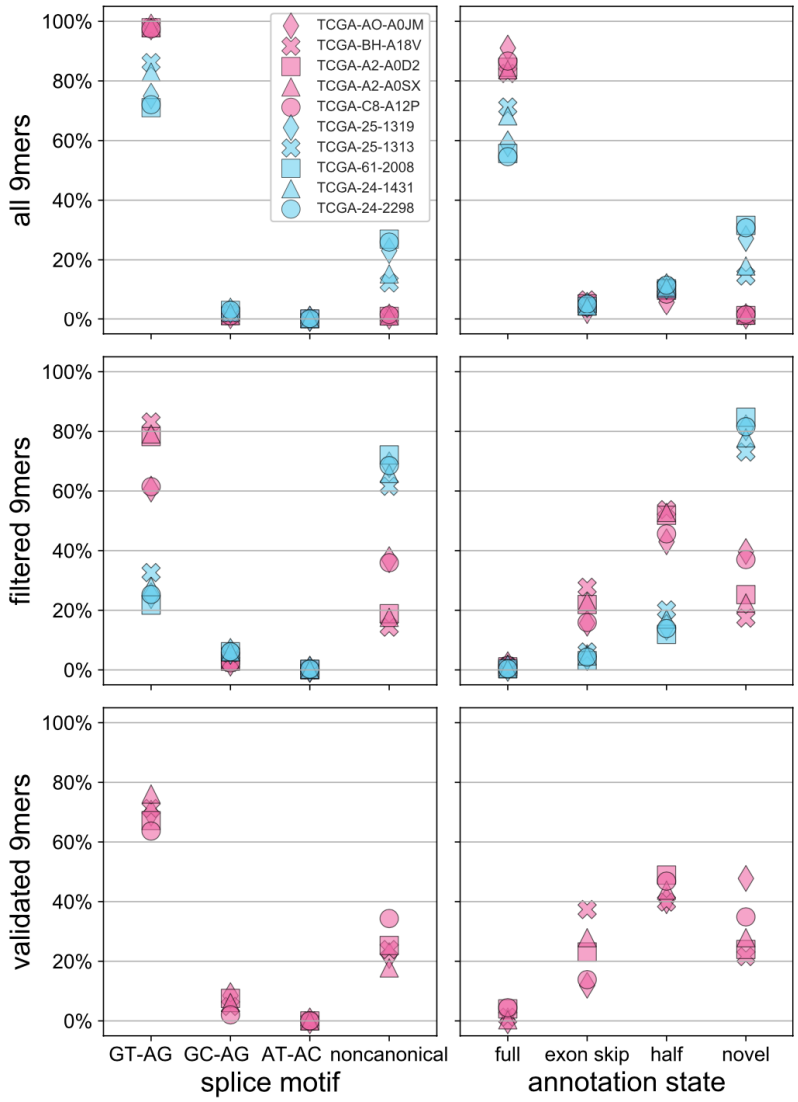


Figure 3.2: Distributions of 9-mer splice motifs and annotation states. For all 9-mers generated by the JP (top panels), those passing cancer-specific filters (middle panels), and those validated in the CPTAC MS data (bottom panels), the left and right columns show breakdown by splice motifs and annotation states. Motif options are the three canonical splice motifs, GT-AG, GC-AG, and AT-AC, and all other (noncanonical) motifs. Annotation options are fully annotated in GENCODE; exon skip, or both ends annotated but not together; half, or one end only is annotated; and novel, where neither end is annotated.

Across all BRCA samples, the JP translates 70% fewer junctions than the GP, and only 91% and 27% respectively of the JP and GP total translated junctions are mutually translated by both pipelines (Supplementary Table 3.3). From the set of mutually translated junctions (MJ), the GP results in 75% more unique junction peptides than the JP (Supplementary Table 3.3). The

translated peptides arising from the two pipelines have high concordance: of all peptide-junction pairs arising from the set of mutually translated junctions, 98% of the JP peptides have a match within the GP set, and 78% of the GP set have a match within the JP set. Altogether, the GP yields more peptides per junction than the JP, with averages of 1.9 peptides per junction overall and 2.7 peptides per MJ, vs. 1.4 peptides per junction in both categories for the JP.

3.3.3 *Summary of filter experiment results*

Filter experiments were performed with the JP and the GP, and junction-spanning 9-mers (pASNs) were collected and compared across experiments. Across the 5 BRCA samples, an average of 4,237 translated junctions (3.0%) in protein-coding regions passed at least one set of filters, with the large majority of these (an average of 92.5% across samples) not fully annotated in GENCODE v.32 (Figure 3.2, Supplementary Table 3.4). The filtered junctions yielded an average of 34,295 junction-overlapping 9-mers, of which 85% were not found in the normal human proteome (Supplementary Table 3.4). The largest differences in output counts between filter experiments were due to the choice of normal cohort (Figures 3.3 and 3.4, Supplementary Figures 3.1, 3.2, and 3.3), where the most lenient normal cohort yielded on average 495–25,421 pASNs, while the most stringent yielded 12–6,212 pASNs (Figure 3.4, Supplementary Figure 3.2). Most pASNs arose from junctions with canonical splice motifs, although an unusually high proportion (average of 21% across samples) had non-canonical splicing (Figure 3.2).

Filter stringency had a much lower effect on the number of pASNs output by the GP (Supplementary Figure 3.1), which ranged from an average across samples of 20,569–22,900 from the most lenient to the most stringent filtering experiments. Altogether, the GP gave on average 5.6 times more peptides than the JP for the core GTEx filter set (an average of 21,584 vs. 3,873 across samples). Very few pASNs were returned by both pipelines (Supplementary

Figure 3.4), and very few of either pipeline's pASNs are even generated by the other (Table 3.1).

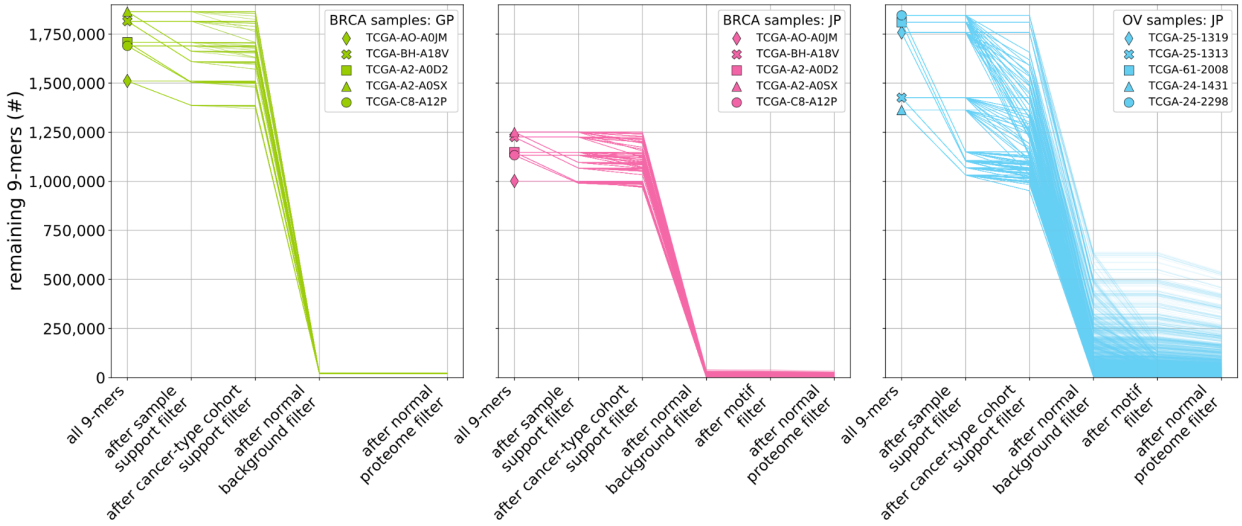


Figure 3.3: Effect of filters on remaining 9-mer counts. Each panel depicts, for one cancer type (left and middle, BRCA; right, OV) and pipeline (left, GP; middle and right, JP), the number of initial junction-spanning 9-mers generated for each sample (marked on the left by a unique shape, see legend) and the number of 9-mers remaining (y-axis) for each filter experiment after each filter stage (x-axis).

The OV samples yielded an order of magnitude (>13x) more translated junctions and pASNs across filter experiments (Supplementary Table 3.4), as well as a larger proportion (27% of all OV junctions, vs. 3% for BRCA). An average of 3,452 to 396,401 pASNs passed the most stringent to the most lenient JP filter experiments across the 5 OV samples. Filtered junctions contained a high proportion of fully unannotated junctions (79%) and those with non-canonical splice motifs (67%) (Figure 3.2, Supplementary Table 3.4). While the normal filter still had the largest effect on the output pASN counts as seen for BRCA, the OV results were also strongly affected by the canonical motif filter (as expected due to the high proportion of non-canonical motifs, Supplementary Table 3.4), the cancer-type cohort filter (perhaps due to the smaller number of OV samples than BRCA samples across TCGA), and the sample support filter (due to 4 of the 5 target OV samples having a single read scaled to a normalized expression value <1, see Methods) (Figure 3.3, Supplementary Figure 3.1).

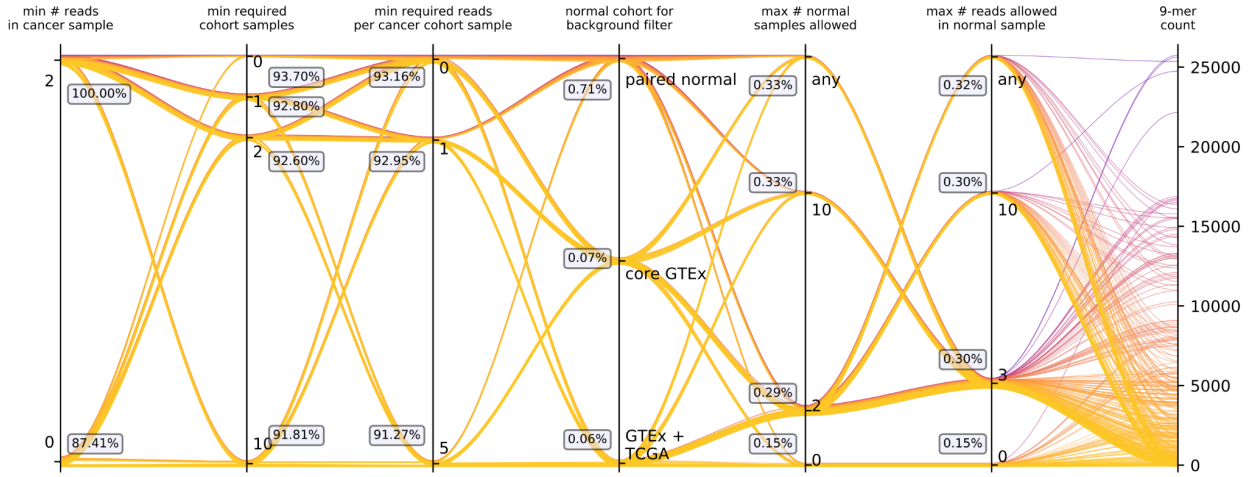


Figure 3.4: Effect of JP filters on final 9-mer count for BRCA sample TCGA-C8-A12P. Each vertical axis except the rightmost represents one filter, showing parameter options from most stringent (bottom) to most lenient (top). Each colored line represents one JP filter experiment, with its path passing through the filters parameters it uses and its color mapped to the final number of 9-mers passing the full set of filters (yellow == low, purple == high). The rightmost axis shows final filtered 9-mer counts for each filter experiment, with each filter experiment colored line terminating at its final value. Floating gray boxes show, across experiments passing through the corresponding filter parameter, the mean of the ratio of remaining 9-mers after that filter parameter has been applied to the sample's total initial generated 9-mers.

Table 3.1: Proportion of filtered 9-mers generated by both pipelines.

sample	GP-only filtered 9-mers (#, % generated by JP)	JP-only filtered 9-mers (#, % generated by GP)
TCGA-A2-A0D2	22,746 (993, 4.4%)	6,033 (98, 1.6%)
TCGA-A2-A0SX	23,350 (1,031, 4.4%)	5,702 (112, 2.0%)
TCGA-AO-A0JM	21,398 (971, 4.5%)	4,062 (123, 3.0%)
TCGA-BH-A18V	23,575 (1,145, 4.9%)	6,836 (125, 1.8%)
TCGA-C8-A12P	22,616 (993, 4.4%)	9,063 (109, 1.2%)

3.3.4 *Mass spectrometry queries of BRCA peptides across experiments*

We queried junction-overlapping trypsin-digested peptides arising from the filtering experiments against sample-matched MS data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Most JP filter experiments (an average of 575/720 across samples), comprising the more stringent filter sets, had no validated pASNs due to fewer peptides

predicted than are required for discovery at a 5% false discovery rate (FDR). The remaining experiments yielded an average of 131 validated pASNs per experiment per sample (1.84% on average of predicted pASNs, the "validation ratio"), arising from an average of 18.4 junction peptides. These had a similar distribution of splice motifs and annotation states to predicted pASNs, with a large proportion (24.8% on average across samples) having non-canonical splice motifs (Figure 3.2, Supplementary Table 3.5). The distribution of both validation counts and ratios across filter parameters (Supplementary Figures 3.5 and 3.6) reflected the distribution of filtered 9-mers (Supplementary Figure 3.3). More validated 9-mers arose from lenient filter sets, although validation ratios tended to be higher for more stringent filters (Figure 3.5, Supplementary Figure 3.7).

While the GP generated over an order of magnitude more pASNs than the JP, the two pipelines had similar numbers of unique junction-overlapping trypsin-digested peptides (Supplementary Figure 3.8). For the subset of matched filter experiments (the core GTEx normal cohort only and no motif filter applied), the GP had relatively high validated peptide counts (an average of 754 per experiment per sample) and low validation ratios (an average of 0.44% across samples) (Supplementary Figure 3.9), although its ratio of validated peptides to unique tryptic peptides tested was much higher (average 17% across samples). The range of validated counts within samples was low (average across samples of 17.8, or 2%) as also seen for filtered peptides (Figure 3.3, Supplementary Figure 3.1). Notably, the GP had higher validation ratios for experiments with more stringent sample expression support required (Supplementary Figures 3.9 and 3.10). As the peptides from these filter experiments are a subset of those with more lenient required sample support, we note that this shows some instability in the proteomics validation.

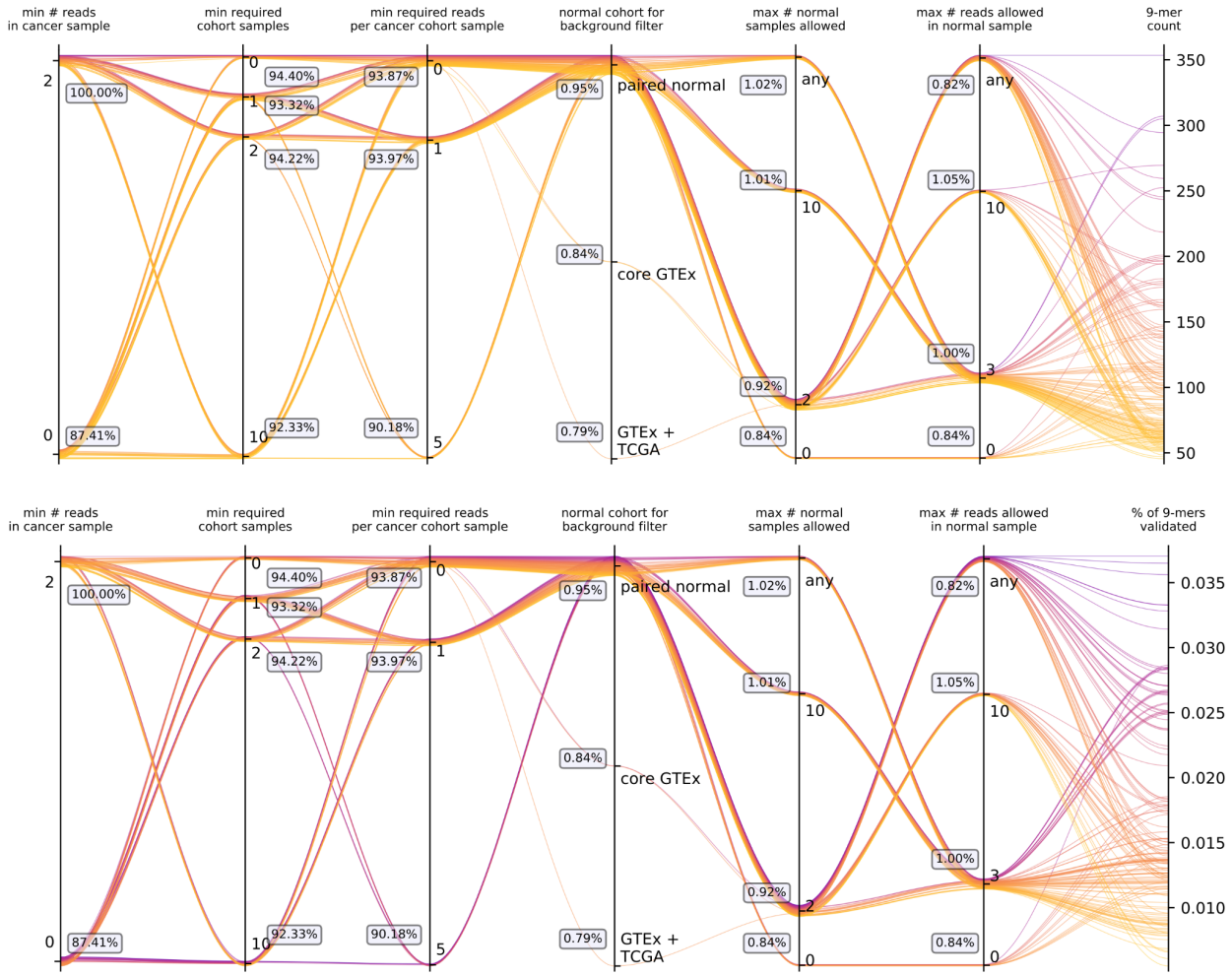


Figure 3.5: Effect of JP filters on validated 9-mer count and validation ratios for BRCA sample TCGA-C8-A12P. Vertical axes except the rightmost represent one filter, with parameter options from most stringent (bottom) to most lenient (top). Each colored line shows one filter experiment, with its path passing through the parameters it uses, its color mapped to its validated 9-mer count (top) or validation ratio (bottom) (yellow = low, purple = high), and terminating at its final value on the rightmost axes. Floating boxes show, for experiments passing through the corresponding filter parameter, the mean of the ratio of remaining 9-mers after that filter parameter has been applied to the total initial sample 9-mers.

4.4 Discussion

With the goal of identifying cancer-specific peptides arising from real junctions (vs. from sequencing or alignment artifacts), we performed a set of cancer-specific splicing discovery experiments across many sets of filter parameters and two fundamentally different pipelines. The junction-based method takes a simpler and more stringent approach, where all junctions are

assumed to be associated with the annotated transcriptome. No context is assessed between short reads, so that any called junction may be associated with any transcript that has appropriate coordinates. This yields a fast and agile pipeline that may neglect some critical contextual information, but also may avoid detection of peptides from spurious transcripts. In contrast, the graph-based method allows for the discovery and use of novel, unannotated transcripts, and accounts for transcript context by phasing input short reads. A drawback of this method is the significant compute resources & time required for implementation. (Benchmarking timing studies were not performed, but an informal analysis shows that matched tasks performed by the JP are several orders of magnitude faster than by the GP using similar computing resources with similar parallelization.)

The discrepancy in number of junctions translated by the two pipelines is partly due to the GP assembling and translating junction peptides from novel transcripts, where the JP translates junctions for which the 5' splice site falls in the coding region of an annotated transcript and rejects translations where the junction peptide is less than 9 amino acids long or the junction creates a stop codon. Non-matching peptide-junction pairs from mutually translated junctions arise from the use of additional reading frames, primarily in novel transcripts generated by the GP. Many GP junction peptides contain degenerate sequence in the junction region, leading to the multiple order of magnitude drop in total to unique tryptic peptides in the GP. The GP's small range of filtered pASNs counts may be partially related to the limited experiments performed, covering only one normal cohort (core GTEx) and not including the motif filter, and to the initial cohort graph support required. However, the GP's lack of output sensitivity to the varying filter parameters was unexpected, as was the extreme similarity in output between samples.

This study was limited by the small number of samples and cancer types studied in depth. We also were limited computationally and by analysis bandwidth to a relatively small set of filter parameters, although we attempted to cover the range of reasonable choices. Implementation of the junction support and motif filters differed between the GP and the JP. We did not take into account potential somatic mutations occurring in the cancer samples, some of which (e.g. upstream frameshift indels) may have affected the translated junction peptide sequences. We used a large but still limited set of normal samples and tissue types in our normal cohorts, leaving open the possibility that some normal splicing may be declared "cancer specific" here.[1] Finally, we additionally did not test MHC binding of our pASNs and our MS queries were limited to intracellular data available from CPTAC instead of surface-presented peptides, so we have no information about the pASNs' functionality as neoantigens.

Ultimately, the large proportion of filtered 9-mers arising from junctions with non-canonical splice motifs identified by the JP (Supplementary Table 3.4) calls into question the biological significance of filtered junction peptides. The lack of overlap between the two pipelines' outputs (Supplementary Figure 3.4), and the fact that each pipeline's pASNs are, for the most part, novelly generated by only that pipeline (Table 3.1), suggests that filter results may include a set of nonbiological artifacts of the alignment, translation, and filtering methods. Most junction-overlapping peptides initially generated are shared between both pipelines, but these may largely arise from transcripts that are easy to handle (e.g. annotated transcripts), with fewer pipeline-specific assumptions imposed on peptide generation. Finally, we note the instability of MS validation of identical peptides between filter experiments, and that the distribution of splice motifs of validated pASNs (Figure 3.2, Supplementary Table 3.5) suggests a high incidence of false positive MS detections.

3.4 Methods

Experimental overview & filter parameters

Table 3.2: Filter parameter values across pipelines.

		Filter Stringency			
Filter type	**pipeline**	**lowest**	**Low2**	**high**	**highest**
sample support	GP		graph confidence 1	graph confidence 2	
	JP		1 normalized read/sample	2 normalized reads/sample	
motif	GP		no filter		
	JP		no filter	canonical motifs only	
minimum matched cohort samples	both	0	1	2	10
Min reads per cohort sample	both	0	1	5	
normal cohort	GP			core GTEx	all GTEx + TCGA
	JP		paired normal	core GTEx	all GTEx + TCGA
Max normal samples	both	any*	10	2	0
Max reads per normal sample	both	any*	10	3	
proteome	both	resulting 9-mers must not be present in Uniprot			

* "any" filters were applied in conjunction with a limiting value for the complementary filter, e.g. 3 samples with any expression, or any number of samples with 2 or fewer normalized reads.

We analyzed the output of two pipelines, junction- (JP) and graph- (GP) based, that aim at detecting cancer-specific peptides arising from exon-exon junctions. Each pipeline executes multiple filters that allow for a range of parameters, which we explored across the space of parameter options by analyzing their effects on final outcomes (Table 3.2). Five independent

filters fall broadly into two categories, those supporting the detection of true junctions, and those supporting the cancer-specificity of the resulting peptides, as follows.

Junction support filters include 1) RNA expression in the target sample, 2) RNA expression in a matched cancer type cohort, and 3) a filter on junction splice motif (implemented in the JP only). In the JP, junctions are represented by coordinates, while in the GP, junctions are represented by junction-overlapping 9-mers; in both pipelines, RNA expression quantification is based on junction-overlapping primary aligned reads. For filter 1), target sample minimum RNA expression thresholds were set to "any" (>0) or 2 normalized junction reads. For filter 2), a matched sample-type cohort is defined as the full set of TCGA matched tumor samples (breast or ovarian cancer, as appropriate). Junctions were then required to be expressed in 0 (no cohort requirement), or 1, 2, or 10 cohort samples, with 1 or 5 normalized reads required in each external sample. For filter 3), the JP toggles a requirement (or not) for the junction's splice motif to be canonical (GT-AG, GC-AG, or AT-AC).

Cancer-specificity filters comprise 4) full or partial absence of the junction from a set of normal samples, and 5) absence of resulting peptides from a normal human proteome database. For filter 4), three normal tissue cohorts were established to represent the breadth of normal alternative splicing. The more lenient set included paired normal samples only, comprising all normal tissue-matched samples across GTEx and TCGA, called the "paired normal" cohort. A more stringent set included all normal samples across all GTEx tissue types except for immune privileged tissues (testis and brain), called the "core GTEx" cohort. Finally, the most stringent normal cohort was all normal samples across GTEx and TCGA, called the "all GTEx + TCGA" cohort. Maximum RNA expression criteria in the normal cohort were established where a junction needed to be expressed in no more than a maximum number of samples across the

normal cohort, with a maximum expression level within each of the allowed normal samples. In the JP cancer junctions are filtered against normal junctions, while in the GP overlapping junction 9-mers are filtered against all cohort 9-mers, regardless of junction presence. A set of "lenient" conditions values (maximum sample count, maximum expression level per sample) were set as (2, 3), (2, 10), (2, any), (10, 3), and (any, 3). A "stringent" condition was also used, in which no expression across the normal cohort was allowed. Finally, the proteome filter 5) has no free parameters. In the GP, this filtering was performed with Leucine and Isoleucine equivalent.

Selection of target samples

Five samples from each of the TCGA breast and ovarian cancer cohorts were selected as follows. These two cancer types were chosen for having paired intracellular proteomics mass spectrometry CPTAC[262] data for pASN validation, and for having been previously used for ASN detection and validation.[156] Shortlist subsets of these cohorts were selected, including samples for which mass spectrometry intracellular proteomics analysis was done (as identified from the CPTAC[262] iTRAQ sample file) and which previously passed quality control protocols.[156] Five samples were then selected from each shortlist in a reproducibly random way, using the absolute value of the murmurhash (implemented with mmh3 v.2.4) of each TCGA cancer name string to seed the random selection for that cancer type, using python the random.sample method (python v.3.8.1).

Reference data download

A GTEx v6 phenotype table was downloaded from recount2[197] at https://jhubiostatistics.shinyapps.io/recount. TCGA phenotype data was obtained from the GDC at https://portal.gdc.cancer.gov/repository. GENCODE v.32 gene annotations[49] were downloaded from https://www.gencodegenes.org/human/release_32.html for use with the GRCH38 reference genome.

The human proteome was downloaded from Uniprot[241] (https://www.uniprot.org/proteomes /UP000005640).

Sequencing data download and initial preparation

All data was downloaded from the GDC data portal (https://portal.gdc.cancer.gov/), using a complete metadata dump from January 27th 2020. Download of all samples followed in the two weeks after. Alignments were performed with STAR, with the following parameters:

--sjdbOverhang 100, --runThreadN 8, --outFilterMultimapScoreRange 1, --outFilterMultimapNmax 20, --outFilterMismatchNmax 10, --alignIntronMax 500000, --alignMatesGapMax 1000000, --sjdbScore 2, --alignSJDBoverhangMin 1, --genomeLoad NoSharedMemory, --readFilesCommand zcat, --outFilterMatchNminOverLread 0.33, --outFilterScoreMinOverLread 0.33, --outSAMstrandField intronMotif, --outSAMmode Full, --limitBAMsortRAM 7000000000, --outSAMattributes NH HI NM MD AS XS, --outSAMunmapped Within, --limitSjdbInsertNsj 2000000, --outSAMtype BAM Unsorted, --outSAMheaderHD @HD VN:1.4, --outSAMmultNmax 1"

The library sizes are computed based on sequential STAR and SplAdder[219] quantifications. (SplAdder, part of the GP, will be described in the next sections.) Expression quantification is first performed by STAR, then passed on to SplAdder. The read support for each portion of the SplAdder graph in each sample is stored after discarding reads from regions with low coverage. These quantifications are used to calculate protein coding gene expressions. This set of genes is used to extract 75th quantile library sizes for samples of each cohort.

Data preparation for junction-based pipeline

GTEx and TCGA sample IDs were mapped to tissue and sample types using the GTEx and TCGA phenotype tables, respectively. All exon-exon junctions with sample coverage >0 were extracted from the SplAdder graphs generated for all GTEx and TCGA ovarian and breast cancer samples. All junction coverages were normalized by dividing by the 75th quantile library sizes (as calculated above) and multiplying by $4x10^5$. (Four BRCA and one OV sample had a

single read scale normalized to between 1 and 2; the remaining BRCA and OV samples had one read scale to over 2 and less than 1, respectively.) All junctions with nonzero coverage were collected for the 10 target samples, comprising a shortlist set of 1,347,000 junctions. For each shortlist junction, counts were collected of cohort samples with any nonzero coverage and of samples with coverage over each target threshold in the relevant filter type for cancer type, paired normal, and core GTEx sample cohorts. Splice site motifs were extracted for each junction from the reference genome with samtools v1.9.[263] Annotated left, right, and matched splice site coordinates were extracted from the GENCODE gtf, and sample junction splice site coordinates compared against these, with labels 3, 2, 1, and 0 respectively representing annotated, exon-skip annotated, left- or right- splice site-only annotated, or unannotated junctions. Genome coordinates for protein coding regions were extracted from the GENCODE gtf, and each junction was labeled with potential gene IDs for any protein coding genes overlapping its left and right splice sites. Each junction was further labeled with all transcript IDs for which its upstream, 5' splice site overlapped the transcript's CDS.

Translation of MS query peptides and junction-overlapping kmers via junction-based pipeline

A custom class was written for junction translation, i.e. to translate junction-overlapping reference DNA sequence into junction-overlapping protein sequence. For each junction, all transcripts were previously identified for which the junction's 5' splice site was located within the transcripts' coding regions. For BRCA and OV respectively, 20% and 31% of total junctions on average are not located in protein coding genes, and 26% and 22% on average are in protein coding genes but are untranslated for other reasons such as having upstream splice sites not located in the protein coding boundaries of an exon (Supplementary Table 3.2). Each junction-transcript pair was then processed as follows. The junction was computationally inserted into the

transcript, with the junction's upstream splice site either truncating the upstream exon in which it was located, or with no change if it already matched an annotated 5' splice site in the transcript. The junction's downstream splice site would then fall into one of the following categories: 1) matching an annotated 3' splice site in the transcript, in which case no changes were made to exon coordinates; 2) falling in the middle of an annotated exon, in which the upstream end of the exon was truncated so that the exon's new 5' end coincided with the 3' end of the junction; 3) falling between exons in the coding region of the transcript, in which case the 5' end of the exon immediately downstream from the junction is adjusted to match the junction's 3' end; or 4) falling beyond the end of the transcript's CDS, in which case an artificial exon was added to the end of the transcript, 150 bases long and whose 5' end matched the junction's 3' end. In every case, all exons in the junction boundaries were removed from the transcript.

Each artificial junction-modified transcript was then translated as follows, with a target length of 50 amino acids up- and downstream of the junction to maximize the possibility of obtaining a trypsin-digested junction-overlapping peptide size-appropriate for mass spectrometry queries. Therefore a target sequence length of 150 bases up- and downstream of the junction was collected, or fewer bases if the end of the transcript was reached, as well as the translation reading frame as propagated from the 5' end of the original annotated transcript's coding region. The sequence was then translated *in silico* in all three reading frames for potential reading frame-agnostic peptide analysis, with the annotated reading frame, the junction position within the peptide, 5' or 3' transcript end overlap, and whether or not the junction split a codon or occurred within a codon noted for future use. The immediate peptide sequence around the junction was kmerized into up to 9 junction-overlapping 9-mers. The Uniprot human proteome was kmerized

into 9-mers, and the junction 9-mers not occurring in the normal human proteome set labeled as potential neoepitopes to undergo further filtering.

Identification and filtering of cancer-specific peptides via junction-based pipeline

For each target sample, the set of filter experiments described above was performed as follows. Junctions with nonzero expression in the target sample were collected; any junction falling outside of the target gene list accessible to the graph-based pipeline was removed. Filters were applied to the set of remaining junctions in a single order: target sample expression support, foreground cohort-matched sample support, background normal cohort normal filtering, and canonical splice motif identity. At each filter stage, the number of unique junction-overlapping 9-mers was collected; any 9-mer arising from more than one filter-passing junction was counted only once. After the junction filtering was complete, a Uniprot normal human proteome filter was applied, to collect the number of unique potential neoepitope 9-mers. Finally, a list of potential sample-specific junction peptides resulting from each filter experiment was collected.

Data preparation for graph-based pipeline

The RNA-seq data was used to build cohort-wise splicing graphs with SplAdder. Foreground graphs for OV and BRCA were built, as well as background graphs for GTEx and TCGA, with STAR alignments of RNA-seq data and annotation file as input. Graph segments represent exons and edges represent exon-exon junctions. The annotation graph is then expanded sample-wise with the exons, introns, and junctions supported by the sample's RNA-seq. The graphs are then merged across samples into a joint graph which recapitulates splicing across the cohort. Cross-sample filtering criteria are applied at the level of the graph. In graph "confidence level" 2,[219] first, an intron is added to the graph if it has at least 5 average reads per nucleotide position in any of the cancer cohort samples, a sufficient fraction of intron positions are covered,

and the intron coverage relative to flanking exons falls within a defined range. Secondly, an unannotated edge is added if there are at least 2 reads supporting that edge in any of the cancer cohort samples, and if that edge meets anchor length requirements with low mismatches in the anchor regions. Given these inclusion criteria, each junction is associated with the number of primary reads aligned overlapping the junction. This metric is used later to quantify 9-mers.

Translation of MS query peptides and junction-overlapping kmers via graph-based pipeline

Each of the cohort graphs were translated with the ImmunoPepper (unpublished) software tool (https://github.com/ratschlab/immunopepper). ImmunoPepper is an algorithm which traverses a splicing graph to extract bi-exon peptides. In the following we refer to a single gene graph with cross sample information. First, the annotated CDSs of the gene were collected. Then, these were applied onto the graph and their reading frames were propagated downstream in the graph. Finally, all possible exon-exon pairs were extracted and translated according to the propagated CDSs. The translation was stopped when encountering a stop codon. The bi-exon peptides were cut into 9-mers. In the event of a bi-exon pair where the amino acid length of the second exon is smaller than the kmer length, the bi-exon peptide was expanded to the right with a third exon. For each sample, 9-mers were then quantified based on their region of origin. Junction overlapping 9-mers were quantified by the number of reads spanning the junction (edge expression value). Non-junction 9-mers were mapped to the segments from the splicing graph that span the 9-mer positions. Each segment structure in the splicing graph stores an expression value extracted from the alignment. The expression of the 9-mer was calculated as a sum of the different segment expressions weighted by the number of base pairs which originate from this segment (segment expression value). All 9-mers expressions were normalized by dividing by the 75th quantile library sizes (see above) and multiplying by $4x10^5$.

Identification and filtering of cancer-specific peptides via graph-based pipeline

The filtering was performed at the level of the 9-mers peptide sequences. For each target sample, we consider as the initial set, the junction overlapping 9-mers translated from the OV or BRCA graphs (which share information across the cohort), minus the 9-mers not expressed in the sample. For synonymous 9-mers the maximum expression value was used. These junction 9-mers were then filtered for target sample expression, and cancer-type matched cohort support, as described previously. For each background cohort, we extracted both the junction, and non-junction overlapping 9-mers from the graphs, as we wish to remove all normal 9-mers passing the expression and sample criteria. Duplicated 9-mers were made unique by taking the maximum expression across junction and non-junction instances. The background normal cohort normal filter was applied as outlined previously. The annotated 9-mers were removed to focus on novel 9-mers. Finally, the 9-mers were filtered out, Isoleucine and Leucine were made equivalent by substituting isoleucine characters by leucines.

Comparison of translation between GP and JP

Peptides from junctions mutually translated by both pipelines were compared. Each peptide may arise from multiple junctions, and the same junction may give rise to multiple peptides, so the comparison was done for each peptide-junction pair, for each pipeline. A match was called for a pair in a pipeline if any of the following were true: 1) an exact match between the peptide and a peptide generated for the same junction in the opposite pipeline; 2) the peptide falling wholly within a peptide generated for the same junction in the opposite pipeline; or 3) the beginning or the end of the peptide overlapping with, respectively, the end or the beginning of a peptide generated for the same junction in the opposite pipeline.

Proteomics validation of cancer-specific peptides with subset-neighbor search

Following the translation of transcript to amino acid sequence, we then attempted to validate the existence of the cancer-specific peptides via proteomics analysis. In proteomics analysis, peptides are detected via database searching with experimental spectra and a peptide database given as input. Since the samples analyzed by mass spectrometry were digested using trypsin, the input peptide database must also contain tryptic peptides. Therefore, we extract a fully tryptic peptide from each transcript. This process was successful if the junction-spanning amino acid was between two tryptic sites. In addition, the process was successful if the junction-spanning amino acid was between the beginning of the protein and the first tryptic site or the last tryptic site and the end of a protein. After this process we removed any peptides shorter than six or longer than 50 amino acids as these peptides are unlikely to be detected by proteomics. For each pipeline and cancer sample, we extracted the full set of tryptic peptides arising from transcripts from any filter experiment, encompassing the most lenient possible set of pASNs. We then concatenated these sequences to the human reference protein database for use in a database search. The human proteome was downloaded from Uniprot (https://www.uniprot.org /proteomes/UP000005640)[241] on January 18th, 2022. In January 2022 we also downloaded 124 raw mass spectrometry data from the Proteomics Data Commons (https://pdc.cancer.gov/pdc/ study/PDC000173) to use in the database search. Each of the raw files was converted to mgf file format using ThermoRawFileParser.[264]

To detect peptides in our proteomics samples, we employed the subset-neighbor database search strategy[260] using the XCorr score in Crux (version 3.2).[265] Cancer-specific peptides were considered relevant while reference human peptides were considered to be irrelevant. An irrelevant peptide was relabeled as a neighbor peptide if the precursor mass was within 40ppm of

a relevant peptide and if the two peptides have at least 25% of fragment peaks in common. Peptides that were found to be both relevant and irrelevant were considered relevant. The modification used in the construction of our database included carbamidomethylation as well as a static iTRAQ labeling on lysines and the N-terminus of peptides. The precursor tolerance was set to 40ppm and was estimated using Param-medic.[266] All other parameters were set to their default value. The false discovery rate of the resulting PSMs was estimated using target-decoy competition[267] and filtered to a 5% FDR. FDR correction was applied to the specific set of peptides arising from each filter experiment individually, for each sample and pipeline. The resulting list of peptides were labeled as the set of confidently detected peptides.

Acknowledgements

We thank Ryan Kuck for his helpful advice on writing the novel junction-modified transcript class.

Chapter 4: Testing the reliability of retained intron detection

This work has been formatted for inclusion in this dissertation from the manuscript "Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads" by Julianne K. David*, Sean K. Maden*, Mary A. Wood, Reid F. Thompson, and Abhinav Nellore, under review (2022).[2] The author of this dissertation is the co-primary author of the manuscript. (*co-first authors)

4.1 Abstract

A number of bioinformatics tools have been developed specifically to detect retained introns (RIs) from short-read RNA sequencing (RNA-seq) data, and they have been used to make confident statements about retained introns across a variety of biological circumstances. However, overlapping genes and transcripts and the presence of partially processed RNA in sequenced samples can lead to uncertainty in the detection of RIs, particularly from short-read data. We assembled a dataset to test RI detection, consisting of complementary publicly available short- and deep long-read RNA-seq data from the same biological specimens. Then we evaluated 5 short-read RI detection tools and found 1) significant disagreement (Fleiss' $\kappa = 0.231$) such that ~52% of called RIs were called by single tools only; 2) that no tool achieved greater than 20% precision or 35% recall under generous conditions; and 3) that RI detectability was adversely affected by greater intron length and overlap with annotated exons.

4.2 Background

During RNA transcription, multiple spliceosomes may act on the same transcript in parallel to remove segments of sequence called introns and splice together flanking exons.[97] Most splicing occurs stochastically[104] during transcription,[56–58] although up to 20% of splicing may occur after transcription and polyadenylation[58,268] (Supplementary Figure S4.1). Introns are spliced by several known spliceosome types, of which the most studied are called U2 and

U12.[269] Splicing is known to occur primarily in the nucleus,[270] though there is evidence of cytoplasmic splicing.[60,62,271,272]

Intron retention (IR) is a form of alternative splicing where an intron normally spliced out during transcript processing remains after processing is complete. IR occurs in up to 80% of protein-coding genes in humans[76] and may affect gene expression regulation[80–86] as well as response to stress.[90–92] Transcripts containing introns may also be stably detained in the nucleus before undergoing delayed splicing ("intron detention", or ID), with implications for temporal gene expression.[94] In cancers, high levels of IR[150–152] can generate aberrant splicing products with known and potential biological consequences for gene expression and cell survival.[125] IR rarely gives rise to a protein product,[79,93] but novel peptides derived from transcripts with RIs are increasingly being studied in disease contexts such as cancer.[154–157,273]

Despite its biological relevance, detection of IR from bulk RNA-seq data remains challenging for two principal reasons: 1) A short RNA-seq read (e.g., from Illumina's HiSeq, NovaSeq, or MiSeq platforms) is almost never long enough to resolve a full intron or its context in a transcript, particularly in genome regions with multiple overlapping transcripts; 2) RNA-seq data may contain intronic sequence from unprocessed or partially processed transcripts, DNA contamination, and non-messenger RNA such as circular RNAs (cRNAs),[57,209] potentially yielding spurious IR calls, independent of read length.

Existing tools designed specifically for RI detection make simplifying assumptions to address the above issues. These tools include keep me around (KMA),[212] IntEREst,[214] iREAD,[215] superintronic,[216] and IRFinder[76] and its most recent implementation as IRFinder-S.[217] Some mitigate the first challenge by ignoring from consideration any intronic regions that overlap other

features (KMA, IntEREst, iREAD), leaving biological blind spots in RI detection.[212,214,215] Some attempt to mitigate the second challenge by recommending that a user provides poly(A)-selected data as their input,[76,212,215,216] assuming that poly(A) selected data represents fully processed, mature RNA. However, poly(A) selection during library preparation has been shown not to remove all immature post-transcriptionally spliced RNA molecules, and intronic sequences are commonly found in poly(A)-selected RNA-sequencing data.[54,75] To clarify the quality of and best practices for RI detection, we performed tests on poly(A)-selected, sample-matched long- and short-read sequencing runs for two biological specimens, with processed long-read data providing a standard against which we evaluated short read-based RI detection.

4.3 Results

4.3.1 *Sample-paired short- and deep long-read RNA-seq data can robustly test RI detection*

To generate a dataset to test RI detection, we identified two human biological specimens on the SRA with RNA-seq data from both Illumina short-read (SR) and PacBio Iso-Seq RS II long-read (LR) platforms (Figure 4.1). These were a human whole blood sample (HX1)[274] and a human induced pluripotent stem cell line sample (iPSC),[275] with, respectively, 46 and 27 Iso-Seq runs, 24.4 and 91.3 million aligned short reads, and 945 and 840 thousand aligned long reads (Supplementary Table S4.1). To confine attention to robustly represented loci, we identified a set of 4,369 and 4,639 target genes in HX1 and iPSC samples, respectively, each with ≥2 short reads per base median coverage across the full gene length and ≥5 long reads assigned to at least one isoform of the gene (Supplementary Figure S4.2).

We sought to quantify IR in each biological specimen using LR data, accounting for random splicing and sample contamination that may lead to noisy splicing patterns. For a given intron i and transcript *t*, we defined persistence $P_{i,t}$ as

$$P_{i,t} = d_i \cdot \sum_{\{r \in M^t\}} \frac{R_{r,i} \cdot SF_{r,i} \cdot H_{r,i}}{|M^t|}, \qquad (4.1)$$

where r is a read among the set of all reads M^t assigned as best matches to transcript t, information density d_i is the proportion of M^t covering intron i, the binary variable $R_{r,i}$ is 1 if and only if r provides evidence for the retention of i, and the spliced fraction $SF_{r,i}$ and scaled Hamming similarity $H_{r,i}$ are defined in Methods (see Equations 4.3 and 4.4). In brief, the intron persistence $P_{i,t}$ incorporates the extent and similarity of splicing across transcript reads, accounting for stochastic splicing initiation and progression (Supplementary Figure S4.1). Finally, to address ambiguity in transcripts of origin in short-read data, we defined intron i's persistence P_i as its maximum persistence across all isoforms T_i that contain i:

$$P_i = max_{t \in T_i}\left(P_{i,t}\right). \qquad (4.2)$$

Going forward, we define a “persistent intron" as an intron for which $P_i >= 0.1$.

Across all transcripts studied in both samples, a substantial majority (83.7%) of introns were fully spliced out ($P_{i,t} = 0$), and a small minority (0.15%) of introns were always unspliced within a transcript ($P_{i,t} = 1$) (Figure 4.2A and Supplementary Figure S4.3). These extreme values are in keeping with our qualitative understanding of splicing patterns; however, the range of intermediate persistence values appears to represent a spectrum with varying extents of inconsistent splicing across and between reads. While we tested short-read RI detection on a per-sample basis, we also compared intron persistence patterns between HX1 and iPSC samples and found significant similarity in splicing patterns across matched transcripts (Supplementary Figures S4.3 and S4.4).

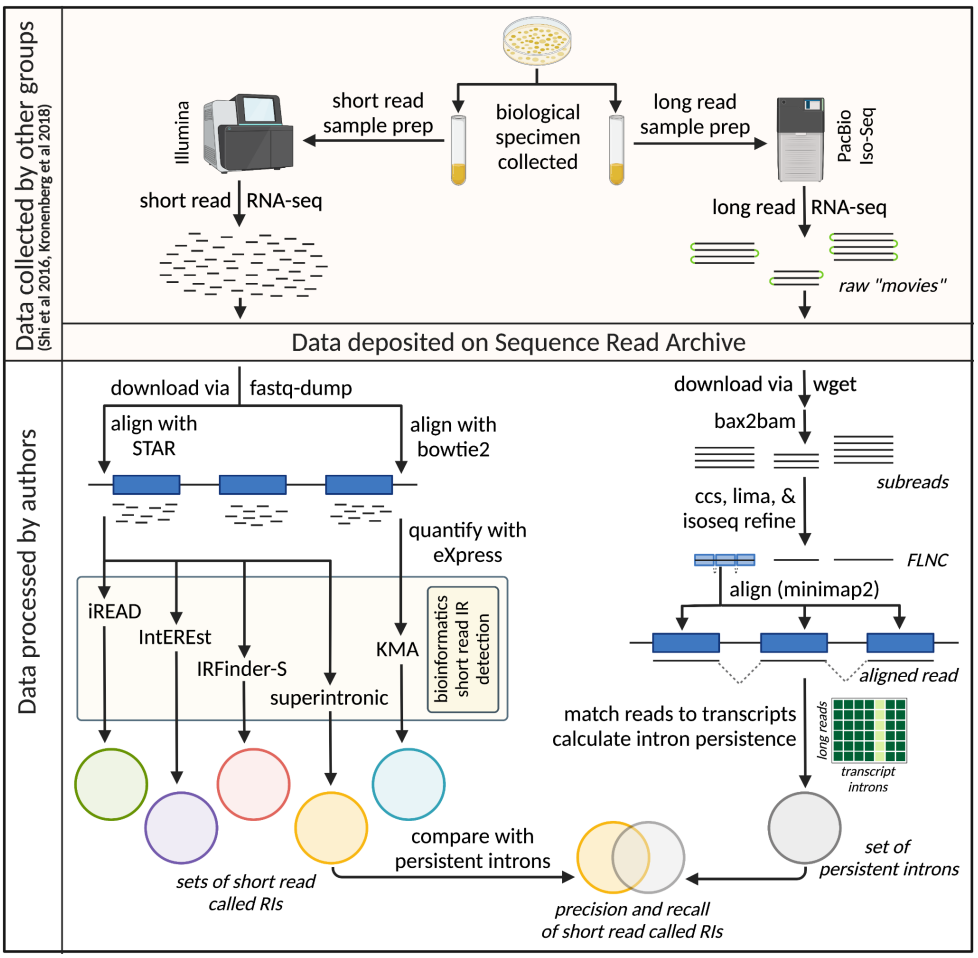


Figure 4.1: Overview of experimental plan. Created with BioRender.com. Long and short read RNA-seq data from the same biological specimen[274,275] were downloaded from the SRA and subject to processing and analysis. Short reads (left path) were aligned and quantified according to the requirements of five short read RI detection tools,[212,214–217] and retained introns were called by each of these. The raw long Iso-Seq reads (right path) were processed to the stage of full-length non-concatemer (FLNC) reads, but left unclustered. After long reads were aligned to the reference genome, each aligned read was assigned to a best match transcript or discarded, and intron persistence was calculated. The called RI output of each short read detection tool was compared against the set of persistent introns identified in the long-read data (where $P_i >= 0.1$).

4.3.2 *Intron properties explain similarities across short-read RI detection tool outputs*

We compared RIs called by five detection tools for short-read data (Table 4.1). While most introns were consistently spliced out, 39.9% (1,743/4,369) and 31.4% (1,457/4,639) of target genes in HX1 and iPSC, respectively, had at least one RI identified in either short- or long-read data. Expression of called RIs varied substantially between tools in both HX1 (Fleiss' $\kappa =$

0.282) and iPSC (Fleiss' $\kappa = 0.162$), though we did observe moderate overall correlation between the output of IntEREst, superintronic, and KMA (Supplementary Figure S4.5). Further, using circBase[276] to probe whether cRNA contamination may have affected RI detection, we identified only a small percent (<5%) of called RIs that appeared to overlap intronic cRNAs (Supplementary Figure S4.6).

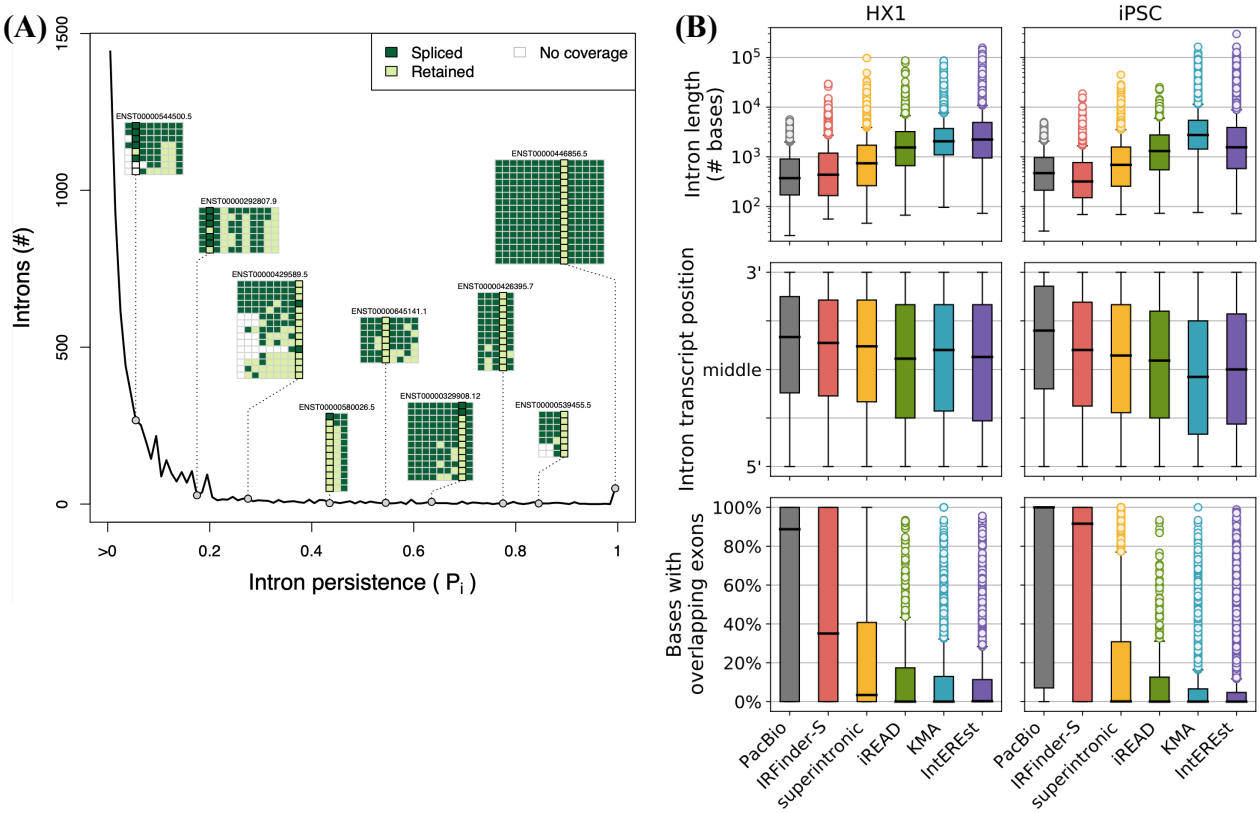


Figure 4.2: Intron persistence and other properties. (A) Distribution of intron persistence ($P_{i,t}$) and representative transcript examples for iPSC. The black line shows the number of introns (y-axis) having a given persistence value (x-axis); note that a large number of introns with $P_i = 0$ are omitted from this analysis. Along the line, gray circles indicate the P_i value corresponding to each of nine introns from representative transcript examples with each transcript labeled by Ensembl ID. Read-level data is shown for each transcript as a colored matrix, where each row is a single long read assigned to the transcript and each column represents a given intron, and color indicates whether an intron is retained (light green), spliced out (dark green), or lacking sequence coverage (white) in a given read. **(B)** Distributions of properties of persistent and called RIs. Each panel contains a series of boxplots depicting the distribution of intron length (top, log-scale), relative position in transcript (middle), and % of intron bases with overlapping annotated exons (bottom) for HX1 (left) and iPSC (right). The distribution of each of these features is shown for long-read persistent introns ("PacBio", gray) and RIs called by each of the five short read tools: IRFinder-S (red), superintronic (yellow), iREAD (green), KMA (blue), IntEREst (purple).

Table 4.1. Short read tools studied.

Tool	IRFinder-S[217]	superintronic[216]	iREAD[215]	KMA[212]	IntEREst[214]
Year	2021	2020	2020	2015	2018
IR measure†	IRratio	$\log_2$ coverage	FPKM	TPM	FPKM or PSI
Language	C++	R	Python	Python, R	R
Host website	GitHub	GitHub	GitHub	GitHub	Bioconductor
Sample data format	BAM or FASTQ	BAM	BAM	FASTQ	BAM
Reference format	GTF	GTF/GFF3	BED	FASTA, GTF/GFF3	GTF/GFF3
Intron definition	All introns	All introns	Independent introns*	Independent introns*	Independent introns*

***** Independent introns are intron regions not overlapping features from other annotated transcripts.
† See Methods for IR measure definitions and details

We next examined the distributions of several intron properties (length, relative position in transcript, and annotated exon overlap) and their relationships with the set of RIs called by each short-read tool and their relative expression levels (Figure 4.2B, Supplementary Figure S4.7). Unsurprisingly, tools that exclude introns with overlapping genomic features (i.e. KMA, IntEREst, iREAD; Table 4.1) had exceedingly low overlap between exons and the IRs they reported. We also note that KMA and IntEREst called extremely long RIs (up to >297 kilobases), compared to those called by other short-read tools or the persistent introns identified from long read data (maximum 6,275 and 5,926 bases in HX1 and iPSC). We observed a slight overall 3' bias among persistent introns from long-read data, as well as the set of RIs from several short-read tools (Figure 4.2B), potentially reflecting the relatively shorter duration of exposure of 3' introns to the cotranscriptional splicing machinery and/or implicit 3' bias of the Clontech sample prep[277] used in both samples.[274,275] Despite this slight 3' tendency, there was no appreciable association between intron persistence and intron position in transcript

(Supplementary Figure S4.8). Among all tools, IRFinder-S called a set of RIs with characteristics most similar to persistent introns from long-read data (Figure 4.2B).

4.3.3 *Precision and recall are poor across short-read RI detection tools*

We tested performance (precision, recall, and F1-score) of RI detection by five short-read tools, comparing sets of called RIs against persistent introns identified from long read data (defined as $P_i \geq 0.1$). Overall tool performance was poor in all cases (Figure 4.3A, Supplementary Table S4.2). Many persistent introns (55% and 48% in iPSC and HX1, respectively, Supplementary Figure S4.9) were not called by any short-read tool, and the majority of called RIs were neither identified among persistent introns in long-read data nor consistently called between short-read tools (Figure 4.3B, Supplementary Figure S4.9). In HX1 and iPSC, respectively, 54% and 49% of called RIs were not called by more than one tool (52.4% overall). IRFinder-S had the best performance across most metrics, possibly due to the similarity between the properties of its called RIs and properties of persistent introns. By contrast, iREAD demonstrated the lowest recall across all tools, likely due to its sparse calling of RIs (Supplementary Figure S4.10). Performance metrics for IntEREst and KMA were very similar across both samples (Figure 4.3C).

To address sensitivity in persistent intron identification, we also considered short-read tool performance on subsets of LR introns with increasing minimum thresholds of intron persistence ($P_i >= 0.1 - 0.9$ in 10% increments). We found that overall performance remained poor across all levels of intron persistence, with uniformly worse precision, recall and F1 score as intron persistence increased (Figure 4.3A, Supplementary Figure S4.11). While individual tool performance varied significantly, IRFinder-S and superintronic were consistently best performers, albeit interchangeably depending on the sample, metric assessed, and intron

persistence threshold. For instance, IRFinder-S demonstrated highest recall in HX1 at the lowest cutoff values ($P_i >= 0.1 - 0.4$), while superintronic demonstrated higher recall across higher thresholds in HX1 and for all cutoffs in iPSC (Supplementary Table S4.2).

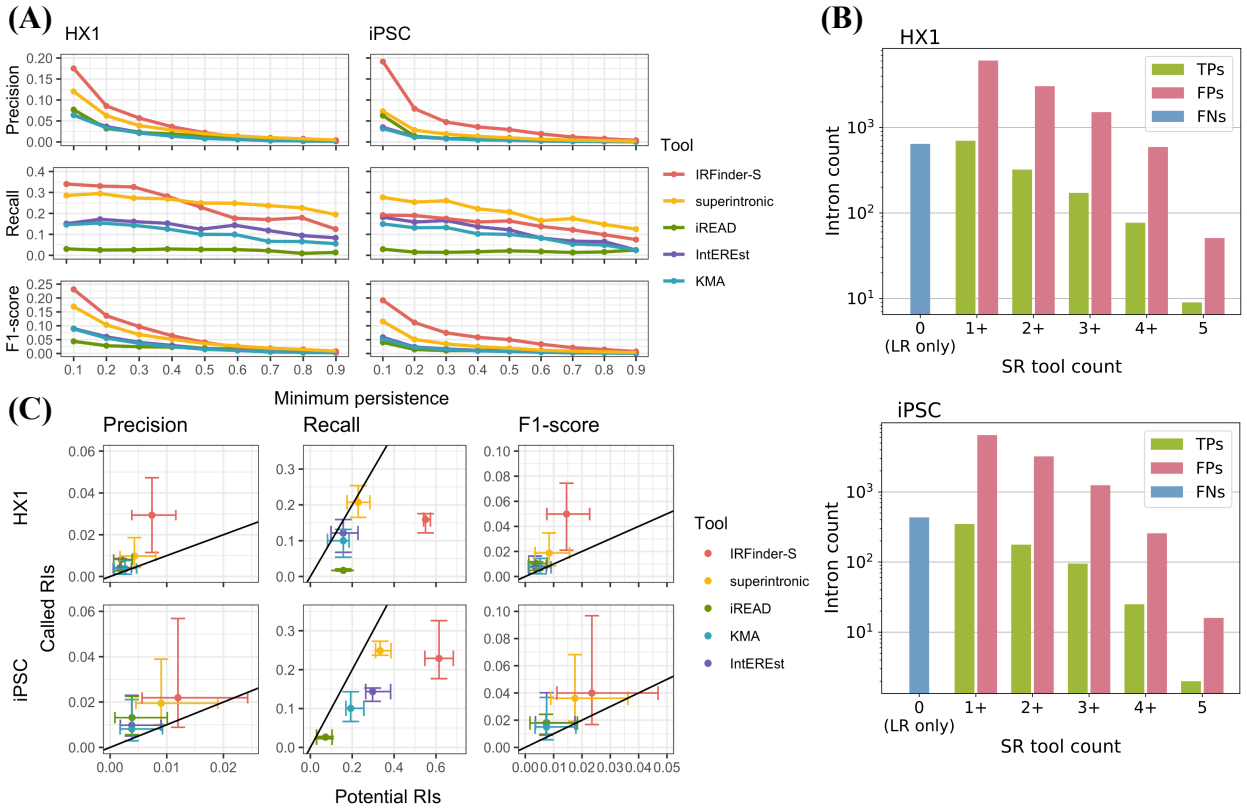


Figure 4.3: Performance of short read tools. (A.) Short-read tool performance across different thresholds of intron persistence. Each panel displays tool performance along the y-axis (measured by either precision, recall, or F1-score as labeled) for a set of introns defined by the indicated threshold for intron persistence along the x-axis. Data for HX1 and iPSC are shown at left and right, respectively, with each tool's per-sample performance depicted in a different color (IRFinder-S [red], superintronic [yellow], iREAD [green], IntEREst [purple], and KMA [blue]). **(B.)** Varying degrees of consensus of retained intron calls among short-read tools. Bar plots depict the number of true positive (green), false positive (pink), and false negative (blue) intron calls (y-axis) consistent across a specified number of short-read (SR) tools (x-axis). Upper and lower panels depict HX1 and iPSC data, respectively. LR denotes long-read data. **(C.)** Variation in short-read tool performance across intron persistence thresholds for potential vs. called RIs. Each panel displays tool performance as measured by precision (left), recall (middle), and F1-score (right) for HX1 (top) and iPSC (bottom) samples. The performances for each tool's potential RIs and called RIs are shown along the x- and y-axes, respectively, with centroid and whiskers denoting, respectively, the median and interquartile range of tool performance across intron persistence thresholds. Tools are depicted by color (IRFinder-S [red], superintronic [yellow], iREAD [green], IntEREst [purple], and KMA [blue]). Reference lines with slope = 1 are shown.

Finally, since each tool is capable of calling RIs with different levels of stringency, we evaluated tool performance on a raw set of all potential RIs (all expressed introns detected by that tool) vs. the corresponding subset of introns called as RIs by that tool. Rather than improving overall performance by retaining persistent RIs and removing false positive ones, stringency filters improved precision at the expense of recall, with a slight corresponding improvement in F1-score across tools (Figure 4.3C, Supplementary Table S4.3).

4.3.4. *Short introns and introns that do not overlap exons are more reliably called*

We next compared the distributions of six intronic properties (length, position, exonic overlap, splice site motifs, U2- vs. U12-type spliceosomes, and uniformity of coverage by mapped reads) between the sets of true positive (TP), false positive (FP) and false negative (FN) RIs for each tool. Every tool except IRFinder-S had difficulty identifying shorter RIs (<600 bases) (Figures 4.4A, 4.4B). FPs tended to be longer than either TPs or FNs, and were distributed more centrally within a transcript compared to persistent introns (both TPs and FNs) across all tools (Figure 4.4A, Supplementary Figure S4.12). Further, there was a relative 3' bias for the small subset of FPs that were shared across all short-read tools, potentially reflective of the minimum coverage filters for most tools combined with sequencing coverage bias[278] (Supplementary Figure S4.13). As expected, the overwhelming majority of introns across all tools had canonical GT-AG splice motifs and splicing by the U2 spliceosome, while FNs showed increased frequencies of other motifs and spliceosome types relative to FPs and TPs (Supplementary Fig S4.14).

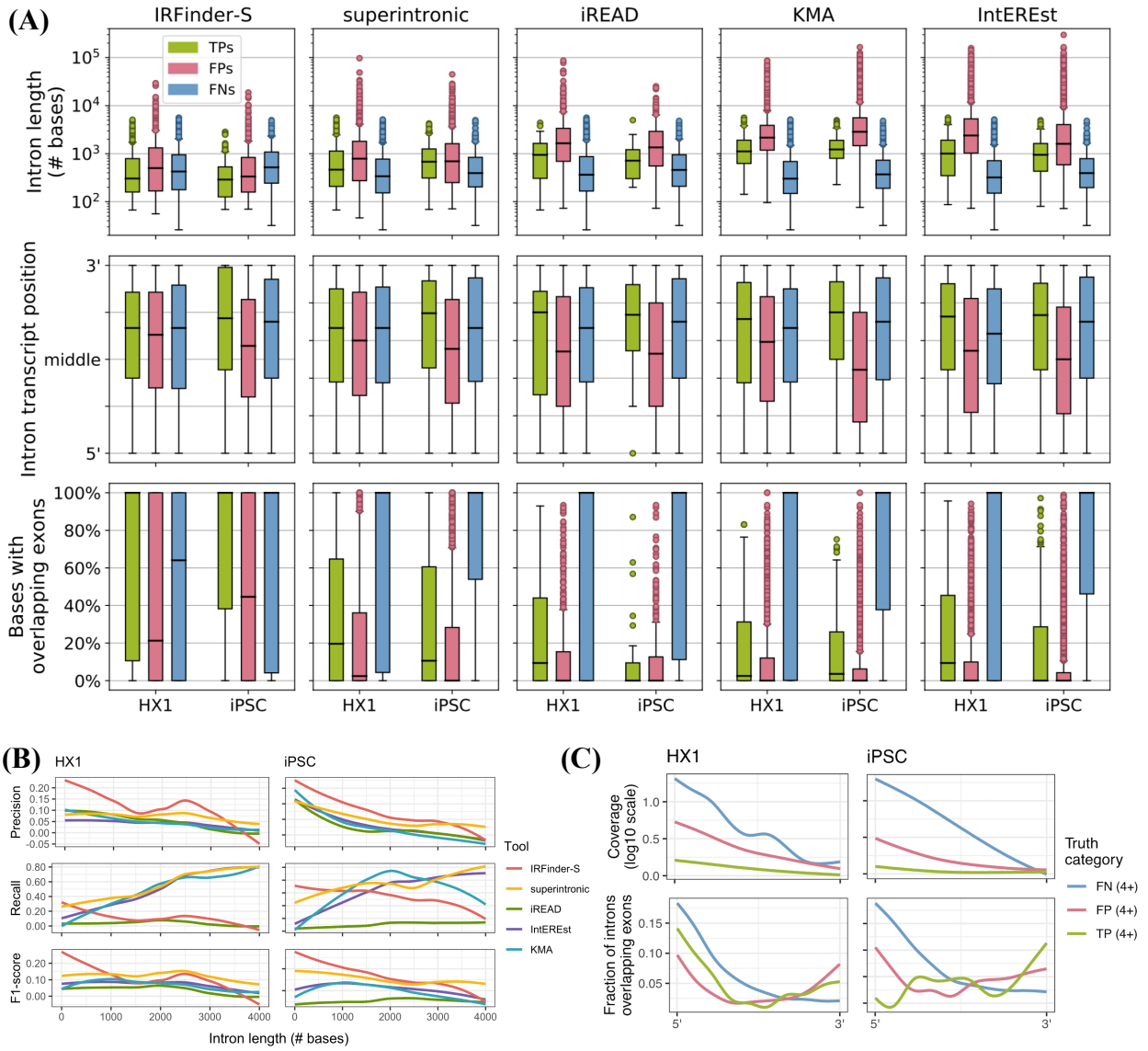


Figure 4.4: Properties of introns across detection truth categories. (A) Distributions of TP, FP, and FN RI properties across short-read detection tools. Panels contain boxplot distributions of intron length (top, log scale), relative position in transcript (middle), and % of intron bases with overlapping annotated exons (bottom) for each of five short-read tools (from left to right: IRFinder-S, superintronic, iREAD, KMA, IntEREst). Y-axes correspond to intron properties; each boxplot along the x-axis corresponds to the TP (green, left boxes), FP (pink, middle boxes), and FN (blue, right boxes) calls for HX1 (left) and iPSC (right). **(B)** Short-read tool performance vs. intron length. Panels depict precision (top), recall (middle), or F1-score (bottom) for five short-read tools applied to either HX1 (left) or iPSC (right). Tool performance (y-axis) for the subset of introns of a given length (x-axis) is colored by tool (red = IRFinder-S, yellow = superintronic, green = iREAD, purple = IntEREst, blue = KMA). (**C.**) Read coverage and exon overlap vs. position within an intron. LOESS-smoothed SR data (see Methods) show median log-scaled coverage (top row, y-axes) and fractions of introns with overlapping exons (bottom row, y-axes) as a function of position (x-axis, 5' → 3' on positive strand) for HX1 (left column) and iPSC (right column). Introns were grouped by truth category membership for at least 4/5 tools (colors, blue = FN, pink = FP, green = TP).

We also probed how much distributional uniformity of mapped read coverage across an intron (coverage "flatness" [215,217]) and incidence of overlapping exons differed among TPs, FPs, and FNs. Coverage of FPs and to a greater degree FNs was nonuniform, where coverage decreased roughly monotonically from 5' to 3' intron ends. Coverage of TPs was comparatively uniform, where coverage was in general substantially lower than for FPs and FNs (Figure 4.4C, top two plots). Closer to their 5' ends, FNs were distinguished by their tendency to overlap exons (Figure 4.4C, bottom two plots). Indeed, for superintronic, iREAD, KMA, and IntEREst, the majority of FNs appear to be accounted for by overlapping exons (Figure 4.4A). Overlapping exons may thus be a key obstacle to improving recall of many short-read RI detection tools.

4.3.5 *Persistent introns or called RIs occur in genes with experimentally validated IR*

Finally, we searched the literature and third-party resources for independent evidence of persistent introns appearing in the HX1 and iPSC samples studied here. We examined RI presence in 9 genes (5 in HX1 and 7 in iPSC) that have experimentally validated IR from a variety of cell types and tissues (Supplementary Table S4.4).[83,279–281] We found that intron retention across these 9 genes varied substantially by sample (no TP introns were observed in both HX1 and iPSC) (Figure 4.5). We also found significant variation between the set of RIs in these genes called by different short-read tools, with only a single TP intron in *IGSF8* identified across all tools for iPSC (Supplementary Figure S4.15). Interestingly, the genes *SRSF7* [94,282] and *AP1G2* [283] appear to be generally enriched for persistent introns, potentially consistent with post-transcriptional splicing.[61,268]

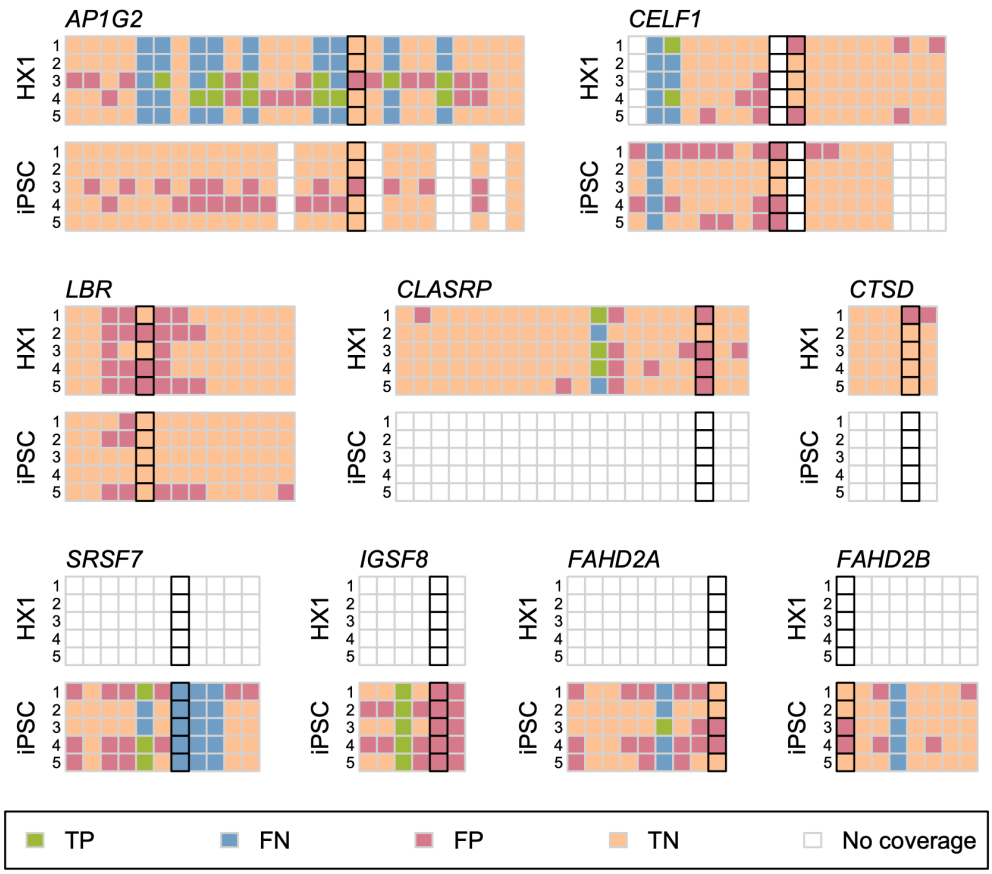


Figure 4.5: Short-read tool performance across nine genes with experimentally validated RIs. Comparison of short-read tool called RIs with introns detected in long-read data are shown as a pair of matrices for each of nine genes (*AP1G2*, *CELF1*, *LBR*, *CLASRP*, *CTSD*, *SRSF7*, *IGSF8*, *FAHD2A*, and *FAHD2B*). The rows in each matrix correspond to the results from each of five short-read tools (from top to bottom: 1: IntEREst, 2: iREAD, 3:IRFinder-S, 4: superintronic, 5: KMA) applied to either HX1 (top) or iPSC (bottom) data; columns correspond to all introns found across all annotated transcript isoforms of the indicated gene, ordered by left and then right genomic coordinates. Each cell in the matrix depicts the presence or absence of an intron in short-read and/or long-read data as a TP (green), FN (blue), FP (pink), and TN (peach) assessment; white boxes indicate introns found only in transcripts with <5 assigned long reads. Black outlines indicate the experimentally validated RI(s) in each gene.

4.4 Discussion

This is the first study to evaluate the quality of short-read RI detection using short- and long-read RNA-seq data from the same biological specimen. This study also establishes a novel metric capturing the persistence of an intron in a transcript as it is processed using deep long read RNA-seq, and it is the first to interrogate the potential effects of splicing progression during

transcript processing and spurious sources of intronic sequence. We find that short-read tools detect IR with poor recall and even worse precision, calling into question the completeness and validity of a large percentage of putatively retained introns called by commonly used methods. While our results indicate that it may be possible to improve precision slightly by applying expression filters to potential RIs, this appears to come at significant expense to recall.

This work raises fundamental questions regarding how results from short-read RI detection tools should be interpreted. We have taken IR to mean the persistence of an intron in a transcript after processing is complete, in alignment with the biological literature on IR. Short-read RI detection tools are commonly thought to identify such retained introns, with the assumption that poly(A) selection is sufficient to guarantee fully spliced and mature transcripts for sequencing; however, these tools are not inherently designed to distinguish intron retention from contaminating events such as partial transcript processing. This disconnect between how tool developers and tool users employ the same language may be responsible for false assertions in the published literature about which introns are retained. We note, for instance, that the prediction of putative neoepitopes arising from IR[154–157,273] requires confidence in the detection of stable, persistent IR with a high likelihood of translation and a low likelihood of undergoing NMD, none of which is assured by short-read RI detection tools.

Limitations of this work include the small number of biological specimens with matched short and deep long read RNA-seq available in the public domain, the lack of replicates of short-read RNA-seq data in this setting, and the limited depth of the long-read sequencing data. As a result, we were unable to study the patterns of IR across tissue type and other distinguishing sample characteristics. We confined attention to introns that occur in genes with high coverage in both short and long read data, and did not address either confidence in IR as a function of read

depth or systematic biases in gene coverage as a function of sequencing platform. While an improvement, our intron persistence metric only partially accounts for admixed splicing patterns from different cell types in a mixed-cell sample such as HX1. Like other RI detection studies,[75,81–83,150,154,157] our approach is explicitly linked to annotation (here, GENCODE v35) and therefore reports IR only relative to annotated transcripts, ignoring potential unannotated transcripts. We also did not explore the entanglement of biological and technical effects in the length of persistent introns: shorter introns are more likely to be retained,[75,120,124] but the length limit of PacBio Iso-Seq reads of up to 10 kilobases means that any molecules with longer persistent introns were not considered in this study. Furthermore, we calculated length-weighted median expression to harmonize short-read tool outputs to LR intron ranges (Supplementary Figure S4.16), and this stringent approach may have inflated false negative rates in regions returning high expression magnitudes and variances. Finally, we were only able to evaluate a small subset of the tools available for short read-based RI detection, as many of these tools harbor substantial software implementation and reproducibility challenges.

While there is evidence for cytoplasmic splicing, the phenomenon is rare in many tissues and cell types.[60,62,271,272] It may be worth exploring the extent to which sequencing only cytoplasmic RNAs focuses attention on fully processed RNA transcripts in future work.

4.5 Methods

Identification of paired short- and long-read data

Two advanced-search queries were performed on the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) on July 13, 2021, and all experiment accession numbers were collected from the query results by downloading the resulting “RunInfo” csv files. For both searches, the query terms included organism “human,” source “transcriptomic,” strategy “rna

seq," and access "public," with platform varying between the two searches: "pacbio smrt" for the long-read query and "illumina" for the short-read query. The RunInfo files were merged and projects with both Illumina and PacBio sequencing performed on the same National Center for Biotechnology Information (NCBI) Biosample (biological specimen) were identified. Due to relatively low sequencing depth of PacBio experiments, all projects with less than 20 PacBio sequencing runs were eliminated. PacBio experiments conducted on any PacBio platform earlier than RS II were also removed. Two remaining biosamples were chosen as data on which to test RI detection: 1) biosample SAMN07611993, an iPS cell line collected and processed by bioproject PRJNA475610, study SRP098984, with 1 short-read and 27 long-read runs,[275] and 2) biosample SAMN04251426 (HX1), a whole blood sample collected and processed by bioproject PRJNA301527, study SRP065930, with 1 short-read and 46 long-read runs.[274] (See the project repository at https://github.com/pdxgx/ri-tests for accession numbers.)

Long-read data collection, initial processing, and alignment

Raw Iso-Seq RS II data were downloaded from the SRA trace site (https://trace.ncbi.nlm.nih.gov/Traces/sra), via the "Original format" links under the "Data access" tab for each run. These comprised three .bax.h5 files for both samples, with an additional .bas.h5 and metadata file for each HX1 run. For both samples, individual runs were processed separately as follows, with differences in handling of the two samples as noted. Subreads were extracted to BAM files from the raw movie files using bax2bam (v0.0.8). Circular consensus sequences were extracted using ccs (v3.4.0) with --minPasses set to 1 and --minPredictedAccuracy set to 0.90. Barcodes were removed from ccs reads and samples were demultiplexed with lima (v2.2.0). For HX1, the input barcode fasta files were generated from the Clontech_5p and NEB_Clontech_3p lines from "Example 1" primer.fasta (https://github.com/PacificBiosciences/IsoSeq/blob/master/

isoseq-deduplication.md). For iPSC, forward and reverse barcode fasta files were downloaded from the study's GitHub page (https://github.com/EichlerLab/isoseq_pipeline/tree/master/data) and merged into a single fasta file per the lima input requirements. Since lima generates an output file for each 5'-3' primer set, these were merged using samtools merge (samtools and htslib v1.9). Demultiplexed reads were refined and poly-A tails removed using isoseq3 refine (isoseq v3.4.0) to generate full length non-concatemer (FLNC) reads. FLNC reads were extracted to fastq files using bedtools bamtofastq (bedtools v2.30.0), and aligned to GRCh38 with minimap2 (v2.20-r1061) using setting -ax splice:hq. Sequence download and processing scripts, and alignment statistics, are available at https://github.com/pdxgx/ri-tests.

Assignment of long reads to transcripts

The long-read alignment files were parsed as follows. GENCODE v.35[49] annotated transcripts' introns, strand, and start/end positions were extracted from the gencode v.35 GTF file. Then for each aligned long read, spliced-out introns, strand and start/end positions were extracted using pysam (v.0.16.0.1, using samtools v.1.10).[263,284] A set of possible annotated transcripts was generated, comprising transcripts for which the read's set of introns exactly matched the annotated transcripts' introns sets ("all introns"), or if no such transcripts were found, transcripts for which the read's introns were a subset of the transcripts' intron sets ("skipped splicing"). Then the best transcript match was chosen from the shortlist of potential matches as the transcript whose length most closely matched the read length. Some reads did not cover the full length of their best-matched transcripts, defined by the read alignment start and end position encompassing all introns in the annotated transcript ("full length"); in the case where not all intron coordinates were covered, these were labeled "partial" reads.

Intron persistence calculation

Intron persistence was calculated only for every transcript that was assigned as the best match for at least 5 reads. We calculated persistence for each intron within these transcripts as the information density of the intron d_i (i.e., the proportion of reads assigned to the transcript that cover intron i) multiplied by the mean of the product of three terms across all long reads assigned to that isoform:

1) The **retention**, or presence, $R_{r,i}$ of a given intron i is 1 if the read wholly contains i or 0 if it is absent/spliced out as annotated in read r.

2) The **spliced fraction** ($SF_{r,i}$) for a given intron i and read r is defined as

$$SF_{r,i} = \frac{|\{i' \epsilon I: R_{r,i'} = 0\}| + R_{r,i} - 1}{|I| - 1} \quad (4.3),$$

where I is the set of introns spanned by r and $R_{r,i}$ is defined above. This fraction of spliced introns in a read, with the target intron excluded, represents the splicing progression of the read. A mature RNA molecule should tend to have fewer unspliced introns present than an RNA from the same transcript at an earlier point in splicing progression.

3) The scaled **Hamming similarity** ($H_{r,i}$) for a given read r and intron i is defined as the average number of spliced or unspliced introns that match between the target read and other reads assigned to the transcript that have intron i spliced the same as in read r, scaled to the number of introns in the isoform:

$$H_{r,i} = \frac{1}{|\{r' \epsilon M^t: R_{r',i} = R_{r,i}\}|} \cdot \left(\sum_{\{r' \epsilon M^t: R_{r',i} = R_{r,i}\}} \frac{|\{i' \epsilon (I_{r'} \cap I_r): R_{r',i'} = R_{r,i'}\}|}{|I_{r'} \cap I_r|} \right) \quad (4.4),$$

where I_r is the set of introns spanned by r, $(I_{r'} \cap I_r)$ is the set of introns covered by both r and r', M^t is the set of reads assigned as best matches to the same transcript as r and span the target

intron i, and $R_{r,i'}$ is as defined above. Any partial reads that are assigned to the transcript as a best match but do not span the target intron are not included in this calculation, and the scaled Hamming similarity between two reads is only calculated for introns covered by both reads. This term accounts for the stochasticity of splicing initiation and progression, since a collection of reads would be more likely to have a dissimilar pattern of unspliced introns if the splicing process remained incomplete.

Persistence $P_{i,t}$ was calculated for each intron i in a given transcript isoform (t) as information density of the intron d_i times the mean of the product of the three terms above per Equation (4.1). Since short reads are not assignable to specific transcripts or isoforms, and certain introns fully or partially recur across multiple transcripts, we set the **intron persistence** (P_i) for a given intron i as the maximum $P_{i,t}$ found for that intron across all transcripts in which it occurs per Equation (4.2).

Alignment and BAM generation for short-read data

FASTQs were previously generated by other groups using either Illumina's NextSeq 500 (iPSC,[275] run id SRR6026510) or HiSeq 2000 (HX1,[274] run id SRR2911306), and files were obtained from the SRA using the "fastq-dump" command from the SRA Toolkit software (v2.10.8). A STAR (v2.7.6a)[200] index was generated based on the GRCh38 primary assembly genome FASTA (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/GRCh38. primary_assembly.genome.fa.gz) and GTF (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/gencode.v35.primary_assembly.annotation.gtf.gz) files from GENCODE version 35.[49] Reads were aligned with STAR to this index using the "--outSAMstrandField intronMotif" option. Primary alignments were retained for reads mapping

to multiple genome regions. SAM alignment files from STAR were converted to both sorted and unsorted BAM files using samtools (v1.3.1)[263] sort and view, respectively.

Additionally, for use with KMA,[212] bowtie2 (v2.3.4.3)[285] alignments were performed. Alignment statistics may be found in the project repository (https://github.com/pdxgx/ri-tests) and are summarized in Supplementary Figure S4.17. A FASTA file with intron sequences was generated based on the GRCh38 primary assembly genome FASTA and GTF files from GENCODE version 35 using the generate_introns.py script from the KMA package setting 0 bp for the extension flag. These intron sequences were combined with the GRCh38 transcript sequence FASTA file from GENCODE version 35 (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/gencode.v35.transcripts.fa.gz), and this combined FASTA was used to create a bowtie2 index. Reads were aligned to this index using bowtie2 according to specifications from KMA.[286] To quantify expression from the Bowtie 2 alignments, eXpress (v1.5.1)[287,288] was used.

Selection of target gene subset

Due to variable short- and long-read coverage across the genome, we selected a subset of genes to use for our test dataset to ensure adequate sequencing coverage for RI detection on both platforms. For the short-read data, we chose a coverage cutoff based on the requirements of the short-read RI detection tools used. The two tools with clear coverage requirements are iREAD, which requires coverage of 20 reads across an intron for RI detection, and superintronic, which requires 3 reads per region. Since these are short-reads (126 bases for iPSC and 90 for HX1) required over potentially long intronic regions, we chose a median gene-wide coverage (including both intronic and exonic regions) of 2 reads per base, ensuring either consistent coverage across the gene or high coverage in some areas. For the PacBio data, we selected 5 long

reads per gene, and at least 5 reads aligned to a single transcript of the gene, as giving enough information for comparing splicing progression and splicing patterns between reads. The target gene sets, 4,639 genes for iPSC and 4,369 for HX1, were chosen from the aligned data, naive to potential RI detection, and then for both short and long read data, the gene subset was applied as a filter after running metric calculations or RI detection by short read tools. Within these genes, only transcripts with at least 5 long reads were studied.

Intron feature annotation

For the set of target genes, transcripts with at least 5 long reads were selected for analysis. Features of each intron in these transcripts including intron lengths, splice motif sequences, relative transcript position, spliceosome category, and transcript feature overlap properties were extracted as follows. Length was calculated as the difference between the right and left genomic coordinates of the intron ends. Relative position within the transcript is an intron-count normalized fraction where 0 represents the transcript's 5' end and 1 represents the 3' end. Splice motifs were assigned to each intron by querying the GRCh38 reference genome with samtools faidx (samtools v.1.10) for the two coordinate positions at each end of the intron, and assigned to one of three canonical motif sequences (GT-AG, GC-AG, and AT-AC, and their reverse complements for - strand genes) or labeled as "other" for noncanonical motifs. Three feature overlap properties were studied: the total number of exons from other transcripts with any overlap of the intron region; the percent of intron bases with at least one overlapping exon from another transcript; and the maximum number of exons overlapping a single base in the intron. These were calculated by extracting all exon coordinates from the GENCODE v.35 annotation file, and using an interval tree to query each intron base position against the set of annotated exon coordinates. Spliceosome category was determined from recent U2 and U12 intron

annotations.[269] BED files of U2 and U12 introns for GRCh38 were downloaded from the Intron Annotation and Orthology Database (https://introndb.lerner.ccf.org/) on 1/25/22. Introns were labeled "U2" or "U12" if they only overlapped ranges from one of either spliceosome category, and remaining introns were labeled "other".

Selection of short-read RI detection algorithms and identification of likely RIs

We successfully downloaded and ran five IR-specific detection tools for short-read data on our remote server using the CentOS v7 operating system. To run superintronic, KMA, IntEREst, and iREAD, we used conda virtual environments (see https://github.com/pdxgx/ri-tests). We ran IRFinder-S from a fully self-contained Dropbox image per the tool's instructions (see below). IntEREst and superintronic are provided as R libraries which we ran from interactive R sessions, while iREAD, IRFinder-S, and KMA were run from command line, and a separate R package was used for RI detection for KMA. Outputs from all tools were read into R and harmonized to a single set of intron ranges after applying minimum coverage filters based on both short-read and long-read data. After running tools according to their provided documentation, we consulted literature and documentation on a tool-by-tool basis to devise starting filter criteria based on expression magnitude and other properties. We used these starting criteria to find the subset of most likely RIs, then we modified filter criteria to ensure filtered intron quantities were roughly one order of magnitude lower than unfiltered introns in both iPSC and HX1.

IR quantification with IntEREst

To run IntEREst (v1.6.2),[214] the referencePrepare function from the package was used to generate a reference from the GENCODE version 35 primary assembly GTF file.[49] This reference was used along with the sorted STAR BAM alignment from each sample to detect

intron retention with the interest function, considering all reads and not just those that map to junctions. We used the interest() function with the "IntRet" setting, which takes into account both intron-spanning and intron-exon junction reads and returns expression as a normalized fragments per kilobase of exon per million mapped fragments (FPKM). The filter FPKM >= 3, recommended for iREAD, left >90% of introns in both samples, so we increased the minimum filter to FPKM >= 45, and this retained (5038/32544 =) 15% of introns in HX1 and (6832/21820 =) 31% of introns in iPSC (Supplementary Figure S4.10).

IR quantification with keep me around (KMA)

To run KMA,[212] we used devtools to install a patched version of the software which resolves a bug unaddressed by the authors, available at https://github.com/adamtongji/kma. The read_express function was used to load expression quantification data output from eXpress, and the newIntronRetention function was used to detect intron retention. Returned intron expression was scaled as transcripts per million (TPM). We noted the recommended filters of unique counts >= 3 and TPM >= 1 left just 7.2% of introns in iPSC versus 19% in HX1, so we a less stringent filter of unique counts >= 10 for both samples, which left (6437/14155) 45% of introns in iPSC and (5089/20484) 25% of introns in HX1 (Supplementary Figure S4.10).

IR quantification with iREAD

To run iREAD (v0.8.5),[76] a custom intron BED file was made from the GENCODE version 35 primary assembly GTF file using GTFtools (v0.6.9).[289] The total number of mapped reads in each sorted STAR BAM alignment was determined using samtools, and used as input to the iREAD python script to detect intron retention. Intron expression was calculated as FPKM. To identify the most likely RIs, we applied previously published filter recommendations for entropy score (>= 0.9) and junction reads (>= 1). Since there were relatively few introns

remaining after applying published filters to the iPSC SR data (313/19316 = 1.6% versus 583/7748 = 7.5% in HX1), we applied lower filters for FPKM (>= 1 versus 3) and read fragments (>= 10 versus 20) (Supplementary Figure S4.10).

IR quantification with superintronic

To run superintronic (v0.99.4),[216] intronic and exonic regions were gathered from the GENCODE version 35 primary assembly GTF file[49] using the collect_parts function. The compute_coverage function was used to compute coverage scores for each sample from sorted STAR BAM alignments, and the join_parts function was used to convert these scores to per-feature coverage scores. Intron expression was returned as log2-scaled coverage, and we identified retained intron ranges as those overlapping long read-normalized ranges with LWM >= 3, per the expressed introns filter described in Lee et al.[216] (Supplementary Figure S4.10).

IR quantification with IRFinder-S

We ran IRFinder-S v.2.0-beta using the Docker image obtained from https://github.com/RitchieLabIGH/IRFinder. We prepared the IRFinder reference files using the Gencode v35 genome sequence reference and intron annotations.[49] Our analyses focused on the coverage and IRratio metrics, and the intron expression profile flags returned under warnings. Intron expression was returned as an IRratio, which is similar to percent spliced in (PSI), and we identified likely retained introns as having IRratio >= 0.5 without any flags per the methods in Lorenzi et al.[217] (Supplementary Figure S4.10).

Harmonization of intron retention metrics across algorithms and runs

Prior to analysis, we harmonized algorithm outputs on intron ranges returned by analysis of available long read runs. We harmonized intron expressions from short read RI detection tools

to intron ranges remaining after long reads were uniquely mapped to transcript isoforms. For each short-read RI detection tool, we calculated the region median intron expression value after weighting values on overlapping range lengths (a.k.a. length-weighted medians [LWM]). Calculation of LWMs is shown for an example intron in Supplementary Fig S4.16. Inter-rater agreement among the output from different short-read algorithms was assessed by Fleiss' kappa[290] using the R package irr v.0.84.1.[291]

Calculation of performances by intron length bins

We calculated called RI performance metrics across five short-read tools for a series of overlapping intron length bins. In total, 41 bins were calculated for each sample by sliding 300 bp-wide windows from 0 to 4300 bp lengths at 100 bp intervals. Plots were generated by computing LOESS smooths of the binned performance results.

Calculation of normalized binned coverages

We evaluated binned intron characteristics across intron truth metric categories for each sample. We assigned introns to truth categories if they were recurrent in that category for >=4 of 5 short-read tools (e.g. an intron that was recurrent TP for four tools in iPSC, etc.). We then calculated the $\log_{10}$ median short-read coverage for 1,000 evenly-spaced bins per intron for each truth category. We further calculated percent of introns overlapping an intron for each bin using the GENCODE v35 GTF. Plots were generated by computing the LOESS smooths of the binned results.

Comparison of detected RIs with circular RNA and validated RIs

We downloaded a database of human circular RNAs from circbase[276] (http://www.circbase.org/download/hsa_hg19_circRNA.txt), most recently updated in 2017. We

extracted all cRNAs labeled with the "intronic" flag in the annotation column and performed a liftover of genomic coordinates for these cRNAs from hg19 to GRCh38 using the University of California Santa Cruz (UCSC) Genome Browser liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). For each sample, we determined the percent of introns overlapping at least one cRNA for the 4+ consensus truth metric groups TP, FP, and FN (i.e. TP introns in >= 4 tools).

In order to test introns in this study against experimentally validated RIs, we identified wet lab studies in the literature that had first predicted, and then validated intron retention. We identified 4 such studies[83,279–281] that validated a total of 9 RIs in our sets of target genes as defined above (5 and 7 in HX1 and iPSC respectively) (Supplementary Table S4.4). (The above four plus an additional ten[60,79,88,292–298] studies experimentally validated RIs in an additional 6 and 9 genes that were found in our target gene sets for in HX1 and iPSC respectively, but without evidence of IR in our samples, and 41 and 36 genes, respectively, that did not pass our sample coverage thresholds for inclusion in this study.) The validated intron coordinates (Supplementary Table S4.4) were extracted either from the published intron number,[83,279,280] assuming a count from the gene's 5' to 3' end, or via BLAT queries of the target sequence[281]. In each sample and tool, we determined the truth status (TP, FP, TN, or FN) of all introns of all transcripts in the target gene for transcripts with >= 5 long reads. Adequate intron expression information was available in both samples for the genes *LBR*, *CELF1*, and *AB1G2*, but only one sample each for remaining genes.

Acknowledgements

We thank Jeremy Goecks, Joe Gray, and Kasper Hansen for their helpful feedback as this work was being prepared.

Chapter 5: Conclusion

5.1 *Summary*

In this work, I explored the detection and validation of cancer- and sample-specific splice events from short-read RNA-seq data. I found that even large normal datasets such as GTEx do not fully represent the breadth of normal human splicing. Some tissues are not sampled by GTEx and so their tissue-specific junctions are not included; similarly, GTEx contains only adult samples and so does not include junctions specific to early developmental stages. Bulk tissue samples, such as skin, may not contain certain low-prevalence cell types such as melanocytes, and even though some melanocyte-specific junctions are present in the GTEx data, some are still neglected. I also found that the detection of cancer-specific junctions and associated peptides is highly sensitive to filtering methods and parameter values chosen, and that MS validation of these may have a high rate of false positives. Finally, I found that retained intron detection in short-read RNA-seq data is highly inconsistent, and the tools tested are each likely to call many introns that are not truly retained while missing true RIs.

5.2 *Future directions*

Many interesting experiments suggest themselves as extensions of this work. First, other types of publicly available data can be leveraged to explore validity of detected junctions or RIs or deepen our biological understanding of these. Ribosome profiling (RNA-seq performed on transcripts being actively translated) could provide another layer of insight on whether or not target junctions are translated, and the possibility of using peptides arising from target junctions for immunotherapy or other applications. Deep long-read RNA-seq would provide rich transcriptional context for understanding the association of aberrant splice sites within transcripts and allow for more accurate generation of junction-associated peptide sequence.[299] Using, in

particular, longer reads than currently obtainable with PacBio Iso-Seq would provide a better scope of intron persistence across all introns. Paired nascent RNA-seq data would provide insight into co- and post-transcriptional splicing, and whether transcript- or gene-specific patterns exist, increasing understanding of potential transcript processing states in RI detection studies.

Furthermore, all RI detection currently is linked to annotation, but the development of annotation-free RI detection would open up new frontiers of exploration. This could potentially involve observing occurrence and recurrence of short sequences, i.e. kmerizing sequenced reads from large normal datasets as well the target sample, and identifying and reassembling kmers occurring only in the target sample.

Finally, many outstanding disease-specific questions remain, such as whether splicing changes in response to cancer treatment, which could be explored with pre- and post-treatment RNA-seq tumor data. As more RNA-seq data is eventually generated and becomes publicly available the scope of cancer-specific junction detection could be extended, with deeper exploration of cancer subtypes or stage-specific splicing. More data would also increase detection power for analyses such as observing the effect of splicing factor mutations on observed tumor junctions. Additional normal RNA-seq samples, such as the recently released GTEx v8 data (comprising >17,000 normal samples) will allow for increasingly detailed association of splicing with specific tissues and disease states.

5.3 *Concluding remarks*

The overall message of this dissertation is to urge caution when attempting to identify cancer- or sample-specific splicing from short-read RNA-seq data. In junction detection, many potentially cancer-specific junctions are simply rare in normal tissues or occur in normal tissues

not frequently sampled. This work also raises concerns about the confidence with which protein products of rare junctions can be detected with current methods and data. The high proportion of noncanonical splice motifs for splicing neoepitopes peptides validated by MS data indicates that either these are falsely called junctions, or that they represent modes of splicing significantly outside of current understanding; the former is a real possibility, especially with lowly-expressed junctions. Other mutations can also be misidentified as splicing by an aligner, such as the known long deletion of exons 2-7 in *EGFR* in GBM, identified as a junction in Chapter 2.[1] For intron retention detection, uncertainties lie in the processing state of any given transcript, and convolution of overlapping annotated transcript features. Viewing transcript processing as a continuum instead of a set of binary states (fully processed or unprocessed) adds valuable nuance to RI detection that is currently not used in other RI detection studies.[150,154,155] Altogether, a number of recent studies have confidently declared detection of neoepitopes arising from cancer-specific junctions[155,156,186,258] and intron retention[154–156] from short-read RNA-seq data, supported by MS validation of associated peptides. I propose that these should be viewed with some lack of certainty based on the results set forth here.

Ultimately, using long-read sequencing to examine transcript splicing context and reduce the probability of misalignment would add clarity and confidence to the detection of novel junctions and retained introns, and associated proteins.[299] Current limitations of long-read sequencing experiments such as read length, sequencing quality, and cost will continue to be mitigated as technology improves.[207]

References

1. David, J. K., Maden, S. K., Weeder, B. R., Thompson, R. F. & Nellore, A. Putatively cancer-specific exon-exon junctions are shared across patients and present in developmental and other non-cancer cells. *NAR Cancer* **2**, zcaa001 (2020).

2. David, J. K., Maden, S. K., Wood, M. A., Thompson, R. F. & Nellore, A. Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. (2022) doi:10.1101/2022.03.11.484016.

3. Wood, M. A. *et al.* neoepiscope improves neoepitope prediction with multivariant phasing. *Bioinformatics* **36**, 713–720 (2020).

4. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).

5. Weeder, B. R., Wood, M. A., Li, E., Nellore, A. & Thompson, R. F. pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab628.

6. Nielsen, M., Lundegaard, C., Lund, O. & Keşmir, C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57**, 33–41 (2005).

7. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).

8. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).

9. Shao, X. M. *et al.* High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res* **8**, 396–408 (2020).

10. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**, i237–i244 (2021).

11. Lu, T. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence* vol. 3 864–875 (2021).

12. Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M. & Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput. Biol.* **17**, e1008814 (2021).

13. Milighetti, M., Shawe-Taylor, J. & Chain, B. Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-

Major Histocompatibility Complexes. *Front. Physiol.* **12**, 730908 (2021).

14. Loeb, K. R. & Loeb, L. A. Significance of multiple mutations in cancer. *Carcinogenesis* **21**, 379–385 (2000).

15. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

16. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

17. Coventry, B. J. & Henneberg, M. The immune system and responses to cancer: Coordinated evolution. *F1000Res.* **4**, 552 (2021).

18. Yarchoan, M., Johnson, B. A., 3rd, Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* **17**, 569 (2017).

19. Nurieva, R., Wang, J. & Sahoo, A. T-cell tolerance in cancer. *Immunotherapy* **5**, 513–531 (2013).

20. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).

21. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).

22. Cogdill, A. P., Andrews, M. C. & Wargo, J. A. Hallmarks of response to immune checkpoint blockade. *British Journal of Cancer* vol. 117 1–7 (2017).

23. Kalinich, M. *et al.* Prediction of severe immune-related adverse events requiring hospital admission in patients on immune checkpoint inhibitors: study of a population level insurance claims database from the USA. *J Immunother Cancer* **9**, (2021).

24. Kartolo, A., Sattar, J., Sahai, V., Baetz, T. & Lakoff, J. M. Predictors of immunotherapy-induced immune-related adverse events. *Curr. Oncol.* **25**, e403–e410 (2018).

25. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).

26. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).

27. Pritchard, A. L. Targeting Neoantigens for Personalised Immunotherapy. *BioDrugs* **32**, 99–109 (2018).

28. Lennerz, V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 16013–16018 (2005).

29. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).

30. Etzioni, R. *et al.* The case for early detection. *Nat. Rev. Cancer* **3**, 243–252 (2003).

31. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* **20**, 71–88 (2019).

32. Lokshin, A., Bast, R. C. & Rodland, K. Circulating Cancer Biomarkers. *Cancers* **13**, (2021).

33. Oshi, M. *et al.* Urine as a Source of Liquid Biopsy for Cancer. *Cancers* **13**, (2021).

34. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404 (2016).

35. Valentini, D. *et al.* Identification of neoepitopes recognized by tumor-infiltrating lymphocytes (TILs) from patients with glioma. *Oncotarget* **9**, 19469–19480 (2018).

36. Lu, Y.-C. *et al.* Direct identification of neoantigen-specific TCRs from tumor specimens by high-throughput single-cell sequencing. *J Immunother Cancer* **9**, (2021).

37. De Mattos-Arruda, L. *et al.* Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. *Ann. Oncol.* **31**, 978–990 (2020).

38. Zhang, X., Qi, Y., Zhang, Q. & Liu, W. Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomed. Pharmacother.* **120**, 109542 (2019).

39. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576 (2014).

40. Zhou, Z. *et al.* TSNAD v2.0: A one-stop software solution for tumor-specific neoantigen detection. *Comput. Struct. Biotechnol. J.* **19**, 4510–4516 (2021).

41. Wood, M. A. *et al.* Population-level distribution and putative immunogenicity of cancer neoepitopes. *BMC Cancer* **18**, 414 (2018).

42. Richman, L. P., Vonderheide, R. H. & Rech, A. J. Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Syst* **9**, 375–382.e4 (2019).

43. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).

44. Yang, W. *et al.* Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat. Med.* **25**, 767–775 (2019).

45. Slansky, J. E. & Spellman, P. T. Alternative Splicing in Tumors - A Path to Immunogenicity? *N. Engl. J. Med.* **380**, 877–880 (2019).

46. Crick, F. Central Dogma of Molecular Biology. *Nature* vol. 227 561–563 (1970).

47. Bustamante, C., Cheng, W. & Mejia, Y. X. Revisiting the Central Dogma One Molecule at a Time. *Cell* vol. 145 160 (2011).

48. Shapiro, J. A. Revisiting the central dogma in the 21st century. *Ann. N. Y. Acad. Sci.* **1178**, 6–28 (2009).

49. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

50. Alberts, B. *et al.* Molecular Biology of the Cell. (2007) doi:10.1201/9780203833445.

51. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* vol. 74 3171–3175 (1977).

52. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* vol. 12 1–8 (1977).

53. Early, P. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* **20**, 313–319 (1980).

54. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).

55. Herzel, L. & Neugebauer, K. M. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods* **85**, 36–43 (2015).

56. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).

57. Alpert, T., Herzel, L. & Neugebauer, K. M. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip. Rev. RNA* **8**, (2017).

58. Reimer, K. Rapid and efficient co-transcriptional splicing enhances mammalian gene expression. (2020) doi:10.26226/morressier.5ebd45acffea6f735881ae83.

59. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* **16**, 1128–1133 (2009).

60. Denis, M. M. *et al.* Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets. *Cell* **122**, 379–391 (2005).

61. Bhatt, D. M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).

62. Buckley, P. T., Khaladkar, M., Kim, J. & Eberwine, J. Cytoplasmic intron retention, function, splicing, and the sentinel RNA hypothesis. *Wiley Interdiscip. Rev. RNA* **5**, 223–230 (2014).

63. Rosenfeld, M. G. *et al.* Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 1717–1721 (1982).

64. Blencowe, B. J. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in biochemical sciences* vol. 42 407–408 (2017).

65. Strobelt, H. *et al.* Vials: Visualizing Alternative Splicing of Genes. *IEEE Trans. Vis. Comput. Graph.* **22**, 399–408 (2016).

66. Doxakis, E. RNA binding proteins: a common denominator of neuronal function and dysfunction. *Neurosci. Bull.* **30**, 610–626 (2014).

67. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

68. Rodriguez, J. M., Pozo, F., di Domenico, T., Vazquez, J. & Tress, M. L. An analysis of tissue-specific alternative splicing at the protein level. *PLOS Computational Biology* vol. 16 e1008287 (2020).

69. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4**, 2771 (2013).

70. Lindholm, M. E. *et al.* The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J.* **28**, 4571–4581 (2014).

71. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).

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

73. Mazin, P. V., Khaitovich, P., Cardoso-Moreira, M. & Kaessmann, H. Alternative splicing during mammalian organ development. *Nat. Genet.* **53**, 925–934 (2021).

74. Wang, H. *et al.* Alteration of splicing factors' expression during liver disease progression: impact on hepatocellular carcinoma outcome. *Hepatology International* vol. 13 454–467 (2019).

75. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786 (2014).

76. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).

77. Galante, P. A. F., Sakabe, N. J., Kirschbaum-Slager, N. & de Souza, S. J. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10**, 757–765 (2004).

78. Lejeune, F. & Maquat, L. E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell Biol.* **17**, 309–315 (2005).

79. Buckley, P. T. *et al.* Cytoplasmic Intron Sequence-Retaining Transcripts Can Be Dendritically Targeted via ID Element Retrotransposons. *Neuron* vol. 69 877–884 (2011).

80. Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B. & Makeyev, E. V. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev.* **26**, 1209–1223 (2012).

81. Edwards, C. R. *et al.* A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127**, e24–e34 (2016).

82. Pimentel, H. *et al.* A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research* vol. 44 838–851 (2016).

83. Wong, J. J.-L. *et al.* Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. *Cell* vol. 154 583–595 (2013).

84. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* vol. 446 926–929 (2007).

85. Ge, Y. & Porse, B. T. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. *Bioessays* **36**, 236–243 (2014).

86. Naro, C. *et al.* An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev. Cell* **41**, 82–93.e4 (2017).

87. Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Hum. Genet.* **136**, 1043–1057 (2017).

88. Bell, T. J. *et al.* Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21152–21157 (2010).

89. Cho, V. *et al.* The RNA-binding protein hnRNPLL induces a T cell alternative splicing program delineated by differential intron retention in polyadenylated RNA. *Genome Biology* vol. 15 R26 (2014).

90. Memon, D. *et al.* Hypoxia-driven splicing into noncoding isoforms regulates the DNA damage response. *npj Genomic Medicine* vol. 1 (2016).

91. Mauger, O., Lemoine, F. & Scheiffele, P. Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. *Neuron* **92**, 1266–1278 (2016).

92. Ni, T. *et al.* Global intron retention mediated gene regulation during CD4 T cell activation. *Nucleic Acids Research* vol. 44 6817–6829 (2016).

93. de Lima Morais, D. A. & Harrison, P. M. Large-scale evidence for conservation of NMD candidature across mammals. *PLoS One* **5**, e11695 (2010).

94. Boutz, P. L., Bhutkar, A. & Sharp, P. A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* **29**, 63–80 (2015).

95. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).

96. Chen, H.-C. & Cheng, S.-C. Functional roles of protein splicing factors. *Biosci. Rep.* **32**, 345–359 (2012).

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

98. Beyer, A. L. & Osheim, Y. N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.* **2**, 754–765 (1988).

99. Aebi, M., Hornig, H., Padgett, R. A., Reiser, J. & Weissmann, C. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell* **47**, 555–565 (1986).

100. Herzel, L., Straube, K. & Neugebauer, K. M. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* **28**, 1008–1019 (2018).

101. Kessler, O., Jiang, Y. & Chasin, L. A. Order of intron removal during splicing of endogenous adenine phosphoribosyltransferase and dihydrofolate reductase pre-mRNA. *Molecular and Cellular Biology* vol. 13 6211–6222 (1993).

102. Schwarze, U., Starman, B. J. & Byers, P. H. Redefinition of Exon 7 in the COL1A1 Gene of Type I Collagen by an Intron 8 Splice-Donor–Site Mutation in a Form of Osteogenesis Imperfecta: Influence of Intron Splice Order on Outcome of Splice-Site Mutation. *The American Journal of Human Genetics* vol. 65 336–344 (1999).

103. de la Mata, M., Lafaille, C. & Kornblihtt, A. R. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA* **16**, 904–912 (2010).

104. Wan, Y. *et al.* Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell* **184**, 2878–2895.e20 (2021).

105. Hu, J., Boritz, E., Wylie, W. & Douek, D. C. Stochastic principles governing alternative splicing of RNA. *PLoS Comput. Biol.* **13**, e1005761 (2017).

106. Kim, S. W. *et al.* Widespread intra-dependencies in the removal of introns from human

transcripts. *Nucleic Acids Res.* **45**, 9503–9513 (2017).

107. Li, M. Calculating the most likely intron splicing orders in S. pombe, fruit fly, Arabidopsis thaliana, and humans. *BMC Bioinformatics* **21**, 478 (2020).

108. Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110 (2000).

109. Yu, Y. *et al.* Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell* **135**, 1224–1236 (2008).

110. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–1052 (2012).

111. Del Gatto-Konczak, F., Olive, M., Gesnel, M. C. & Breathnach, R. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. *Mol. Cell. Biol.* **19**, 251–260 (1999).

112. Bradley, T., Cook, M. E. & Blanchette, M. SR proteins control a complex network of RNA-processing events. *RNA* **21**, 75–92 (2015).

113. Zhu, J., Mayeda, A. & Krainer, A. R. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell* **8**, 1351–1361 (2001).

114. Long, J. C. & Caceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal* vol. 417 15–27 (2009).

115. Shen, H., Kan, J. L. C. & Green, M. R. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol. Cell* **13**, 367–376 (2004).

116. Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C. & Black, D. L. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.* **15**, 183–191 (2008).

117. Wei, G. *et al.* Position-specific intron retention is mediated by the histone methyltransferase SDG725. *BMC Biol.* **16**, 44 (2018).

118. Zhou, H.-L., Luo, G., Wise, J. A. & Lou, H. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* **42**, 701–713 (2014).

119. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428**, 2623–2635 (2016).

120. Schmitz, U. *et al.* Intron retention enhances gene regulatory complexity in vertebrates.

Genome Biol. **18**, 216 (2017).

121. Wong, J. J.-L. *et al.* Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nat. Commun.* **8**, 15134 (2017).

122. Kim, D. *et al.* Population-dependent Intron Retention and DNA Methylation in Breast Cancer. *Mol. Cancer Res.* **16**, 461–469 (2018).

123. Long, S. W., Ooi, J. Y. Y., Yau, P. M. & Jones, P. L. A brain-derived MeCP2 complex supports a role for MeCP2 in RNA processing. *Biosci. Rep.* **31**, 333–343 (2011).

124. Sakabe, N. J. & de Souza, S. J. Sequence features responsible for intron retention in human. *BMC Genomics* **8**, 59 (2007).

125. Monteuuis, G., Wong, J. J. L., Bailey, C. G., Schmitz, U. & Rasko, J. E. J. The changing paradigm of intron retention: regulation, ramifications and recipes. *Nucleic Acids Res.* **47**, 11497–11513 (2019).

126. Hogan, A. L. *et al.* Splicing factor proline and glutamine rich intron retention, reduced expression and aggregate formation are pathological features of amyotrophic lateral sclerosis. *Neuropathol. Appl. Neurobiol.* **47**, 990–1003 (2021).

127. Yao, J. *et al.* Prevalent intron retention fine-tunes gene expression and contributes to cellular senescence. *Aging Cell* **19**, e13276 (2020).

128. Lee, S. C.-W. & Abdel-Wahab, O. Therapeutic targeting of splicing in cancer. *Nature Medicine* vol. 22 976–986 (2016).

129. Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Hum. Mol. Genet.* **20**, 4422–4429 (2011).

130. Escobar-Hoyos, L., Knorr, K. & Abdel-Wahab, O. Aberrant RNA Splicing in Cancer. *Annu Rev Cancer Biol* **3**, 167–185 (2019).

131. Xiong, Y. *et al.* Profiles of alternative splicing in colorectal cancer and their clinical significance: A study based on large-scale sequencing data. *EBioMedicine* **36**, 183–195 (2018).

132. Wei, C. *et al.* Profiles of alternative splicing events in the diagnosis and prognosis of Gastric Cancer. *J. Cancer* **12**, 2982–2992 (2021).

133. Xu, L., Pan, J., Ding, Y. & Pan, H. Survival-Associated Alternative Splicing Events and Prognostic Signatures in Pancreatic Cancer. *Front. Genet.* **11**, 522383 (2020).

134. Zhu, J., Chen, Z. & Yong, L. Systematic profiling of alternative splicing signature reveals prognostic predictor for ovarian cancer. *Gynecol. Oncol.* **148**, 368–374 (2018).

135. Zhang, D., Zou, D., Deng, Y. & Yang, L. Systematic analysis of the relationship between

ovarian cancer prognosis and alternative splicing. *J. Ovarian Res.* **14**, 120 (2021).

136. Ouyang, Y. *et al.* Alternative splicing acts as an independent prognosticator in ovarian carcinoma. *Sci. Rep.* **11**, 10413 (2021).

137. Hu, Y.-X. *et al.* Systematic profiling of alternative splicing signature reveals prognostic predictor for cervical cancer. *J. Transl. Med.* **17**, 379 (2019).

138. Ouyang, D., Yang, P., Cai, J., Sun, S. & Wang, Z. Comprehensive analysis of prognostic alternative splicing signature in cervical cancer. *Cancer Cell Int.* **20**, 221 (2020).

139. Huang, R. *et al.* Identification of prognostic and metastasis-related alternative splicing signatures in hepatocellular carcinoma. *Biosci. Rep.* **40**, (2020).

140. Turpin, J. *et al.* The ErbB2ΔEx16 splice variant is a major oncogenic driver in breast cancer that promotes a pro-metastatic tumor microenvironment. *Oncogene* vol. 35 6053–6064 (2016).

141. Courtois, S. *et al.* DeltaN-p53, a natural isoform of p53 lacking the first transactivation domain, counteracts growth suppression by wild-type p53. *Oncogene* **21**, 6722–6728 (2002).

142. Marcel, V., Fernandes, K., Terrier, O., Lane, D. P. & Bourdon, J.-C. Modulation of p53β and p53γ expression by regulating the alternative splicing of TP53 gene modifies cellular response. *Cell Death Differ.* **21**, 1377–1387 (2014).

143. Brown, R. L. *et al.* CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *Journal of Clinical Investigation* vol. 121 1064–1074 (2011).

144. Sheen-Chen, S.-M., Chen, H.-S., Eng, H.-L. & Chen, W.-J. Circulating soluble Fas in patients with breast cancer. *World J. Surg.* **27**, 10–13 (2003).

145. Kondera-Anasz, Z., Mielczarek-Palacz, A. & Sikora, J. Soluble Fas receptor and soluble Fas ligand in the serum of women with uterine tumors. *Apoptosis* **10**, 1143–1149 (2005).

146. Liu, J. H. *et al.* Blockade of Fas-dependent apoptosis by soluble Fas in LGL leukemia. *Blood* **100**, 1449–1453 (2002).

147. Visakorpi, T. *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat. Genet.* **9**, 401–406 (1995).

148. Dehm, S. M., Schmidt, L. J., Heemers, H. V., Vessella, R. L. & Tindall, D. J. Splicing of a novel androgen receptor exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. *Cancer Res.* **68**, 5469–5477 (2008).

149. Marcias, G. *et al.* Identification of novel truncated androgen receptor (AR) mutants including unreported pre-mRNA splicing variants in the 22Rv1 hormone-refractory prostate

cancer (PCa) cell line. *Hum. Mutat.* **31**, 74–80 (2010).

150. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* **7**, 45 (2015).

151. Zhang, Q. *et al.* The global landscape of intron retentions in lung adenocarcinoma. *BMC Med. Genomics* **7**, 15 (2014).

152. Eswaran, J. *et al.* RNA sequencing of cancer reveals novel splicing alterations. *Sci. Rep.* **3**, 1689 (2013).

153. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).

154. Smart, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.* **36**, 1056–1058 (2018).

155. Trincado, J. L. *et al.* ISOTOPE: ISOform-guided prediction of epiTOPEs in cancer. *PLoS Comput. Biol.* **17**, e1009411 (2021).

156. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211–224.e6 (2018).

157. Dong, C. *et al.* Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene* **40**, 6130–6138 (2021).

158. Jayasinghe, R. G. *et al.* Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.* **23**, 270–281.e3 (2018).

159. Shiraishi, Y. *et al.* A comprehensive characterization of cis-acting splicing-associated variants in human cancer. doi:10.1101/162560.

160. Sun, X. *et al.* Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nat. Genet.* **37**, 407–412 (2005).

161. Diederichs, S. *et al.* The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol. Med.* **8**, 442–457 (2016).

162. Jung, H., Lee, K. S. & Choi, J. K. Comprehensive characterisation of intronic mis-splicing mutations in human cancers. *Oncogene* **40**, 1347–1361 (2021).

163. Anczuków, O. & Krainer, A. R. Splicing-factor alterations in cancers. *RNA* **22**, 1285–1301 (2016).

164. Quesada, V., Ramsay, A. J. & Lopez-Otin, C. Chronic lymphocytic leukemia with SF3B1 mutation. *N. Engl. J. Med.* **366**, 2530 (2012).

165. Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N.*

Engl. J. Med. **365**, 2497–2506 (2011).

166. Papaemmanuil, E. *et al.* SomaticSF3B1Mutation in myelodysplasia with ring sideroblasts. *N. Engl. J. Med.* **365**, 1384–1395 (2011).

167. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).

168. Graubert, T. A. *et al.* Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat. Genet.* **44**, 53–57 (2011).

169. Madan, V. *et al.* Distinct and convergent consequences of splice factor mutations in myelodysplastic syndromes. *Am. J. Hematol.* **95**, 133–143 (2020).

170. Darman, R. B. *et al.* Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep.* **13**, 1033–1045 (2015).

171. DeBoever, C. *et al.* Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. *PLoS Comput. Biol.* **11**, e1004105 (2015).

172. Furney, S. J. *et al.* SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov.* **3**, 1122–1129 (2013).

173. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).

174. Shuai, S. *et al.* The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).

175. Suzuki, H. *et al.* Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature* **574**, 707–711 (2019).

176. Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A. & Skotheim, R. I. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene* **35**, 2413–2427 (2016).

177. Lapuk, A. *et al.* Exon-level microarray analyses identify alternative splicing programs in breast cancer. *Mol. Cancer Res.* **8**, 961–974 (2010).

178. He, X. *et al.* Knockdown of polypyrimidine tract-binding protein suppresses ovarian tumor cell growth and invasiveness in vitro. *Oncogene* **26**, 4961–4968 (2007).

179. Takahashi, H. *et al.* Significance of polypyrimidine tract-binding protein 1 expression in colorectal cancer. *Mol. Cancer Ther.* **14**, 1705–1716 (2015).

180. Bielli, P. *et al.* The splicing factor PTBP1 promotes expression of oncogenic splice variants and predicts poor prognosis in patients with non-muscle-invasive bladder cancer. *Clin. Cancer Res.* **24**, 5422–5432 (2018).

181. Guo, J., Jia, J. & Jia, R. PTBP1 and PTBP2 impaired autoregulation of SRSF3 in cancer cells. *Sci. Rep.* **5**, 14548 (2015).

182. Song, Q. *et al.* CRKL regulates alternative splicing of cancer-related genes in cervical cancer samples and HeLa cell. *BMC Cancer* **19**, 499 (2019).

183. Srebrow, A. & Kornblihtt, A. R. The connection between splicing and cancer. *J. Cell Sci.* **119**, 2635–2641 (2006).

184. Fackenthal, J. D. & Godley, L. A. Aberrant RNA splicing and its functional consequences in cancer cells. *Dis. Model. Mech.* **1**, 37–42 (2008).

185. Coltri, P. P., Dos Santos, M. G. P. & da Silva, G. H. G. Splicing and cancer: Challenges and opportunities. *Wiley Interdiscip. Rev. RNA* **10**, e1527 (2019).

186. Pan, Y. *et al.* IRIS: Big data-informed discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing. *bioRxiv* (2019) doi:10.1101/843268.

187. Gershenson, S. M. Viruses as environmental mutagenic factors. *Mutat. Res.* **167**, 203–213 (1986).

188. National Research Council, Division on Earth and Life Studies, Commission on Life Sciences & Committee on the Biological Effects of Ionizing Radiation (BEIR V). *Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V*. (National Academies, 1990).

189. Ames, B. N. Identifying environmental chemicals causing mutations and cancer. *Science* **204**, 587–593 (1979).

190. Streisinger, G. *et al.* Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 77–84 (1966).

191. Schaaper, R. M. Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. *J. Biol. Chem.* **268**, 23762–23765 (1993).

192. Takeshima, H. & Ushijima, T. Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. *NPJ Precis Oncol* **3**, 7 (2019).

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

194. Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).

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

196. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

197. Collado-Torres, L., Nellore, A. & Jaffe, A. E. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Res.* **6**, 1558 (2017).

198. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).

199. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

200. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

201. Nellore, A. *et al.* Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* **33**, 4033–4040 (2017).

202. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* vol. 37 907–915 (2019).

203. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

204. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, lqab019 (2021).

205. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

206. Deschamps-Francoeur, G., Simoneau, J. & Scott, M. S. Handling multi-mapped reads in RNA-seq. *Comput. Struct. Biotechnol. J.* **18**, 1569–1576 (2020).

207. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

208. Broseus, L. & Ritchie, W. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput. Struct. Biotechnol. J.* **18**, 501–508 (2020).

209. Szabo, L. & Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat. Rev. Genet.* **17**, 679–692 (2016).

210. Xiang, Y. *et al.* Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. *J. Natl. Cancer Inst.* **110**, 379–389 (2018).

211. Yuan, F., Hankey, W., Wagner, E. J., Li, W. & Wang, Q. Alternative polyadenylation of mRNA and its role in cancer. *Genes & Diseases* (2019) doi:10.1016/j.gendis.2019.10.011.

212. Pimentel H, Conboy JG, Pachter L. Keep Me Around: Intron Retention Detection and

Analysis. (2015).

213. Bai, Y., Ji, S. & Wang, Y. IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-Seq data. *BMC Genomics* vol. 16 S9 (2015).

214. Oghabian, A., Greco, D. & Frilander, M. J. IntEREst: intron-exon retention estimator. *BMC Bioinformatics* **19**, 130 (2018).

215. Li, H.-D., Funk, C. C. & Price, N. D. iREAD: A Tool for Intron Retention Detection from RNA-seq Data. doi:10.1101/135624.

216. Lee, S. *et al.* Covering all your bases: incorporating intron signal from RNA-seq data. doi:10.1101/352823.

217. Lorenzi, C. *et al.* IRFinder-S: a comprehensive suite to discover and explore intron retention. *Genome Biol.* **22**, 307 (2021).

218. Tang, S. & Madhavan, S. neoantigenR: An annotation based pipeline for tumor neoantigen identification from sequencing data. doi:10.1101/171843.

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

220. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* vol. 19 (2018).

221. Wang, D. Y. *et al.* Fatal toxic effects associated with immune checkpoint inhibitors: A systematic review and meta-analysis. *JAMA Oncol.* **4**, 1721–1728 (2018).

222. Morgan, R. A. *et al.* Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immunother.* **36**, 133–151 (2013).

223. Linette, G. P. *et al.* Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863–871 (2013).

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

225. Fortier, M.-H. *et al.* The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205**, 595–610 (2008).

226. Sebestyén, E., Zawisza, M. & Eyras, E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.* **43**, 1345–1356 (2015).

227. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of

Alternative Splicing in Cancer. *Cell Rep.* **20**, 2215–2226 (2017).

228. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

229. Marcelino Meliso, F., Hubert, C. G., Favoretto Galante, P. A. & Penalva, L. O. RNA processing as an alternative route to attack glioblastoma. *Hum. Genet.* **136**, 1129–1141 (2017).

230. Robertson, A. G. *et al.* Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell* **33**, 151 (2018).

231. Bjørklund, S. S. *et al.* Widespread alternative exon usage in clinically distinct subtypes of Invasive Ductal Carcinoma. *Sci. Rep.* **7**, 5568 (2017).

232. Zong, Z. *et al.* Genome-Wide Profiling of Prognostic Alternative Splicing Signature in Colorectal Cancer. *Frontiers in Oncology* vol. 8 (2018).

233. Li, Y. *et al.* Prognostic alternative mRNA splicing signature in non-small cell lung cancer. *Cancer Lett.* **393**, 40–51 (2017).

234. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews Immunology* vol. 14 377–391 (2014).

235. Norris, A. D. & Calarco, J. A. Emerging Roles of Alternative Pre-mRNA Splicing Regulation in Neuronal Development and Function. *Frontiers in Neuroscience* vol. 6 (2012).

236. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805–817 (2016).

237. Bernstein, M. N., Doan, A. & Dewey, C. N. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics* **33**, 2914–2923 (2017).

238. Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics* **34**, 114–116 (2018).

239. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).

240. Broad Institute TCGA Genome Data Analysis Center. Firehose stddata run. (2016) doi:10.7908/C11G0KM9.

241. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

242. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across

all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

243. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).

244. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017).

245. Leinonen, R., Sugawara, H., Shumway, M. & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research* vol. 39 D19–D21 (2011).

246. Lin, E. W. *et al.* Comparative transcriptomes of adenocarcinomas and squamous cell carcinomas reveal molecular similarities that span classical anatomic boundaries. *PLoS Genet.* **13**, e1006938 (2017).

247. Naxerova, K. *et al.* Analysis of gene expression in a developmental context emphasizes distinct biological leitmotifs in human cancers. *Genome Biology* vol. 9 R108 (2008).

248. Borczuk, A. C. *et al.* Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am. J. Pathol.* **163**, 1949–1960 (2003).

249. Huang, S., Ernberg, I. & Kauffman, S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* **20**, 869–876 (2009).

250. Kaneko, O. *et al.* A binding domain on mesothelin for CA125/MUC16. *J. Biol. Chem.* **284**, 3739–3749 (2009).

251. Felder, M. *et al.* MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol. Cancer* **13**, 129 (2014).

252. Li, X.-P. *et al.* Overexpression of ribosomal L1 domain containing 1 is associated with an aggressive phenotype and a poor prognosis in patients with prostate cancer. *Oncology Letters* vol. 11 2839–2844 (2016).

253. Godinho, M., Meijer, D., Setyono-Han, B., Dorssers, L. C. J. & van Agthoven, T. Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells. *J. Cell. Physiol.* **226**, 1741–1749 (2011).

254. Wood, M. A., Weeder, B. R., David, J. K., Nellore, A. & Thompson, R. F. Burden of tumor mutations, neoepitopes, and other variants are cautionary predictors of cancer immunotherapy response and overall survival. (2019) doi:10.1101/665026.

255. El Marabti, E. & Younis, I. The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer. *Front Mol Biosci* **5**, 80 (2018).

256. Brinkman, B. M. N. Splice variants as cancer biomarkers. *Clin. Biochem.* **37**, 584–594

(2004).

257. Bessa, C., Matos, P., Jordan, P. & Gonçalves, V. Alternative Splicing: Expanding the Landscape of Cancer Biomarkers and Therapeutics. *Int. J. Mol. Sci.* **21**, (2020).

258. Rivero-Hinojosa, S. *et al.* Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat. Commun.* **12**, 6689 (2021).

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

260. Lin, A., Plubell, D. L., Keich, U. & Noble, W. S. Accurately Assigning Peptides to Spectra When Only a Subset of Peptides Are Relevant. *J. Proteome Res.* **20**, 4153–4164 (2021).

261. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000).

262. Ellis, M. J. *et al.* Connecting Genomic Alterations to Cancer Biology with Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* vol. 3 1108–1112 (2013).

263. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

264. Hulstaert, N. *et al.* ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J. Proteome Res.* **19**, 537–542 (2020).

265. McIlwain, S. *et al.* Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **13**, 4488–4491 (2014).

266. May, D. H., Tamura, K. & Noble, W. S. Detecting Modifications in Proteomics Experiments with Param-Medic. *J. Proteome Res.* **18**, 1902–1906 (2019).

267. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **604**, 55–71 (2010).

268. Girard, C. *et al.* Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat. Commun.* **3**, 994 (2012).

269. Moyer, D. C., Larue, G. E., Hershberger, C. E., Roy, S. W. & Padgett, R. A. Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Res.* **48**, 7066–7078 (2020).

270. Zhang, G., Taneja, K. L., Singer, R. H. & Green, M. R. Localization of pre-mRNA splicing in mammalian nuclei. *Nature* **372**, 809–812 (1994).

271. König, H., Matter, N., Bader, R., Thiele, W. & Müller, F. Splicing segregation: the minor

spliceosome acts outside the nucleus and controls cell proliferation. *Cell* **131**, 718–729 (2007).

272. Uemura, A., Oku, M., Mori, K. & Yoshida, H. Unconventional splicing of XBP1 mRNA occurs in the cytoplasm during the mammalian unfolded protein response. *J. Cell Sci.* **122**, 2877–2886 (2009).

273. Dong, C. *et al.* Intron-retention neoantigen load predicts favorable prognosis in pancreatic cancer. *JCO Clin. Cancer Inform.* **6**, e2100124 (2022).

274. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).

275. Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**, (2018).

276. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).

277. Website. https://www.pacb.com/wp-content/uploads/2015/09/User-Bulletin-Guidelines-for-Preparing-cDNA-Libraries-for-Isoform-Sequencing-Iso-Seq.pdf.

278. Khrameeva, E. E. & Gelfand, M. S. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics* **13 Suppl 6**, S4 (2012).

279. Jeong, J.-E., Seol, B., Kim, H.-S., Kim, J.-Y. & Cho, Y.-S. Exploration of Alternative Splicing Events in Mesenchymal Stem Cells from Human Induced Pluripotent Stem Cells. *Genes* **12**, (2021).

280. Lejeune, F., Cavaloc, Y. & Stevenin, J. Alternative splicing of intron 3 of the Serine/arginine-rich protein 9G8 gene. *J. Biol. Chem.* **276**, 7850–7858 (2001).

281. Li, H.-D. *et al.* Integrative functional genomic analysis of intron retention in human and mouse brain with Alzheimer's disease. *Alzheimers. Dement.* **17**, 984–1004 (2021).

282. Königs, V. *et al.* SRSF7 maintains its homeostasis through the expression of Split-ORFs and nuclear body assembly. *Nat. Struct. Mol. Biol.* **27**, 260–273 (2020).

283. Heinicke, L. A. *et al.* The RNA binding protein RBM38 (RNPC1) regulates splicing during late erythroid differentiation. *PLoS One* **8**, e78031 (2013).

284. Bonfield, J. K. *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, (2021).

285. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

286. pachterlab. pachterlab/kma. https://github.com/pachterlab/kma.

287. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).

288. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

289. Li, H.-D. GTFtools: a Python package for analyzing various modes of gene models. doi:10.1101/263517.

290. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971).

291. irr: Various Coefficients of Interrater Reliability and Agreement. *Comprehensive R Archive Network (CRAN)* https://CRAN.R-project.org/package=irr.

292. Dumbović, G. *et al.* Nuclear compartmentalization of TERT mRNA and TUG1 lncRNA is driven by intron retention. *Nat. Commun.* **12**, 3308 (2021).

293. Inoue, D. *et al.* Minor intron retention drives clonal hematopoietic disorders and diverse cancer predisposition. *Nat. Genet.* **53**, 707–718 (2021).

294. Bell, T. J. *et al.* Cytoplasmic BK(Ca) channel intron-containing mRNAs contribute to the intrinsic excitability of hippocampal neurons. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1901–1906 (2008).

295. Chen, Z., Gore, B. B., Long, H., Ma, L. & Tessier-Lavigne, M. Alternative splicing of the Robo3 axon guidance receptor governs the midline switch from attraction to repulsion. *Neuron* **58**, 325–332 (2008).

296. Zhong, X., Liu, J. R., Kyle, J. W., Hanck, D. A. & Agnew, W. S. A profile of alternative RNA splicing and transcript variation of CACNA1H, a human T-channel gene candidate for idiopathic generalized epilepsies. *Hum. Mol. Genet.* **15**, 1497–1512 (2006).

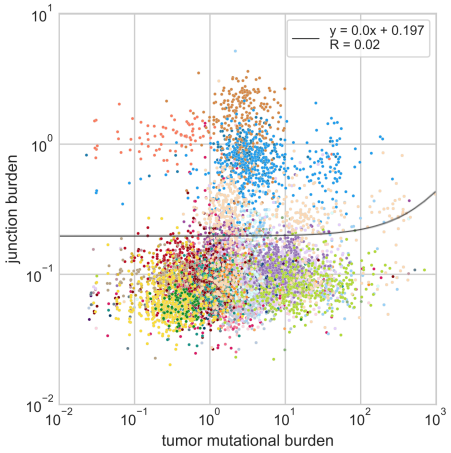
297. Mansilla, A. *et al.* Developmental regulation of a proinsulin messenger RNA generated by intron retention. *EMBO Rep.* **6**, 1182–1187 (2005).

298. Forrest, S. T., Barringhaus, K. G., Perlegas, D., Hammarskjold, M.-L. & McNamara, C. A. Intron retention generates a novel Id3 isoform that inhibits vascular lesion formation. *J. Biol. Chem.* **279**, 32897–32903 (2004).

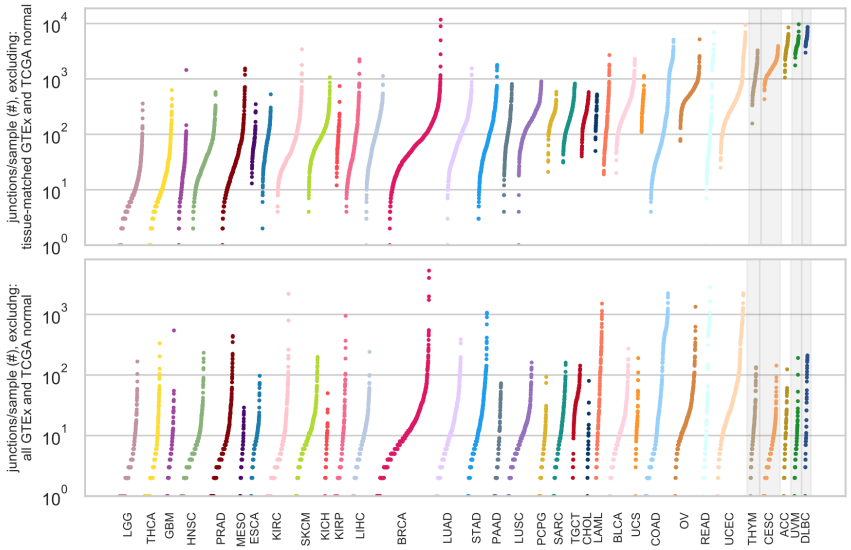
299. Miller, R. M. *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.* **23**, 69 (2022).

Appendix A: Supplementary Figures and Tables

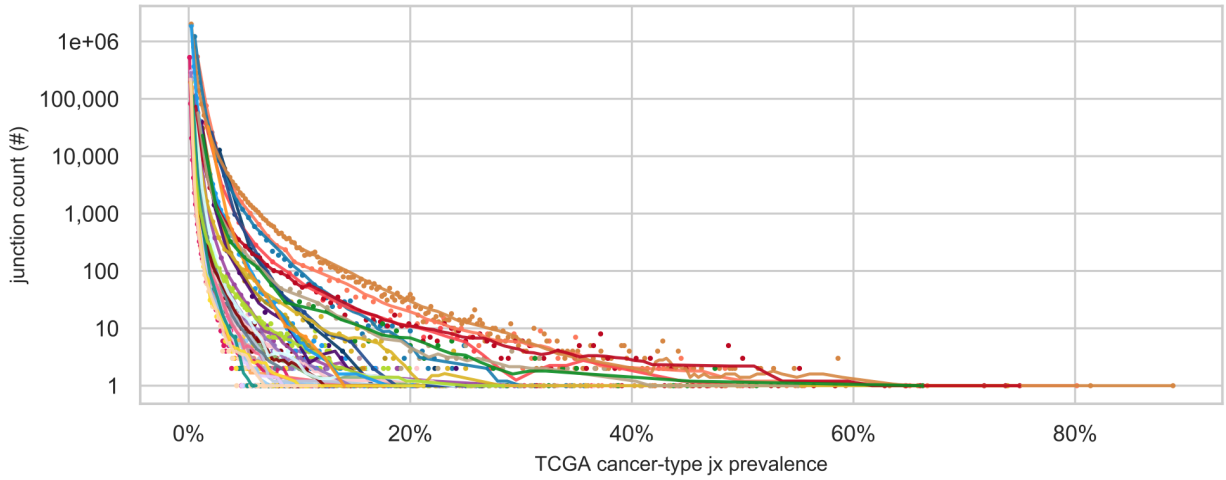
(A)



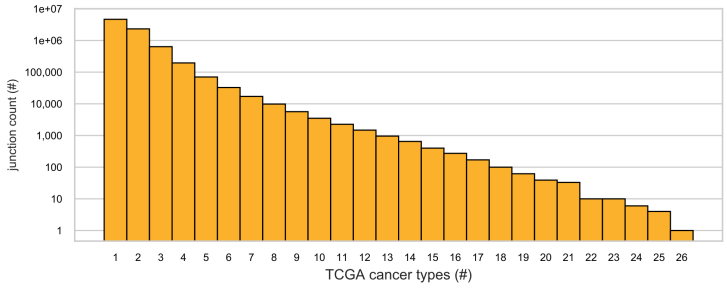
(B)



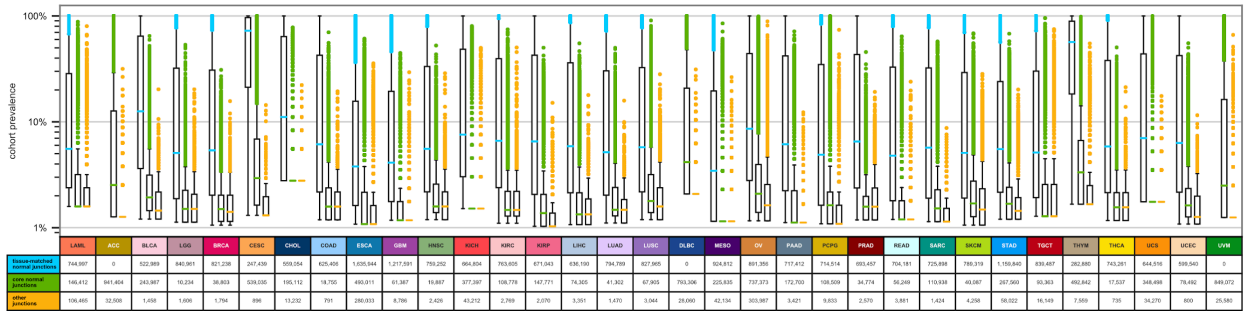
(C)



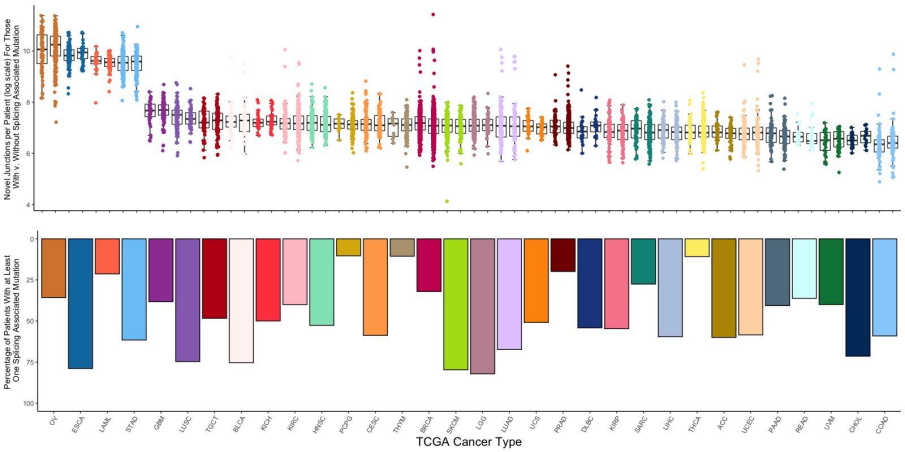
(D)


junction count (#)
1e+07
1e+06
100,000
10,000
1,000
100
10
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
TCGA cancer types (#)

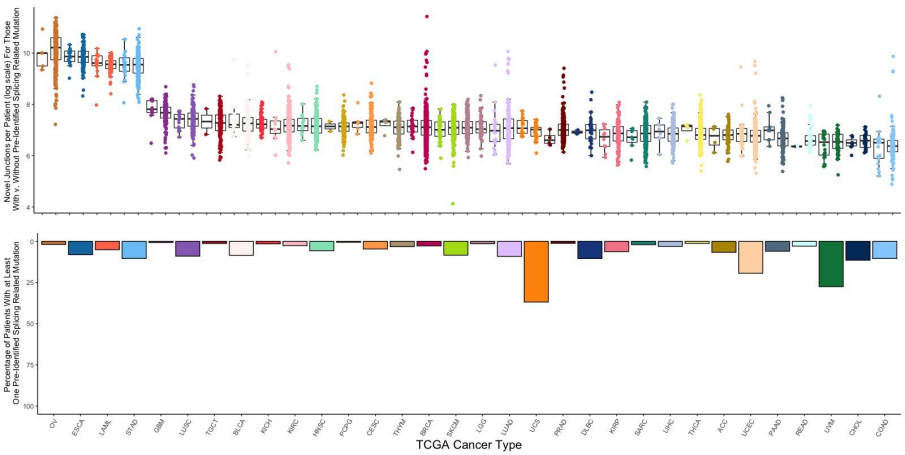
(E)


cohort prevalence
100%
10%
1%
LAML ACC BLCA LGG BRCA CESC CHOL COAD ESCA GBM HNSC KICH KIRC KIRP LIHC LUAD LUSC DLBC MESO OV PAAD PCPG PRAD READ SARC SKCM STAD TGCT THYM THCA UCS UCEC UVM
tissue-matched normal junctions 744,997 0 522,989 840,961 821,236 247,439 559,554 625,406 1,635,644 1,217,591 759,252 664,804 763,605 671,043 636,190 794,769 827,965 0 924,812 891,356 717,412 714,514 693,457 704,181 725,898 789,319 1,159,840 839,487 282,880 743,261 644,516 599,540 0
core normal junctions 140,412 941,404 243,987 10,234 38,803 539,035 195,112 18,700 490,011 61,367 19,887 377,397 108,778 147,771 74,305 41,302 67,905 793,306 225,835 737,373 172,700 108,509 34,774 56,249 110,938 40,087 267,560 93,363 492,842 17,537 348,498 78,492 849,072
other junctions 109,465 32,508 1,458 1,606 1,794 896 13,232 791 280,933 8,786 2,426 43,212 2,769 2,070 3,351 1,470 3,044 28,060 42,134 303,987 3,421 9,833 2,570 3,881 1,424 4,258 58,022 16,149 7,559 735 34,270 800 25,580

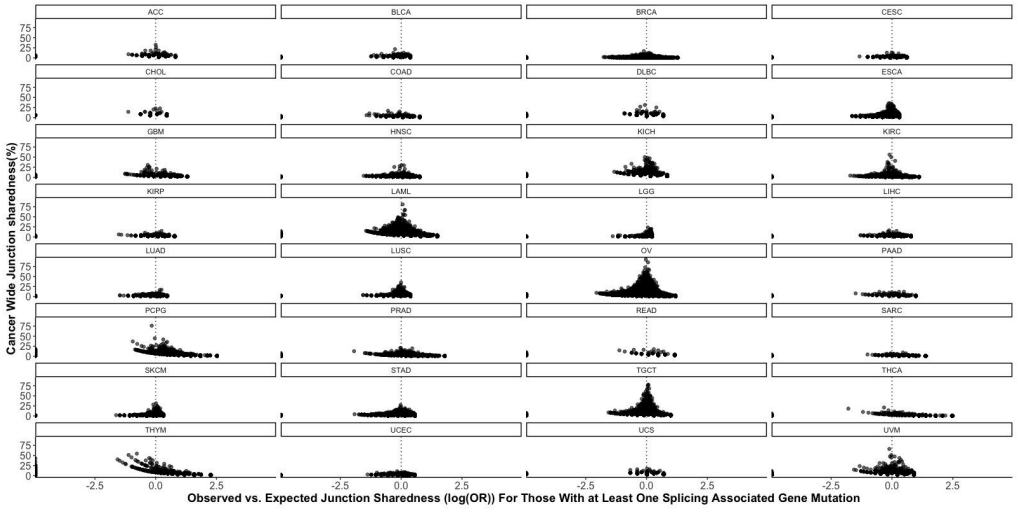
(F)


Novel Junctions per Patient (log scale) For Those With >1 Novel Splicing Associated Mutation
Percentage of Patients With at Least One Splicing Associated Mutation
OV ESCA LAML STAD GBM LUSC TGCT BLCA KICH KIRC HNSC PCPG CESC THYM BRCA SKCM LGG LUAD UCS PRAD DLBC KIRP SARC LIHC THCA ACC UCEC PAAD READ UVM CHOL COAD
TCGA Cancer Type

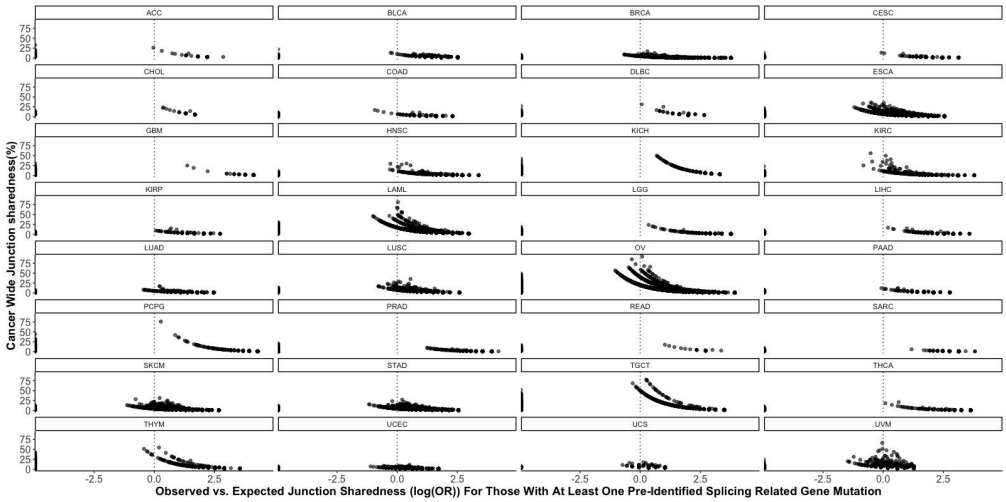
(G)


Novel Junctions per Patient (log scale) For Those With >1 Novel Predicted Splicing Altered Mutation
Percentage of Patients With at Least One Predicted Splicing Altered Mutation
OV ESCA LAML STAD GBM LUSC TGCT BLCA KICH KIRC HNSC PCPG CESC THYM BRCA SKCM LGG LUAD UCS PRAD DLBC KIRP SARC LIHC THCA ACC UCEC PAAD READ UVM CHOL COAD
TCGA Cancer Type

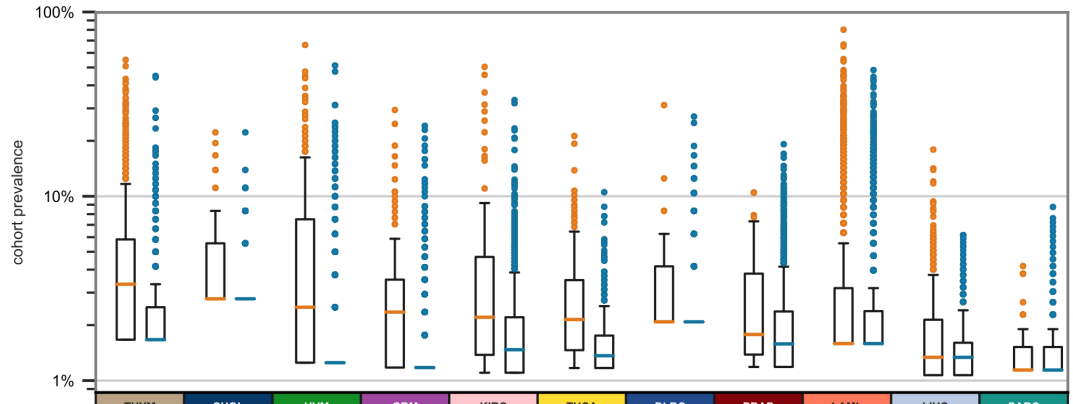
(H)



(I)



(J)



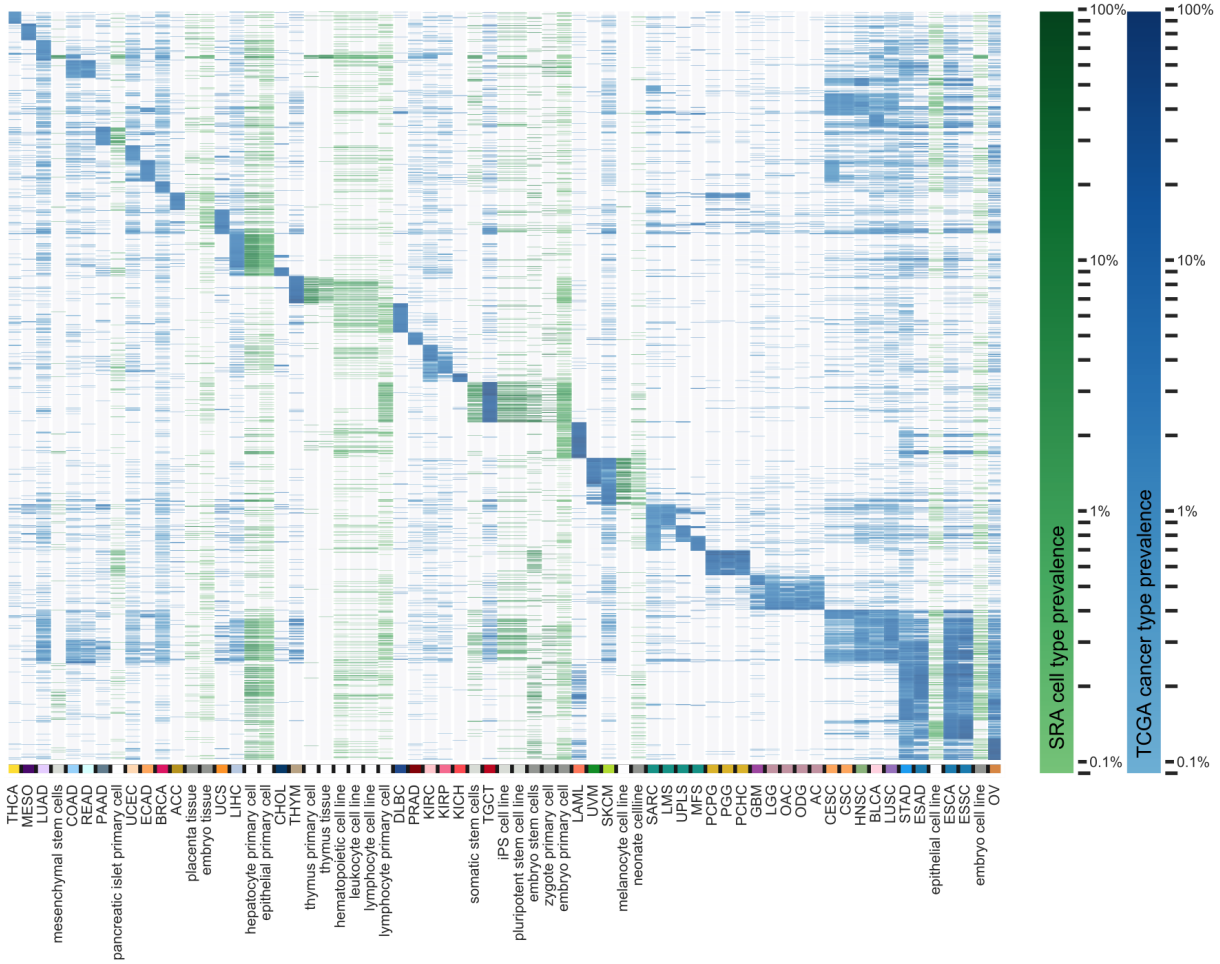
	THYM	CHOL	UVM	GBM	KIRC	THCA	DLBC	PRAD	LAML	LIHC	SARC
SRA expts	15	7	5	20	110	52	2	51	725	115	84
SRA-shared neojunctions	1,516	66	301	112	79	191	117	76	11,572	1,315	61
TCGA-only neojunctions	6,043	25,751	50,594	8,674	2,690	544	56,169	2,494	94,893	2,036	1,363
p value	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.002643	<0.00001	<0.00001	0.04955

Supplementary Figure S2.1: Distribution and prevalences of TCGA cancer sample junctions.
(A) Log-scale scatterplot showing no relationship between junction burden (number of junctions per sample scaled by mapped read count) and tumor mutational burden. Each point represents one TCGA sample, colored based on cancer type (as in Figure 1B and Supplementary Figures S1B and S1C).
(B) Log-scale sorted strip plots representing the number of high-support junctions per sample for each of 33 TCGA cancer types where each point is a single TCGA tumor sample and the width of each strip is proportional to the size of the cancer type cohort.[156] High support requires junction coverage within a sample to be equivalent to at least 5 out of 100 million 100-base pair reads. The upper panel counts junctions not found in GENCODE annotation or in tissue-matched normal GTEx or TCGA samples (Supplementary Table S1); the lower panel counts junctions not found in GENCODE annotation or in any core normal samples, differing from Figure 1B in that here, GENCODE-annotated junctions are also removed and the coverage filter is applied. The gray bars highlight TCGA cancer types with no or few tissue-matched normal samples (Supplementary Table S1); note that there are orders of magnitude fewer GENCODE-annotated junctions than junctions found in tissue-matched normal samples, partially explaining the high values for THYM, CESC, UVM, and DLBC in the upper panel.
(C) Log-scale scatter plot showing, for 33 TCGA cancer types, the number of junctions shared within each cancer-type cohort at each prevalence level, counting only junctions not found in any core normal samples. TCGA cancer type colors are as specified in Figures 1B, 2A, and S1B. Of interest with significant intra-cohort junction sharedness are, among others, ovarian carcinoma (tan), leukemia (pink), testicular germ cell tumors (red), and uveal melanoma (dark green).
(D) Log-scale histogram showing inter-cancer sharedness of junctions not found in core normal samples. Most junctions occur in only one cancer type, but many are shared between 2 or more.
(E) Log-scale box plots representing, for all TCGA cancer types individually, the prevalences within each cancer-type cohort of junctions occurring in at least 1% of cancer-type samples, separated into prevalences for (blue, left) junctions found in GTEx or TCGA tissue-matched normal samples (Supplementary Table S1); (green, center) junctions not found in tissue-matched normals but found in other core normal samples; and (yellow, right) junctions found in no core normal samples. Any junction found in multiple cancer types is represented by multiple data points, one for each cancer type in which it is found. Figure 1C condenses all data from this figure into one pan-cancer set.
(F) Cancer specific splicing junctions in patients with and without splicing associated gene mutations: log count of junctions not found in any core normal samples for each patient are plotted (top) across each cancer type in TCGA. Within each cancer, boxplots represent either patients with mutations in UniProt annotated splicing-related genes (left) or patients without any related mutations (right). Overall prevalence of relevant mutations in splicing-related for each cancer type are plotted below.
(G) Presents data in the same manner as E, but with comparison between patients with mutations only in genes described in the TCGA splicing paper[156] vs all other patients.
(H) Analysis of junction sharedness for patients within mutational cohorts: for each cancer, junctions not found in core normal samples are plotted based on sharedness across all patients in the cancer cohort (y-axis) compared to deviation from expected sharedness (estimated odds ratio (log scale) based on Fisher's exact test) for only patients with mutations in UniProt annotated splicing-related genes (x-axis).
(I) Presents data in the same manner as G, but with deviation from expected sharedness calculated only for patients with mutations in genes described in the TCGA splicing paper[156]. We found that no specific junctions show significantly enriched sharedness in patients carrying relevant mutations (Fisher's exact test FDR > 0.05 for all in both A and B), however there is a consistent shift towards higher than expected

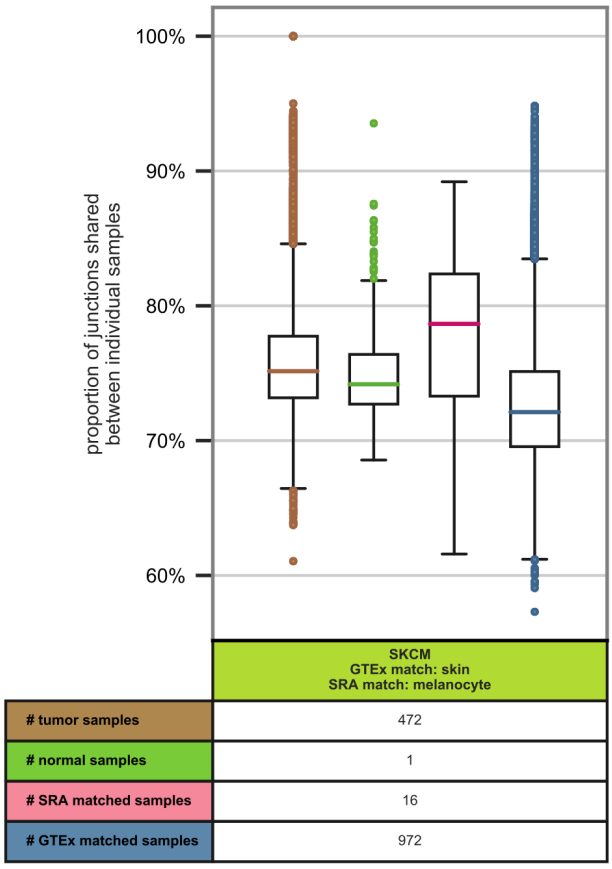
sharedness across the majority of cancers for patients carrying at least one of the mutations defined by the TCGA splicing paper.[156]

(J) Comparison of TCGA-cohort prevalence of junctions occurring vs. not occurring in SRA cancer samples: log-scale box plots representing, for selected TCGA cancer types, the prevalences within each cancer-type cohort of junctions occurring in at least 1% of cancer-type samples, separated into prevalences for junctions (orange, left) found or (blue, right) not found in type-matched cancer sample(s) from the SRA. Selected TCGA cancer types are those for which cancer-matched SRA sample junctions are available from Snaptron [238] and at least 50 TCGA cancer junctions not found in core normal samples are present in the cancer-type matched SRA samples. Most junctions are TCGA-specific, but junctions that are also found in a type-matched SRA cancer cohort have on average higher TCGA-cohort prevalences.

(A)

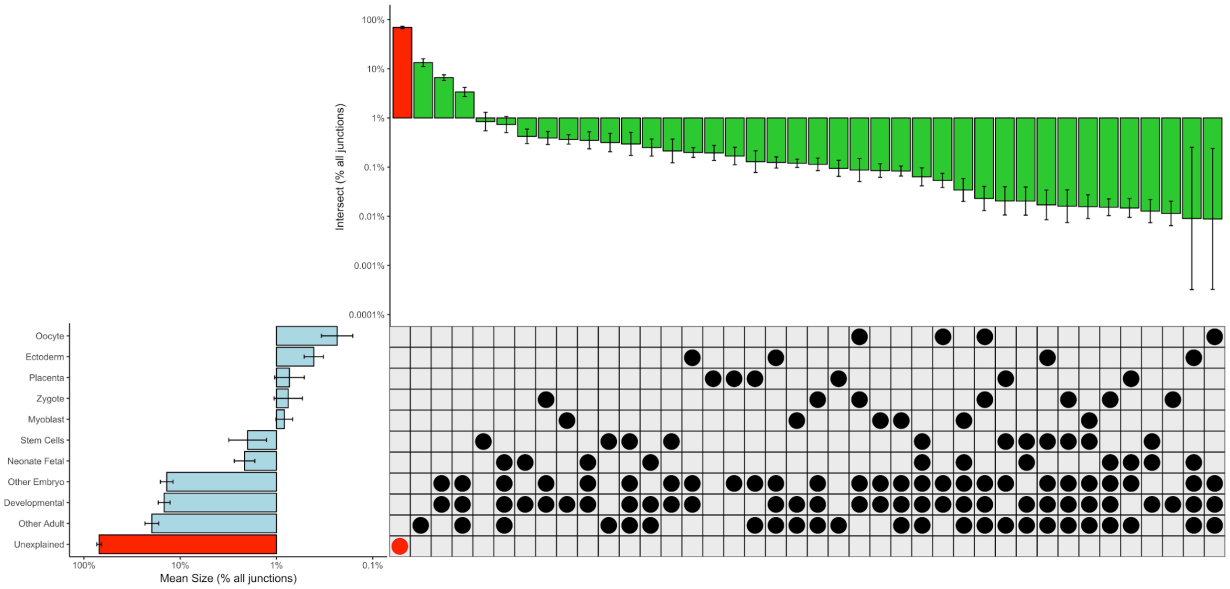

THCA
MESO
LUAD
mesenchymal stem cells
COAD
READ
PAAD
pancreatic islet primary cell
UCEC
ECAD
BRCA
ACC
placenta tissue
embryo tissue
UCS
LIHC
hepatocyte primary cell
epithelial primary cell
CHOL
THYM
thymus primary cell
thymus tissue
hematopoietic cell line
leukocyte cell line
lymphocyte cell line
lymphocyte primary cell
DLBC
PRAD
KIRC
KIRP
KICH
somatic stem cells
TGCT
iPS cell line
pluripotent stem cell line
embryo stem cells
zygote primary cell
embryo primary cell
LAML
UVM
SKCM
melanocyte cell line
neonate cellline
SARC
LMS
UPLS
MFS
PCPG
PGG
PCHC
GBM
LGG
OAC
ODG
AC
CESC
CSC
HNSC
BLCA
LUSC
STAD
ESAD
epithelial cell line
ESCA
ESSC
embryo cell line
OV
SRA cell type prevalence
100%
10%
1%
0.1%
TCGA cancer type prevalence
100%
10%
1%
0.1%

(B)

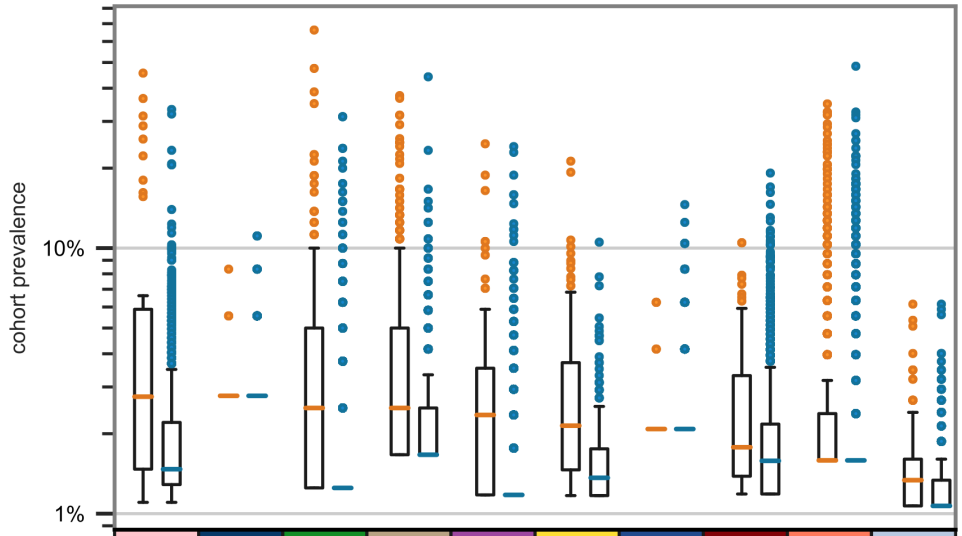


Supplementary Figure S2.2: Similarity of TCGA junctions and non-cancer SRA junctions.
(A) Clustering by cohort prevalence of junctions not found in core normal samples: heatmap showing shared junction prevalences across all TCGA cancer types and associated histological subtypes with at least 20 samples. The clustered junctions are the 200 most prevalent junctions of each cancer type or subtype that are at least 1% prevalent in that subtype and are not found in any core normal samples but are found in at least one of the 22 non-cancer tissue and cell type SRA samples. Each heatmap row represents a junction's prevalence in each of the TCGA and SRA sample-type cohorts. The colorbar beneath the plot shows SRA tissue and cell types colored according to their assigned categories (Supplementary Table S3), where white represents adult normal samples, light gray represents stem cell samples, and darker gray represents developmental samples.
(B) Samplewise comparison of junctions from TCGA melanoma samples and select normal samples: boxplots showing the percent of junctions shared for every pairwise combination of TCGA melanoma tumor samples with (brown) TCGA melanoma tumor samples, (grass green) the single TCGA melanoma paired normal sample, (pink) SRA normal melanocyte samples (see Supplementary Tables S1 and S3), and (blue) GTEx normal skin samples. The percent of junctions shared between two samples is given by $\%\,shared\ =\ (set\ A\ \&\ set\ B)\ /\ min(len(set\ A), len(set\ B))$, where a set comprises all junctions identified in the single cancer or normal sample. TCGA melanoma cancer samples have on average a greater similarity of junctions to SRA normal melanocyte samples than to GTEx or TCGA bulk skin normal samples.

(A)

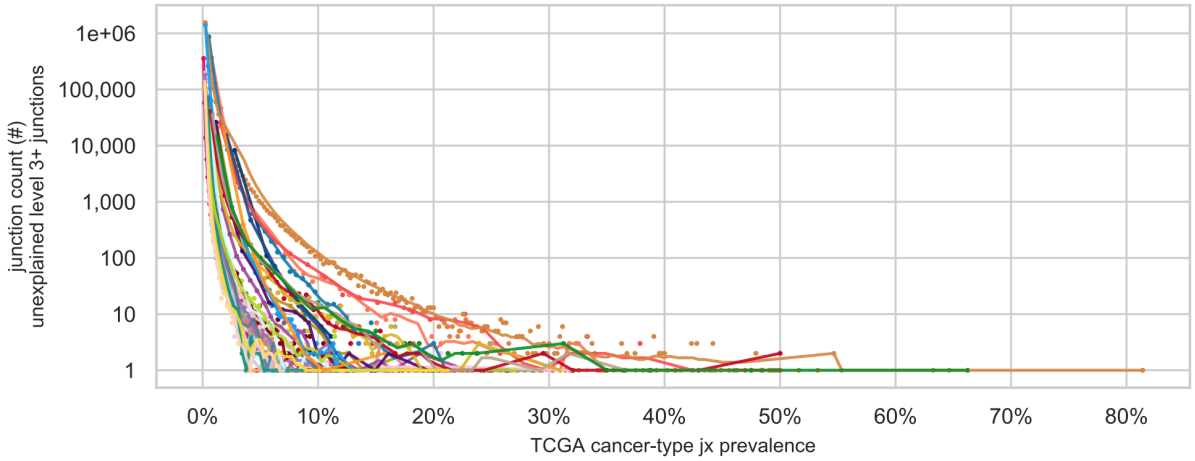


(B)

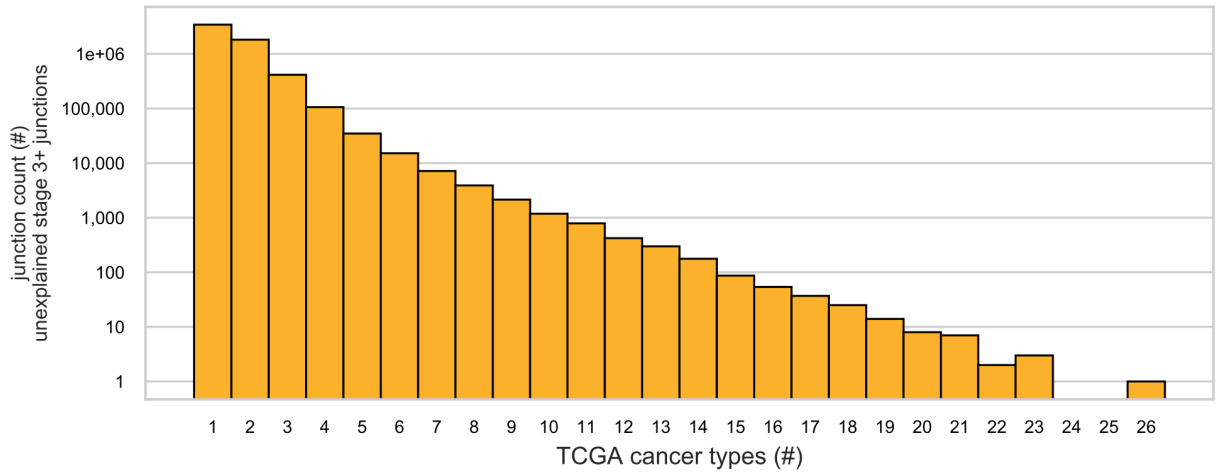


	KIRC	CHOL	UVM	THYM	GBM	THCA	DLBC	PRAD	LAML	LIHC
SRA expts	110	7	5	15	20	52	2	51	725	115
SRA-shared neojunctions	45	31	181	896	65	139	46	62	3,261	315
TCGA-only neojunctions	1,717	8,368	16,103	3,425	6,214	372	16,155	2,024	55,971	1,013
p value	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.007791	<0.00001	2.426e-05

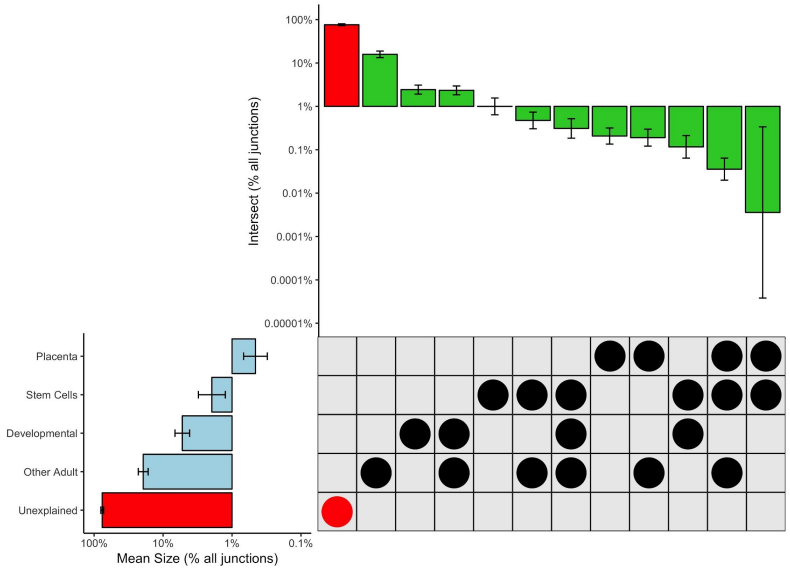
(C)


junction count (#)
unexplained level 3+ junctions
1e+06
100,000
10,000
1,000
100
10
1
0%
10%
20%
30%
40%
50%
60%
70%
80%
TCGA cancer-type jx prevalence

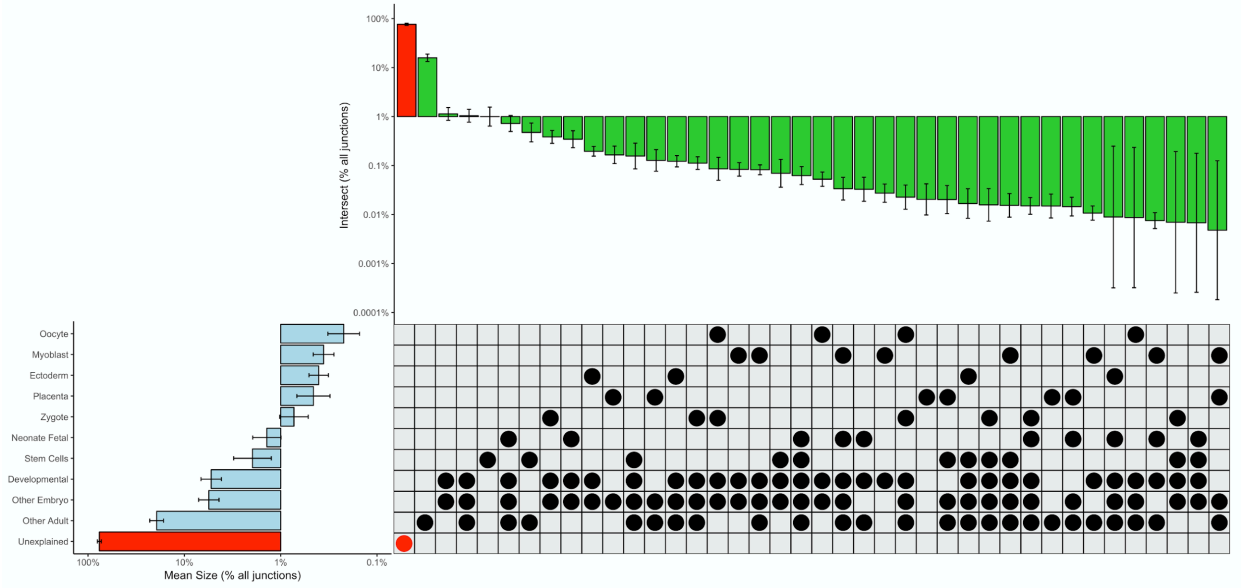
(D)


junction count (#)
unexplained stage 3+ junctions
1e+06
100,000
10,000
1,000
100
10
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
TCGA cancer types (#)

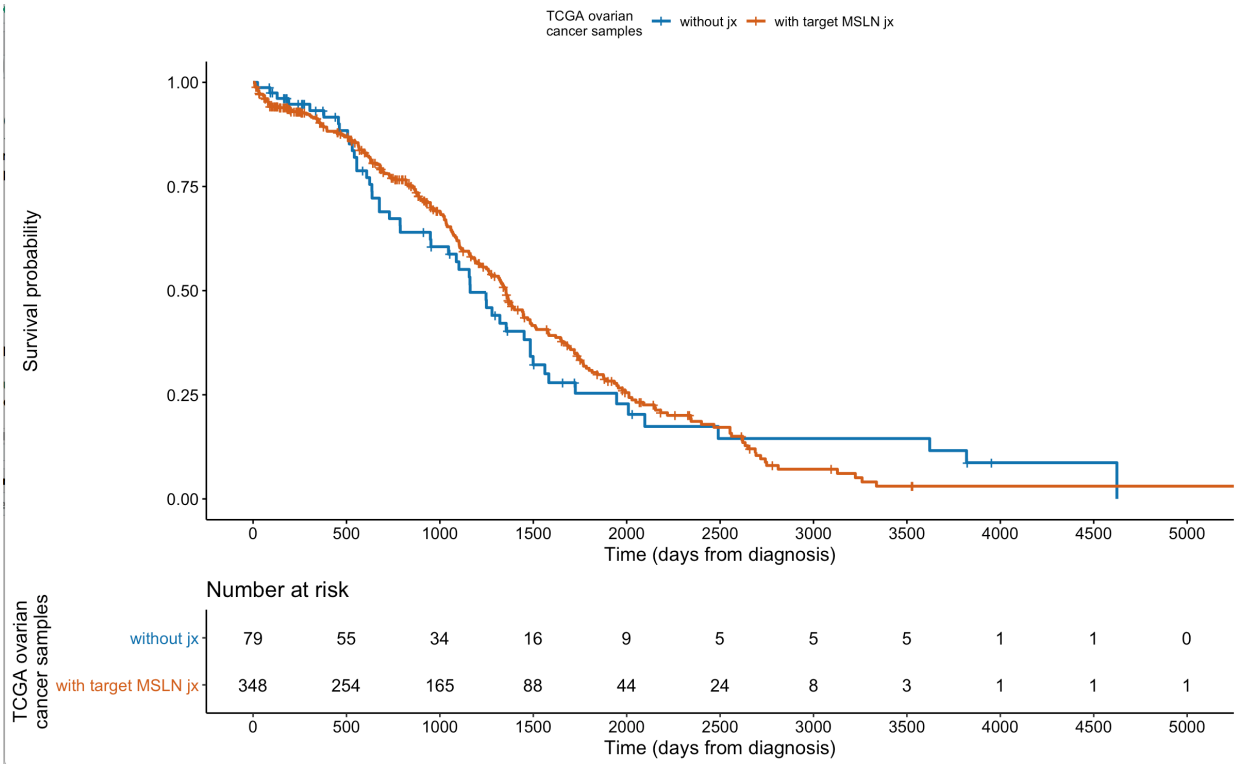
(E)


Intersect (% all junctions)
100%
10%
1%
0.1%
0.01%
0.001%
0.0001%
0.00001%
Placenta
Stem Cells
Developmental
Other Adult
Unexplained
100%
10%
1%
0.1%
Mean Size (% all junctions)

(F)


Intersect (% all junctions)
100%
10%
1%
0.1%
0.01%
0.001%
0.0001%
Oocyte
Myoblast
Ectoderm
Placenta
Zygote
Neonate Fetal
Stem Cells
Developmental
Other Embryo
Other Adult
Unexplained
100%
10%
1%
0.1%
Mean Size (% all junctions)

(G)


TCGA ovarian cancer samples
without jx
with target MSLN jx
Survival probability
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Time (days from diagnosis)
Number at risk
TCGA ovarian cancer samples
without jx 79 55 34 16 9 5 5 5 1 1 0
with target MSLN jx 348 254 165 88 44 24 8 3 1 1 1
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Time (days from diagnosis)

Supplementary Figure S2.3: Distribution of junctions not found in core normal samples and unexplained junctions.

(A) Expanded junction set assignments in normal tissue and cell type categories from the Sequence Read Archive, across cancers: upset-style plot with bar plots showing junction abundances across major sets and subsets (left) and set overlaps (top) across 33 cancers (error bars). Shown junctions are absent from all core normals. Unexplained junctions (red highlights) comprise junctions not present in any set categories studied (see also Figure 3A). The developmental set comprises human development-related junctions not present in the category placenta. Scale is log10 of percent of junctions not found in core normals, calculated for each cancer.

(B) Analysis of inter- and intra-cancer sharedness of stage-3+ "unexplained" junctions: log-scale box plots as in Supplementary Figure S1J but including only stage-3+ unexplained junctions not found in core normal samples or selected SRA normal adult, developmental, or stem cell samples (Table 1). Plots are presented for TCGA cancer types for which cancer-matched SRA sample junctions are available from Snaptron [238] and at least 30 unexplained junctions are present in the cancer-type matched SRA samples. Prevalences are given within each cancer-type cohort of junctions occurring in at least 0.5% of cancer-type samples, separated into prevalences for junctions (orange, left) found or (blue, right) not found in type-matched cancer sample(s) from the SRA. For all cancer types except DLBC, most junctions are TCGA-specific, but junctions that are also found in a type-matched SRA cancer cohort have on average higher TCGA-cohort prevalences.

(C) Log-scale scatter plot showing, for 33 TCGA cancer types, the number of level-3+ unexplained junctions shared within each cancer-type cohort at each prevalence level as in Figure Supplementary S1C. TCGA cancer type colors are as specified in Figures 1B, 2A, and S1B. Again, among others, ovarian carcinoma (tan), leukemia (pink), testicular germ cell tumors (red), and uveal melanoma (dark green) have significant intra-cohort junction sharedness.

(D) Log-scale histogram showing inter-cancer sharedness of stage-3+ unexplained junctions as in Supplementary Figure S1D. Again, most junctions occur in only one cancer type, but many are shared between 2 or more.

(E) Upset-style plot with bar plots showing junction abundances across major sets (left) and set overlaps (top) across 33 cancers (error bars); similar to Figure 3A, but presence in 2 samples across the SRA sample-type category is required for inclusion in a set. Shown junctions are absent from all core normals. Unexplained junctions (red highlights) comprise junctions not present in any set categories studied. The developmental set comprises human development-related junctions not present in the category placenta. Scale is log10 of percent of junctions not found in core normals, calculated for each cancer.

(F) Upset-style plot with bar plots showing junction abundances across major sets and subsets (left) and set overlaps (top) across 33 cancers (error bars); similar to Figure 3A, but presence in 2 samples across the SRA sample-type category is required for inclusion in a set. Shown junctions are absent from all core normals. Unexplained junctions (red highlights) comprise junctions not present in any set categories studied. The developmental set comprises human development-related junctions not present in the category placenta. Scale is log10 of percent of junctions not found in core normals, calculated for each cancer.

(G) Survival curve for patients with or without the target high-prevalence antisense MSLN junction of interest (chr16;766903;768491;-), censored at last registered follow-up appointment.

Supplementary Table S2.1: Sources and counts of tumor and tissue-matched normal samples.

TCGA cancer type (Abbreviation: # tumor samples)	# of TCGA paired normal samples	GTEx matched tissue(s) (# of normal samples)	# TCGA tumor sample junctions not in tissue-matched normal (avg #/sample)	# unique TCGA cancer-type junctions not in tissue-matched normals	Additional matched normals used: SRA cell type of origin (# of samples)	Histological subtypes (Abbreviation: # of tumor samples)	SRA matched cancer: # of SRA samples
Acute Myeloid Leukemia (LAML: 126)	0	Blood (595)	2,836,278 (22,510/sample)	2,264,159	NA	NA	Acute myeloid leukemia: 725
Bladder Urothelial Carcinoma (BLCA: 414)	19	Bladder (11)	5,627,108 (13,592/sample)	2,227,034	NA	NA	NA
Brain Lower Grade Glioma (LGG: 532)	48	Brain (1409)	1,608,421 (3,023/sample)	1,266,167	NA	Astrocytoma (AC: 196); Oligoastrocytoma (OAC: 135); Oligodendroglioma (ODG: 200)	NA
Breast Invasive Carcinoma (BRCA: 1134)	112	Breast (218)	6,287,963 (5,545/sample)	3,480,255	NA	NA	NA
Cervical Squamous Cell Carcinoma & Endocervical Adenocarcinoma (CESC: 306)	3	Cervix Uteri (11)	14,086,434 (46,034 /sample)	2,263,326	NA	Cervical Adenosquamous (CASC: 6); Cervical squamous cell carcinoma (CSC: 253); Endocervical adenocarcinoma (ECAD: 47)	Cervical Carcinoma: 23
Colon Adenocarcinoma (COAD: 505)	41	Colon (376)	1,754,895 (3,475/sample)	1,268,454	NA	NA	NA
Esophageal Carcinoma (ESCA: 185)	13	Esophagus (788)	6,588,463 (35,613/sample)	4,827,828	NA	Esophagus Adenocarcinoma (ESAD: 89); Esophagus Squamous Cell Carcinoma (ESSC: 96)	NA
Glioblastoma Multiforme (GBM: 175)	5	Brain (1409)	922,811 (5,428/sample)	796,026	NA	NA	Glioblastoma multiforme: 20
Kidney Chromophobe (KICH: 66)	25	Kidney (36)	667,651 (10,116/sample)	471,050	NA	NA	NA
Kidney Renal Clear Cell Carcinoma (KIRC: 544)	72	Kidney (36)	4,593,062 (8,443/sample)	2,452,391	NA	NA	Renal Cell Carcinoma: 110
Kidney Renal Papillary Cell Carcinoma (KIRP: 291)	32	Kidney (36)	2,066,360 (7,102/sample)	1,272,647	NA	NA	NA
Liver Hepatocellular Carcinoma (LIHC: 374)	50	Liver (136)	1,968,456 (5,263/sample)	1,187,750	Hepatocyte cell line (7); Hepatocyte primary cells (77)	NA	Hepatocellular carcinoma: 115

Lung Adenocarcinoma (LUAD: 542)	59	Lung (374)	2,827,233 (5,216/sample)	1,885,643	NA	NA	Lung adenocarcinoma: 35
Lung Squamous Cell Carcinoma (LUSC: 504)	51	Lung (374)	3,575,668 (7,095/sample)	1,888,833	NA	NA	NA
Ovarian Serous Cystadenocarcinoma (OV: 430)	0	Ovary (108), Fallopian Tube (7)	30,141,239 (70,096/sample)	11,483,211	NA	NA	NA
Pancreatic Adenocarcinoma (PAAD: 179)	4	Pancreas (197)	1,315,090 (7,347/sample)	830,468	NA	NA	NA
Prostate Adenocarcinoma (PRAD: 506)	52	Prostate (119)	2,476,764 (4,895/sample)	1,643,727	NA	NA	Prostate adenocarcinoma: 51
Skin Cutaneous Melanoma (SKCM: 472)	1	Skin (972)	1,935,123 (4,100/sample)	1,198,791	NA	NA	NA
Rectum Adenocarcinoma (READ: 167)	10	Colon (376)	625,914 (3,748/sample)	496,129	NA	NA	NA
Stomach Adenocarcinoma (STAD: 416)	37	Stomach (203)	12,378,838 (29,757/sample)	7,764,375	NA	NA	NA
Testicular Germ Cell Tumors (TGCT: 156)	0	Testis (203)	935,595 (5,997/sample)	573,824	NA	NA	Testicular cancer: 9
Thyroid Carcinoma (THCA: 513)	59	Thyroid (361)	1,783,492 (3,477/sample)	1,296,521	NA	NA	Thyroid carcinoma: 52
Uterine Carcinosarcoma (UCS: 57)	0	Uterus (90)	593,740 (10,416/sample)	413,469	NA	NA	NA
Uterine Corpus Endometrial Carcinoma (UCEC: 554)	35	Uterus (90)	3,542,334 (6,394/sample)	1,968,191	NA	NA	NA
Sarcoma (SARC: 263)	2	Adipose Tissue (620), Muscle (475)	1,739,153 (6,613/sample)	1,069,549	NA	Desmoid Tumor (DT: 2); Leiomyosarcoma (LMS: 106); Malignant Peripheral Nerve Sheath Tumors (MPNT: 10); Myxofibrosarcoma (MFS: 25); Synovial Sarcoma (SYNS: 10); Undifferentiated Pleomorphic Sarcoma (UPLS: 52)	Sarcoma: 84
Pheochromocytoma & Paraganglioma (PCPG: 184)	3	Adrenal Gland (159), Nerve (335)	1,168,368 (6,350/sample)	662,442	NA	Pheochromocytoma (PCHC: 151); Paraganglioma (PGG: 33)	NA

Adrenocortical Carcinoma (ACC: 79)	0	Adrenal Gland (159)	12,703,775 (160,807 /sample)	1,013,102	NA	NA	NA
Mesothelioma (MESO: 87)	0	Lung (374)	399,050 (4,587/sample)	317,630	NA	NA	NA
Head and Neck Squamous Cell Carcinoma (HNSC: 504)	44	Skin (972)	1,780,841 (3,533/sample)	1,262,383	NA	NA	NA
Cholangiocarcinoma (CHOL: 36)	9	Liver (136)	281,418 (7,817/sample)	220,929	NA	NA	Cholangio-carcinoma: 7
Thymoma (THYM: 120)	2	NA	4,228,476 (35,237 /sample)	1,206,043	Thymus primary cell (20); Thymus tissue (119)	NA	Thymoma: 15
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC: 48)	0	NA	No match: 7,722,928 (160,894 /sample)	849,592	NA	NA	Diffuse Large B-cell Lymphoma: 2
Uveal Melanoma (UVM: 80)	0	NA	No match: 13,315,153 (166,439/sample)	899,967	Melanocyte cell line (22); Melanocyte primary cell (8)	NA	Uveal Melanoma: 5
NA	NA	Heart (489)	NA	NA	NA	NA	NA
NA	NA	Pituitary (124)	NA	NA	NA	NA	NA
NA	NA	Salivary Gland (70)	NA	NA	NA	NA	NA
NA	NA	Small Intestine (104)	NA	NA	NA	NA	NA
NA	NA	Spleen (118)	NA	NA	NA	NA	NA
NA	NA	Vagina (97)	NA	NA	NA	NA	NA
NA	NA	Blood vessel (750)	NA	NA	NA	NA	NA

Supplementary Table S2.2: Percent of junctions not found in core normal samples, averaged across cancer types

SRA category:		Adult	Developmental	Stem Cells	Unexplained
Number of samples per SRA category required for set membership	1	26.5%	15.4%	2.7%	64.9%
	2	19.6%	5.92%	2.2%	76.4%

Supplementary Table S2.3: Selection of additional normal tissue and cell types analyzed

SRA cell or tissue type	Sample types (abbreviation: # of samples)	Assigned category > subcategory (if applicable)
Aorta	tissue (aor_tis: 266)	adult
Astrocyte	cell line (ast_cl: 4), primary cells (ast_pc: 83)	adult
Biliary Tree	combined group of tissue (4 samples) and stem cells (3 samples) (bt_all: 7)	adult
Bone	cell line (bone_cl: 20), tissue (bone_tis: 58)	adult
Ectoderm	cell line (ect_cl: 17)	developmental > embryonic
Embryo	cell line (emb_cl: 929), primary cells (emb_pc: 904), stem cells (emb_sc: 34), tissue (emb_tis: 989)	developmental > embryonic
Epithelial Cell	cell line (ec_cl: 853), primary cell (ec_pc: 621)	adult
	stem cells (ec_sc: 57)	stem cells
Eye	cell line (eye_cl: 42), primary cell (eye_pc: 4), tissue (eye_tis: 53)	adult
Fallopian Tube	tissue (ft_tis: 13)	adult
Fibroblast	cell line (fb_cl: 1660), primary cell (fb_pc: 351)	adult
	stem cells (fb_sc: 9)	stem cells
Glial Cell	cell line (gc_cl: 10), primary cells (gc_pc: 136)	adult
Hematopoietic Cell	cell line (hpc_cl: 603), primary cell (hpc_pc: 2679)	adult
	stem cells (hpc_sc: 18)	stem cells
Hepatocyte	cell line (hep_cl: 7), primary cell (hep_pc: 77)	adult
Induced Pluripotent Stem Cell	cell line (ips_cl: 139)	stem cells
Islet of Langerhans	cell line (ilh_cl: 3), primary cell (ilh_pc: 285)	adult
Leukocyte	cell line (lk_cl: 370), primary cell (lk_pc: 2178)	adult
Lymphocyte	cell line (lym_cl: 255), primary cell (lym_pc: 1073)	adult
Macrophage	cell line (mph_cl: 19), primary cell (mph_pc: 130)	adult
Melanocyte	cell line (mel_cl: 22), primary cell (mel_pc: 8)	adult

Mesenchymal Stem Cell	stem cells (msc_sc: 65)	stem cells
Mesenchyme	stem cells (mes_sc: 19)	stem cells
Mesothelium	cell line (mes_cl: 4)	adult
Myeloid Cell	cell line (myl_cl: 29), primary cell (myl_pc: 794)	adult
Myoblast	cell line (myo_cl: 7), primary cell (myo_pc: 402)	developmental > embryonic
Neonate	cell line (nn_cl: 250), primary cell (nn_pc: 105), tissue (nn_tis: 21)	developmental > fetal
Oligodendrocyte	primary cell (odg_pc: 37)	adult
Oocyte	primary cell (ooc_pc: 11)	developmental > oocyte
Placenta	tissue (plc_tis: 264)	developmental > placental
Platelet	primary cell (plt_pc: 6)	adult
Pluripotent Stem Cell	cell line (pps_cl: 139), stem cells (pps_sc: 6)	stem cells
Somatic Stem Cell	stem cells (ssc_sc: 86)	stem cells
Thymus	primary cell (thym_pc: 20), tissue (thym_tis: 119)	adult
Zygote	primary cell (zyg_pc: 27)	developmental > zygote

Supplementary Table S2.4: Unexplained junctions occurring in >10% of samples in multiple cancer types

Unexplained stage 3+ junction (chr; left splice site; right splice site; strand)	Cancer count	Cancer type: % of cancer-type samples containing the junction
chr18;49501731;49524503;-	3	Esophageal_Carcinoma: 21.0811%; Head_and_Neck_Squamous_Cell_Carcinoma: 28.7698%; Lung_Squamous_Cell_Carcinoma: 13.0952%
chr19;58392528;58393026;+	3	Esophageal_Carcinoma: 11.3514%; Ovarian_Serous_Cystadenocarcinoma: 19.3023%; Stomach_Adenocarcinoma: 12.9808%
chr11;2129478;2395551;+	2	Adrenocortical_Carcinoma: 13.9241%; Uterine_Carcinosarcoma: 10.5263%
chr11;22900551;23057663;-	2	Skin_Cutaneous_Melanoma: 11.4407%; Uveal_Melanoma: 12.5%
chr12;15569284;15578851;+	2	Colon_Adenocarcinoma: 15.4455%; Rectum_Adenocarcinoma: 16.1677%
chr12;15579385;15580037;+	2	Colon_Adenocarcinoma: 16.6337%; Rectum_Adenocarcinoma: 16.7665%
chr14;100376176;100734652;+	2	Adrenocortical_Carcinoma: 17.7215%; Pheochromocytoma_and_Paraganglioma: 22.2826%
chr18;49524635;49578176;-	2	Esophageal_Carcinoma: 15.6757%; Head_and_Neck_Squamous_Cell_Carcinoma: 18.6508%
chr19;3976571;3982811;-	2	Acute_Myeloid_Leukemia: 13.4921%; Ovarian_Serous_Cystadenocarcinoma: 10.4651%
chr19;41683816;41771127;-	2	Esophageal_Carcinoma: 14.5946%; Stomach_Adenocarcinoma: 15.8654%
chr19;41709948;41718251;+	2	Esophageal_Carcinoma: 11.3514%; Stomach_Adenocarcinoma: 10.8173%
chr21;16971056;16971076;+	2	Esophageal_Carcinoma: 13.5135%; Ovarian_Serous_Cystadenocarcinoma: 18.1395%
chr6;116004903;116119064;+	2	Esophageal_Carcinoma: 11.8919%; Stomach_Adenocarcinoma: 11.2981%
chr6;116004910;116119071;+	2	Esophageal_Carcinoma: 11.8919%; Stomach_Adenocarcinoma: 11.2981%
chr6;31944795;31944981;+	2	Ovarian_Serous_Cystadenocarcinoma: 19.5349%; Stomach_Adenocarcinoma: 11.7788%
chrX;4542633;4895036;+	2	Esophageal_Carcinoma: 10.8108%; Stomach_Adenocarcinoma: 11.0577%

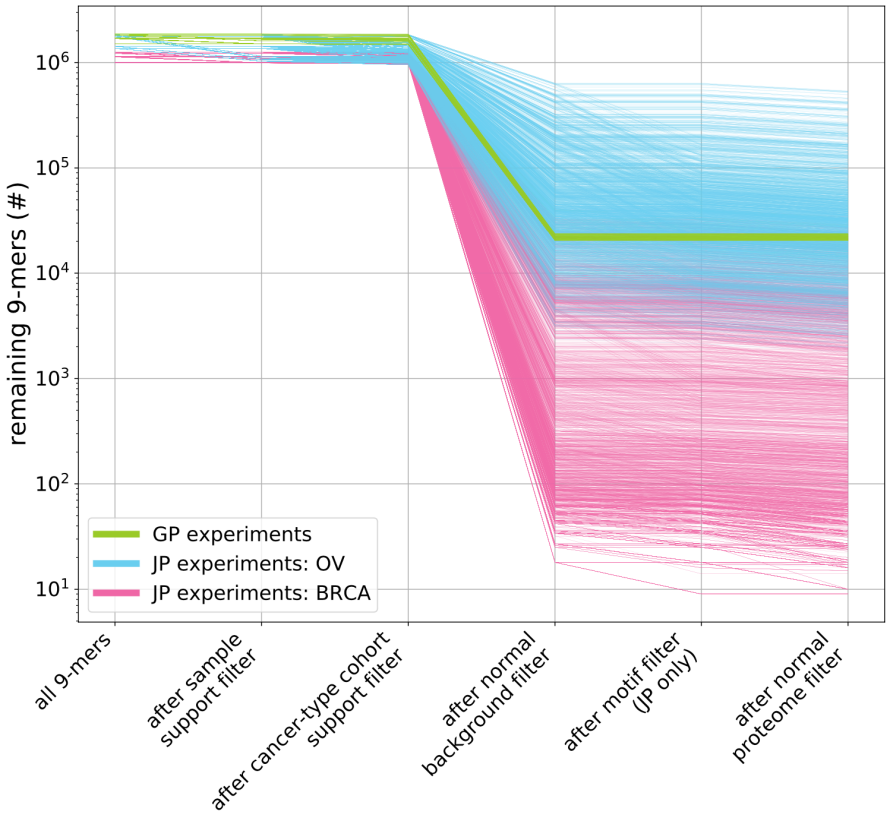
Supplementary Table S2.5: Junction counts and ratio of antisense junctions for TCGA cancer types

	Core normals		Other adult non-cancer		Developmental		Stem cell		Unexplained	
TCGA cancer type	junction count	antisense prevalence	junction count	antisense prevalence	junction count	antisense prevalence	junction count	antisense prevalence	junction count	antisense ratio
Acute Myeloid Leukemia	1,745,361	27%	183,003	46%	46,261	48%	4,580	45%	427,768	47%
Adrenocortical Carcinoma	941,404	20%	7,830	39%	2,369	38%	326	33%	22,053	44%
Bladder Urothelial Carcinoma	2,401,749	24%	58,874	35%	17,652	31%	3,348	26%	155,124	42%
Brain Lower Grade Glioma	2,847,326	26%	64,093	40%	19,716	38%	3,129	32%	165,091	41%
Breast Invasive Carcinoma	4,013,058	27%	170,106	36%	44,855	36%	7,520	30%	441,104	37%
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	2,148,683	24%	45,625	36%	12,353	35%	2,361	27%	111,091	42%
Cholangiocarcinoma	754,166	15%	3,790	33%	900	33%	168	24%	8,399	37%
Colon Adenocarcinoma	2,228,830	25%	64,988	37%	19,894	31%	2,841	32%	155,589	42%
Esophageal Carcinoma	3,939,154	29%	349,351	43%	113,466	45%	13,813	38%	1,039,014	46%
Glioblastoma Multiforme	2,368,563	25%	35,154	36%	12,209	35%	2,024	26%	75,111	37%
Head and Neck Squamous Cell Carcinoma	2,571,867	24%	72,107	35%	19,318	34%	4,210	27%	177,027	40%
Kidney Chromophobe	1,042,201	22%	10,778	42%	3,081	37%	490	26%	28,961	42%
Kidney Renal Clear Cell Carcinoma	2,925,458	26%	93,435	38%	21,535	38%	3,160	32%	197,822	40%
Kidney Renal Papillary Cell Carcinoma	1,896,671	25%	33,239	39%	8,380	38%	1,255	33%	85,284	41%
Liver Hepatocellular Carcinoma	1,889,979	23%	43,871	36%	9,302	33%	1,486	27%	91,086	40%
Lung Adenocarcinoma	2,890,617	26%	95,291	37%	23,848	35%	4,387	30%	229,442	40%
Lung Squamous Cell Carcinoma	3,018,232	24%	92,301	33%	27,381	30%	5,604	22%	212,501	36%
Lymphoid Neoplasm Diffuse Large B cell Lymphoma	793,306	18%	9,851	30%	1,806	32%	277	30%	16,201	36%
Mesothelioma	1,150,647	20%	11,387	38%	3,060	39%	594	29%	27,214	43%
Ovarian Serous Cystadenocarcinoma	5,092,658	32%	542,738	46%	242,074	48%	24,032	41%	2,739,236	49%
Pancreatic Adenocarcinoma	1,593,779	22%	20,854	39%	5,087	41%	876	33%	49,429	44%
Pheochromocytoma and Paraganglioma	1,480,083	23%	19,085	40%	6,003	40%	832	35%	53,453	43%
Prostate Adenocarcinoma	2,423,109	26%	55,616	40%	15,372	40%	2,257	35%	159,220	43%
Rectum Adenocarcinoma	1,381,622	22%	21,823	37%	6,693	33%	945	28%	42,408	42%

Sarcoma	1,985,824	23%	31,671	35%	10,306	33%	1,699	28%	77,762	39%
Skin Cutaneous Melanoma	2,493,051	24%	62,444	36%	17,376	34%	2,973	27%	148,206	38%
Stomach Adenocarcinoma	4,844,395	31%	494,846	44%	172,922	46%	19,099	40%	1,787,745	47%
Testicular Germ Cell Tumors	1,627,329	19%	26,027	26%	12,435	23%	5,054	17%	49,836	31%
Thymoma	1,364,454	21%	19,938	32%	4,892	31%	1,057	26%	40,450	37%
Thyroid Carcinoma	2,501,295	27%	54,285	42%	14,124	42%	2,217	37%	147,977	44%
Uterine Carcinosarcoma	993,014	18%	8,062	35%	3,328	30%	601	29%	22,474	39%
Uterine Corpus Endometrial Carcinoma	2,499,341	26%	70,746	36%	20,596	35%	3,226	31%	155,889	41%
Uveal Melanoma	849,072	20%	7,281	40%	1,799	42%	276	35%	16,284	42%

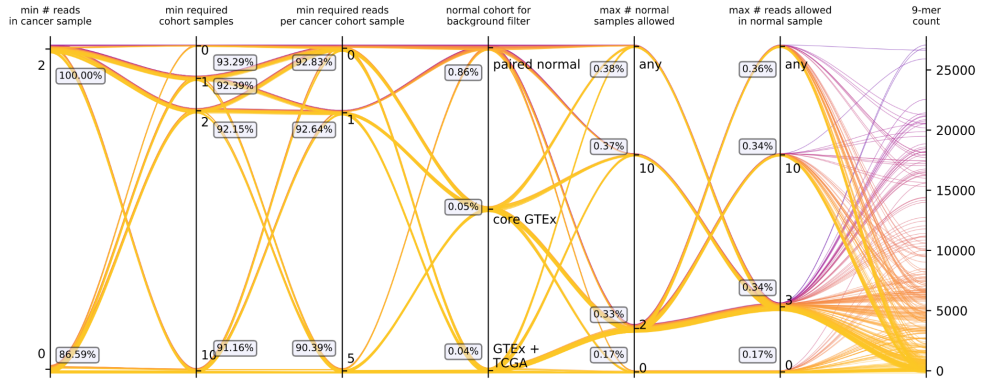
Supplementary Table S2.6: Genes not currently cancer-associated with high novel junction burdens

Gene	Unexplained junctions (#)	Cancer types with >5% of samples containing unexplained junctions in this gene (#)
BCAM	212	1
MPO	121	1
MSLN	102	1
CLDN3	87	1
CHGA	69	1
LGALS3BP	36	2
SSPN	35	1
COL1A2	33	3
PKM	32	1
EEF2	31	2
SPON1	31	1
KRT7	31	1
ZNF503	30	1
C3	29	1
H1FX	28	1
MARCKSL1	27	1
CRIP2	26	1
AC007040.11	25	1
TNFAIP2	25	1
AGRN	25	2

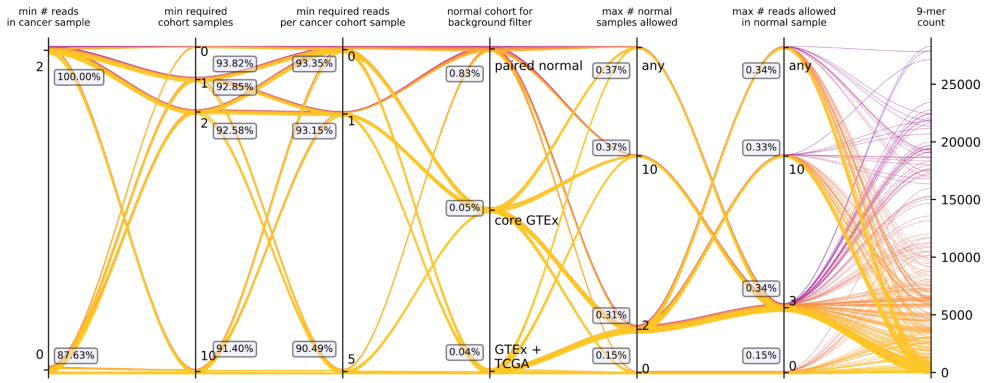


Supplementary Figure 3.1: Effect of filter steps on remaining 9-mer counts in $\log_{10}$ scale. Each line represents the number of initial junction-spanning 9-mers generated for each sample at the far left and the number of 9-mers remaining (y-axis, log scale) for each filter experiment after each filter stage (x-axis). Lines are colored by the sample type and pipeline (green = GP on BRCA samples, pink = JP on BRCA samples, blue = JP on OV samples).

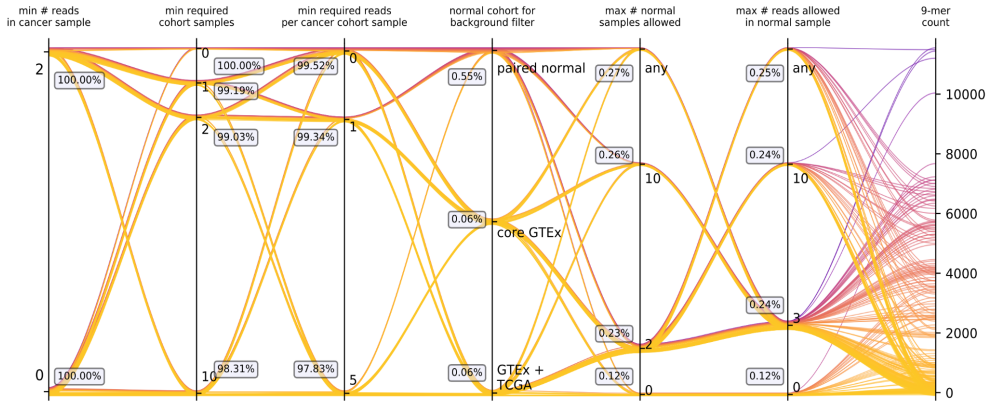
(A)


min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
paired normal
core GTEx
GTEx + TCGA
any

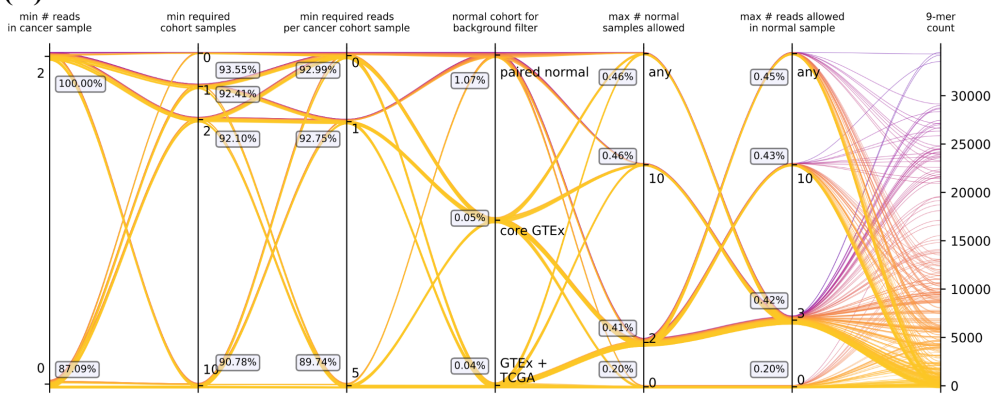
(B)


min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
paired normal
core GTEx
GTEx + TCGA
any

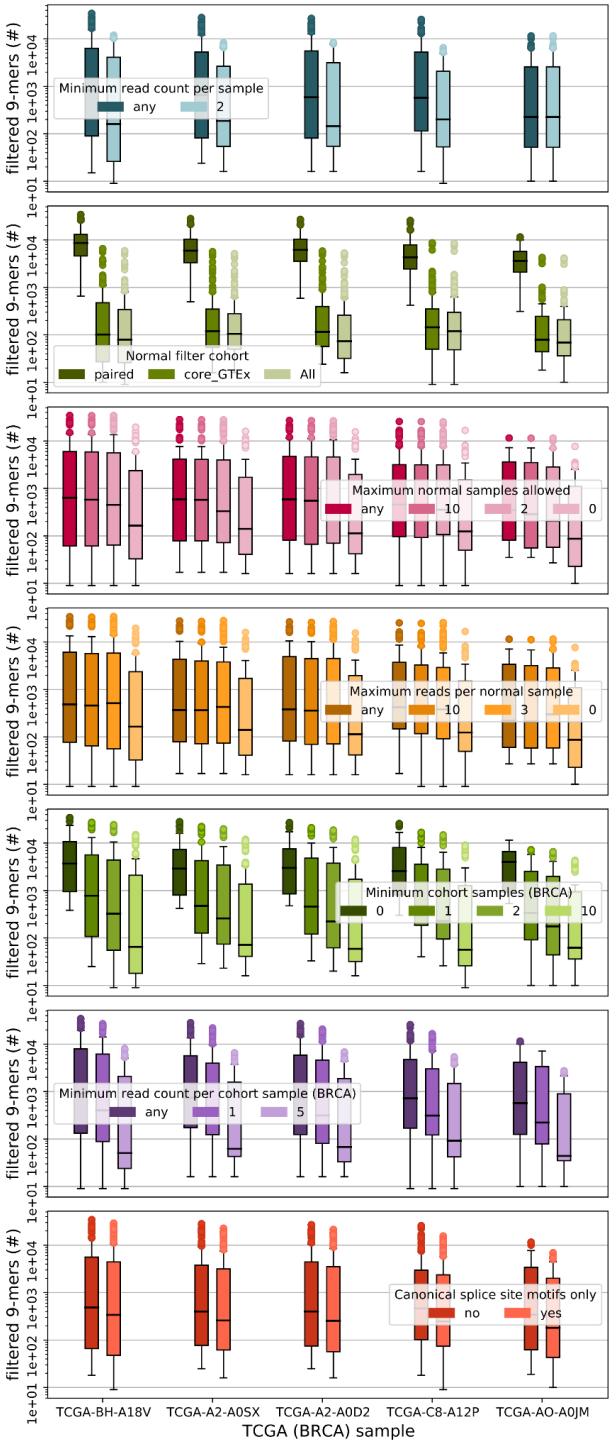
(C)


min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
paired normal
core GTEx
GTEx + TCGA
any

(D)

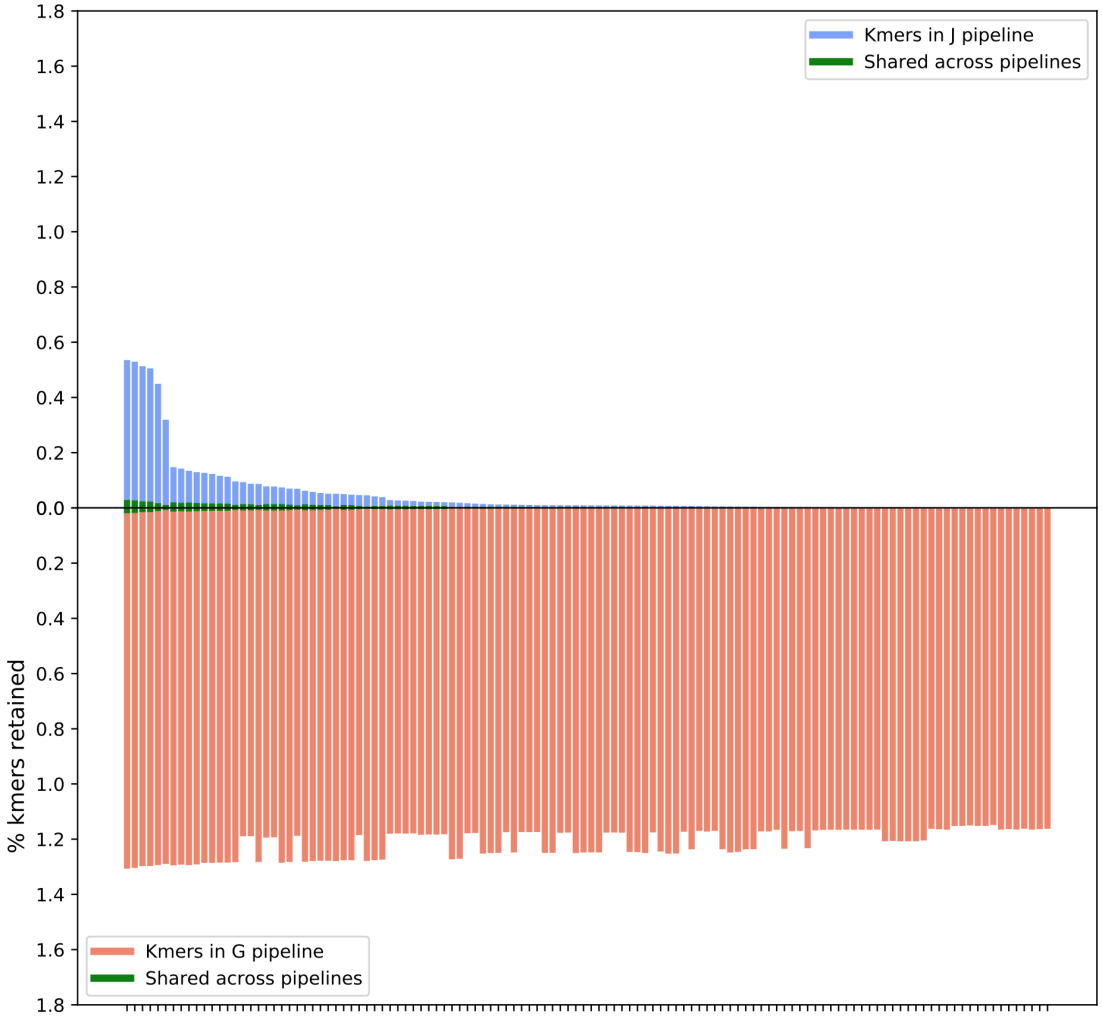

min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
paired normal
core GTEx
GTEx + TCGA
any

Supplementary Figure 3.2: Effect of JP filters on final 9-mer count for BRCA samples. For samples **(A)** TCGA-A2-A0D2, **(B)** TCGA-A2-A0SX, **(C)** TCGA-AO-A0JM, and **(D)** TCGA-BH-A18V. Each vertical axis except the rightmost represents one filter, with its parameter options arranged from most stringent (bottom) to most lenient (top). Each colored line represents one JP filter experiment, with its path passing through the filters parameters it uses and its color mapped to the final number of 9-mers passing the full set of filters (yellow == low, purple == high). The rightmost axis shows final filtered 9-mer counts for each filter experiment, with each filter experiment colored line terminating at its final value. Floating gray boxes show, across experiments passing through the corresponding filter parameter, the mean of the ratio of remaining 9-mers after that filter parameter has been applied to the total initial generated 9-mers for the sample.

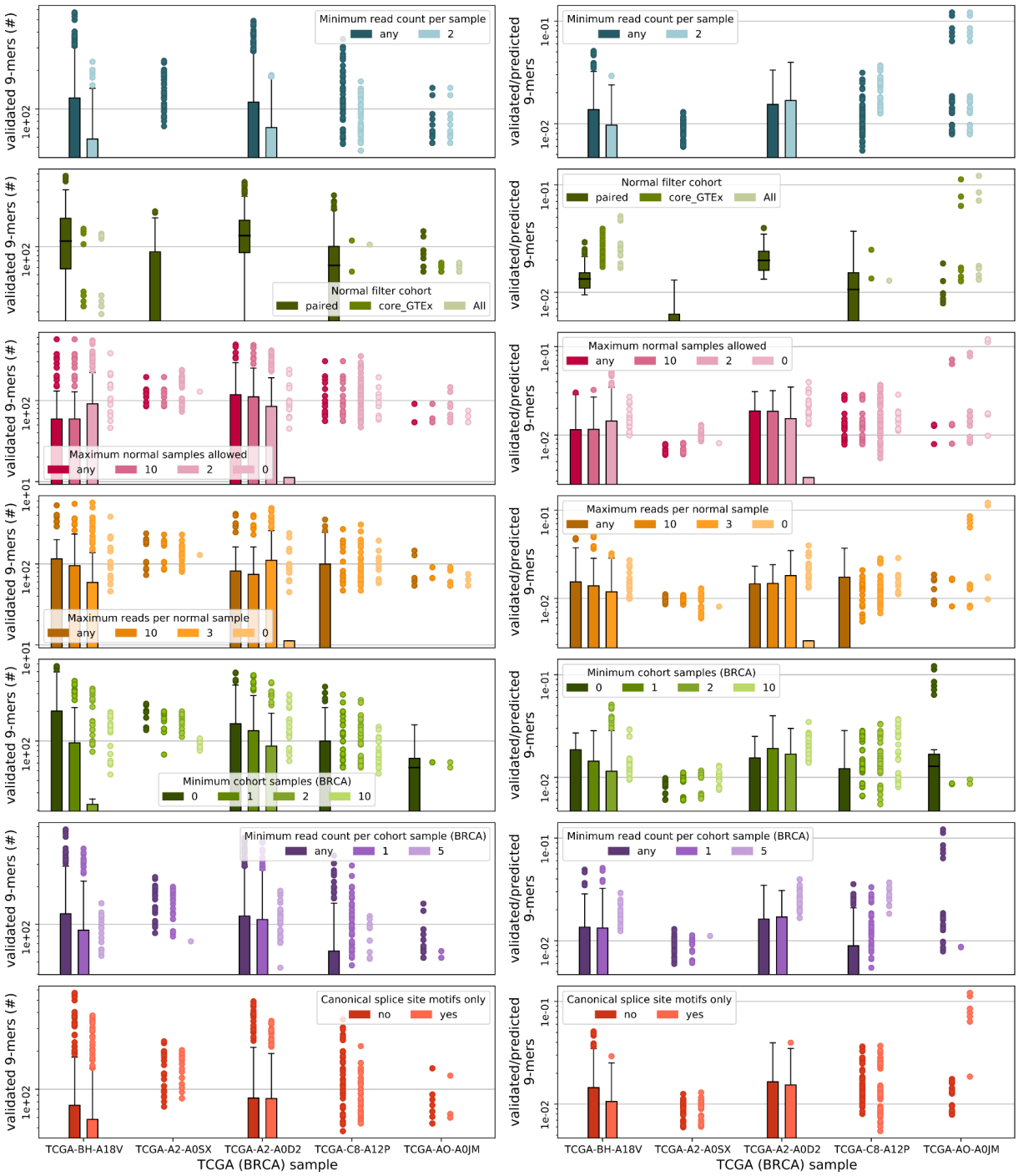


Supplementary Figure 3.3: Aggregate effect of JP filter parameters on filtered 9-mer count across BRCA samples. Panels contain boxplots of filter experiment results, with each point representing the filtered 9-mer count (y-axis) for one filter experiment and one sample (within-panel boxplot sets, left to right, TCGA-BH-A18V, TCGA-A2-A0SX, TCGA-A2-A0D2, TCGA-C8-A12P, and TCGA-AO-A0JM). Each panel represents the same data, with boxplots broken out according to parameter options used for each filter type, top to bottom, the minimum expression (normalized read count) in the target sample; the normal filter cohort used; the maximum number of normal samples allowed with junction expression; the

maximum expression (normalized read count) per normal sample; the minimum number of cancer cohort samples with junction support; the minimum number of reads required per cohort sample; and the canonical motif filter. Within each panel, the box color saturation corresponds to filter leniency (the lightest color represents the most stringent parameter and the darkest color is the most lenient value).

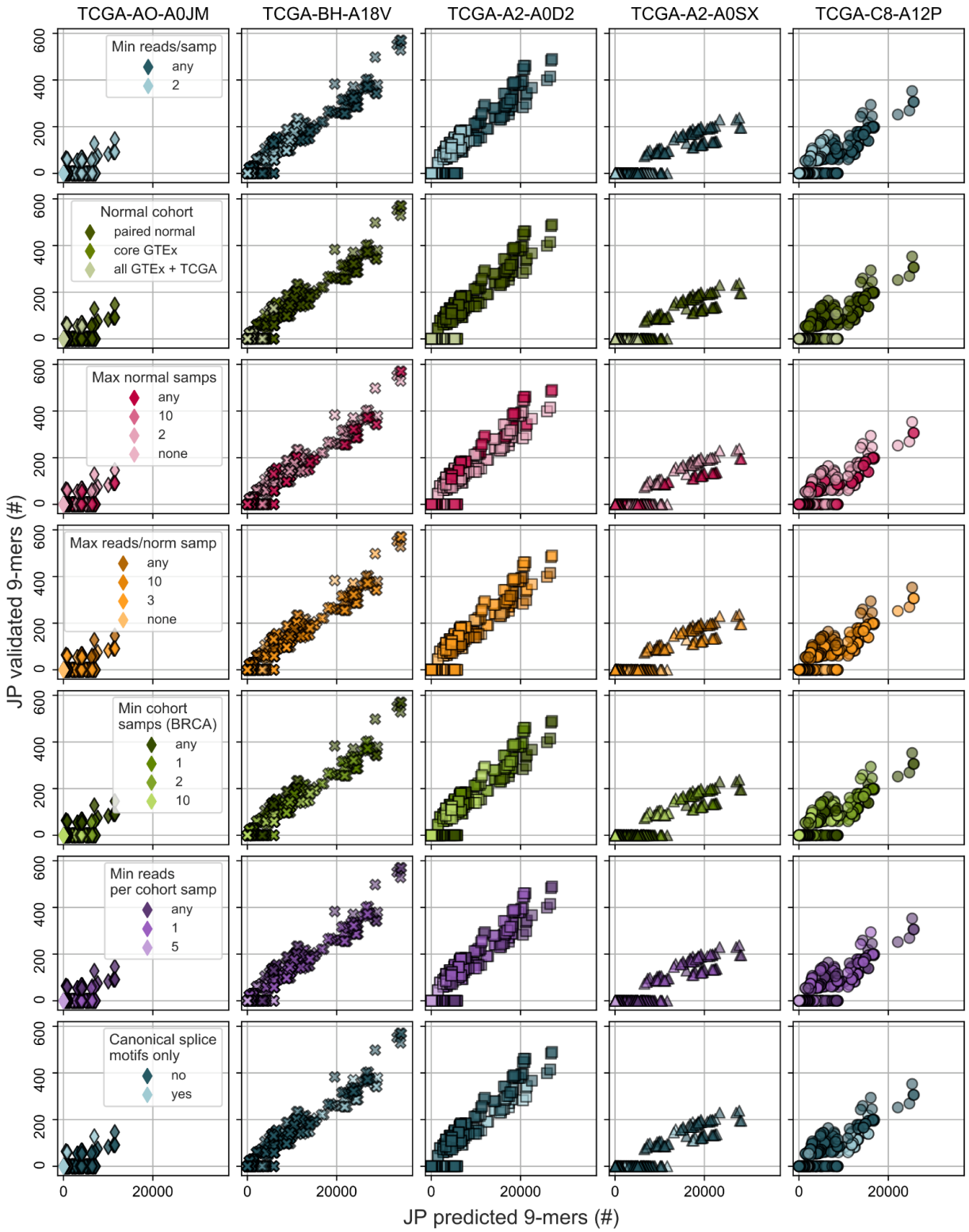


Supplementary Figure 3.4: Proportion and overlap of TCGA-AO-A0JM 9-mers predicted by two pipelines. Bars show proportion of generated 9-mers ("kmers") retained by the JP (top panel) and GP (bottom panel) after filtering for each of the subset of experiments that include the core GTEx normal cohort and no motif filter; experiments are sorted by most to fewest 9-mers output by the JP. Blue indicates 9-mers present after JP filtering only, red shows 9-mers present after GP filtering only, and green on top and bottom shows the shared 9-mers present across both pipelines after filtering for each experiment.



Supplementary Figure 3.5: Aggregate effect of JP filter parameters on validated 9-mer counts and validation ratios across BRCA samples. Panels contain boxplots of filter experiment results, with each point representing, in the left column, the validated k-mer count and in the right column, validation ratio (y-axis) for one filter experiment and one sample (within-panel boxplot sets, left to right, TCGA-BH-A18V, TCGA-A2-A0SX, TCGA-A2-A0D2, TCGA-C8-A12P, and TCGA-AO-A0JM). Each panel represents the same data, with boxplots broken out according to parameter options used for each filter type, top to bottom, the minimum expression (normalized read count) in the target sample; the normal filter cohort used; the maximum number of normal samples allowed with junction expression; the

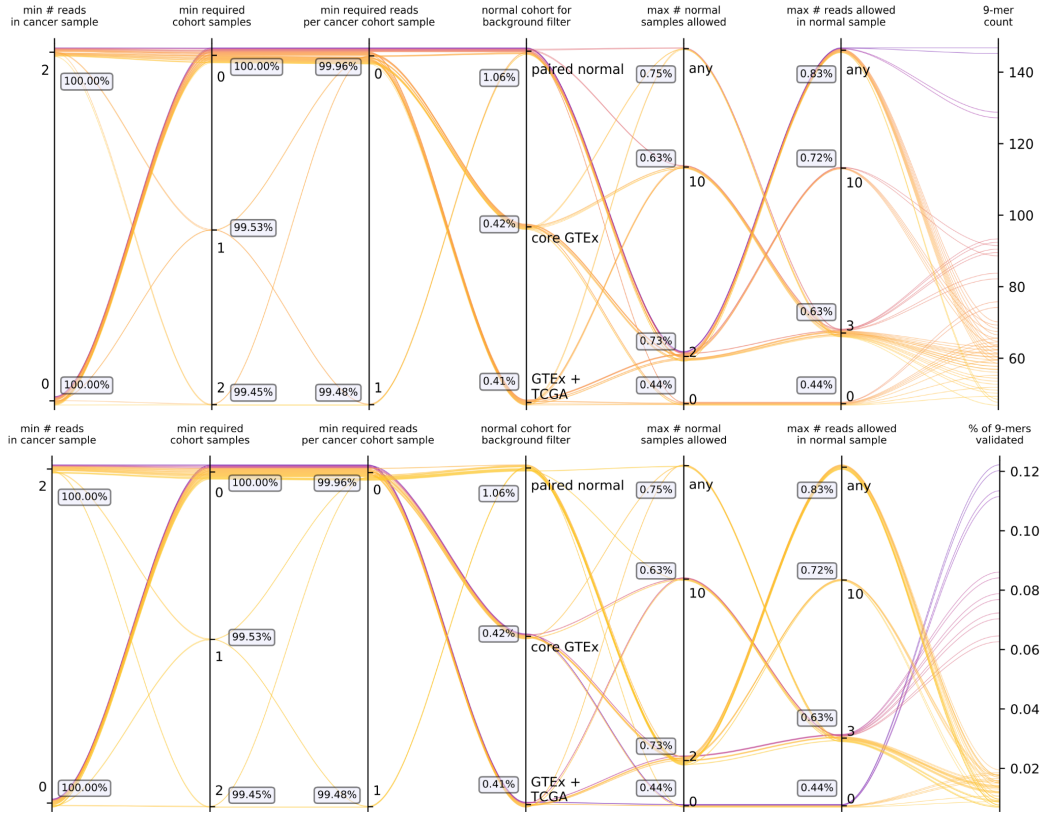
maximum expression (normalized read count) per normal sample; the minimum number of cancer cohort samples with junction support; the minimum number of reads required per cohort sample; and the canonical motif filter. Within each panel, the box color saturation corresponds to filter leniency (the lightest color represents the most stringent parameter and the darkest color is the most lenient value).



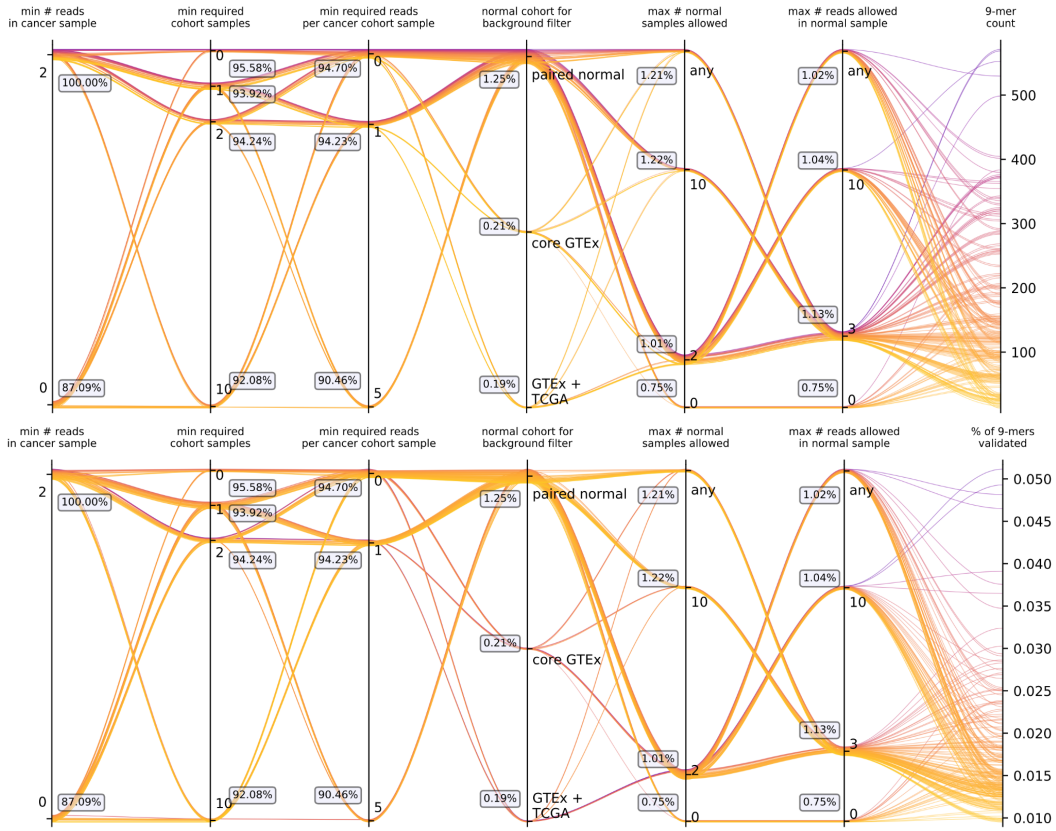
Supplementary Figure 3.6: Effect of JP filter parameters on validated and filtered junction peptide counts across BRCA samples. Panels contain scatter plots of filter experiment results, with each point representing the number of validated junction 9-mers (y-axis) vs. the number of predicted (filtered) junction 9-mers for one filter experiment. Panel columns represent samples, left to right, TCGA-AO-

A0JM, TCGA-BH-A18V, TCGA-A2-A0D2, TCGA-A2-A0SX, and TCGA-C8-A12P, with each column panel showing the same data but colored according to parameter options used for each filter type. Rows top to bottom show the minimum expression (normalized read count) in the target sample; the maximum number of normal samples allowed with junction expression; the maximum expression (normalized read count) per normal sample; the minimum number of cancer cohort samples with junction support; and the minimum number of reads required per cohort sample. Within each panel, the point color saturation corresponds to filter leniency (the lightest color represents the most stringent parameter and the darkest color is the most lenient value).

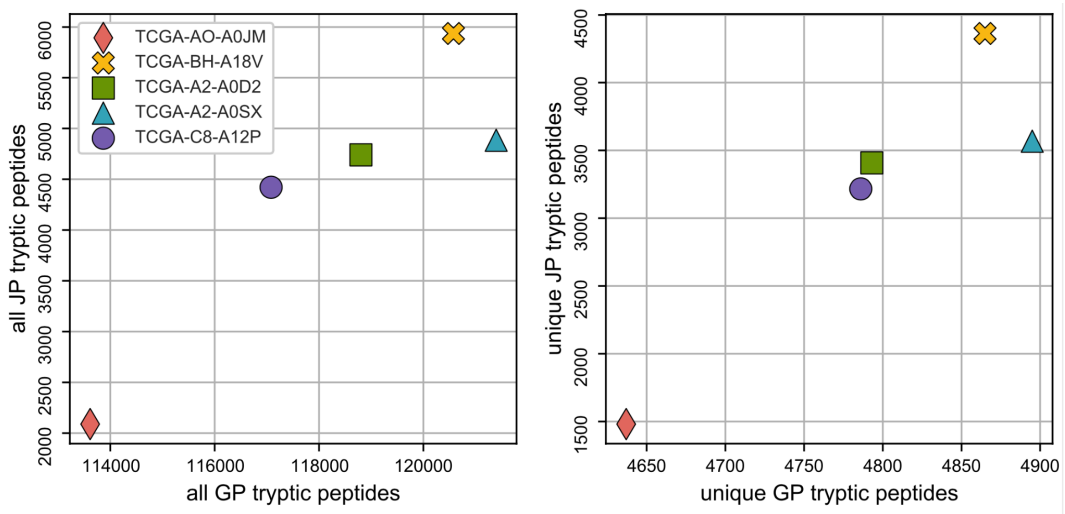
(A)


min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
% of 9-mers validated
paired normal
core GTEx
GTEx + TCGA
any
100.00%
99.96%
99.53%
99.45%
99.48%
1.06%
0.42%
0.41%
0.75%
0.63%
0.73%
0.44%
0.83%
0.72%
0.63%
0.44%

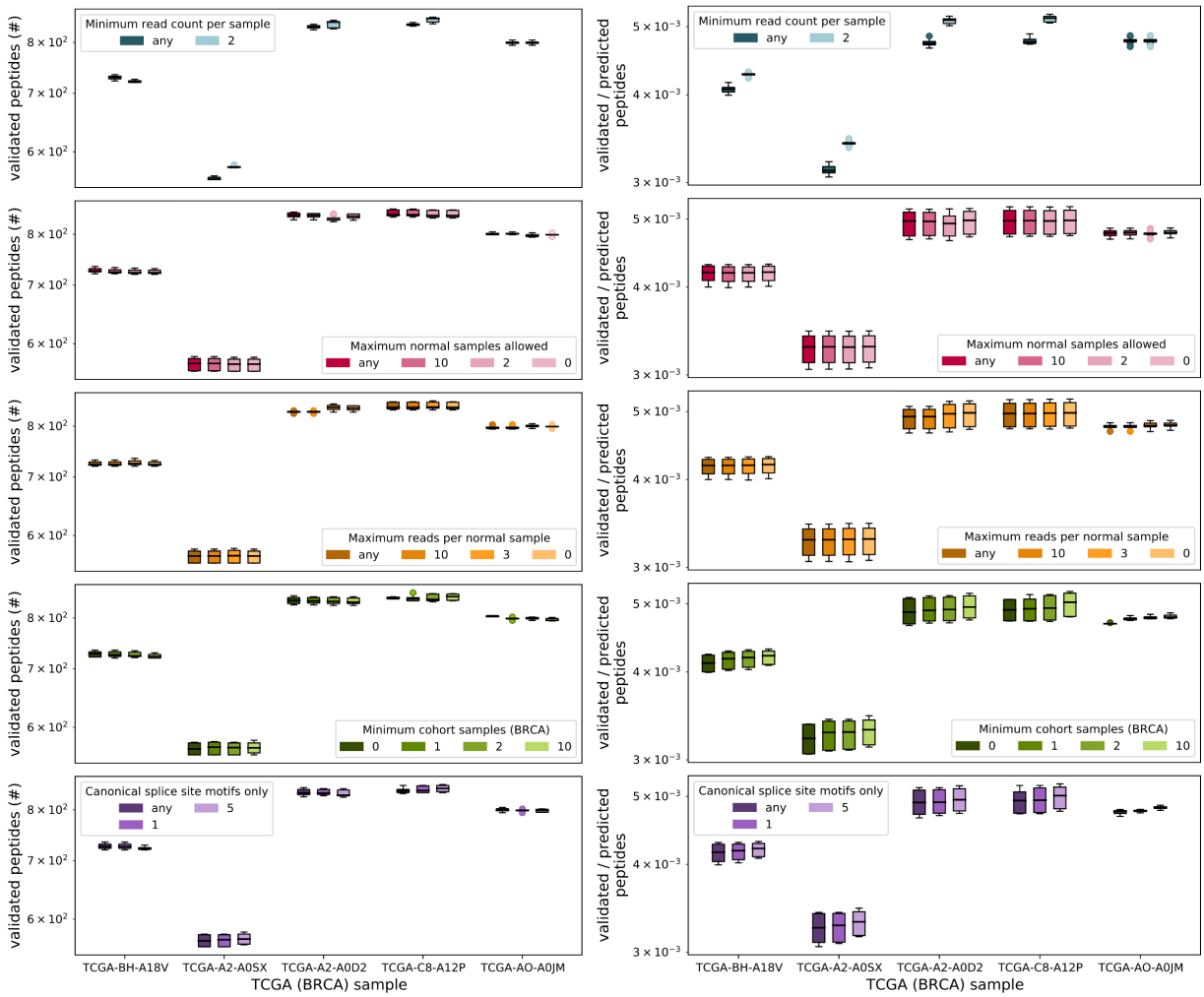
(B)


min # reads in cancer sample
min required cohort samples
min required reads per cancer cohort sample
normal cohort for background filter
max # normal samples allowed
max # reads allowed in normal sample
9-mer count
% of 9-mers validated
paired normal
core GTEx
GTEx + TCGA
any
100.00%
95.58%
93.92%
94.24%
94.70%
94.23%
87.09%
92.08%
90.46%
1.25%
0.21%
0.19%
1.21%
1.22%
1.01%
0.75%
1.02%
1.04%
1.13%

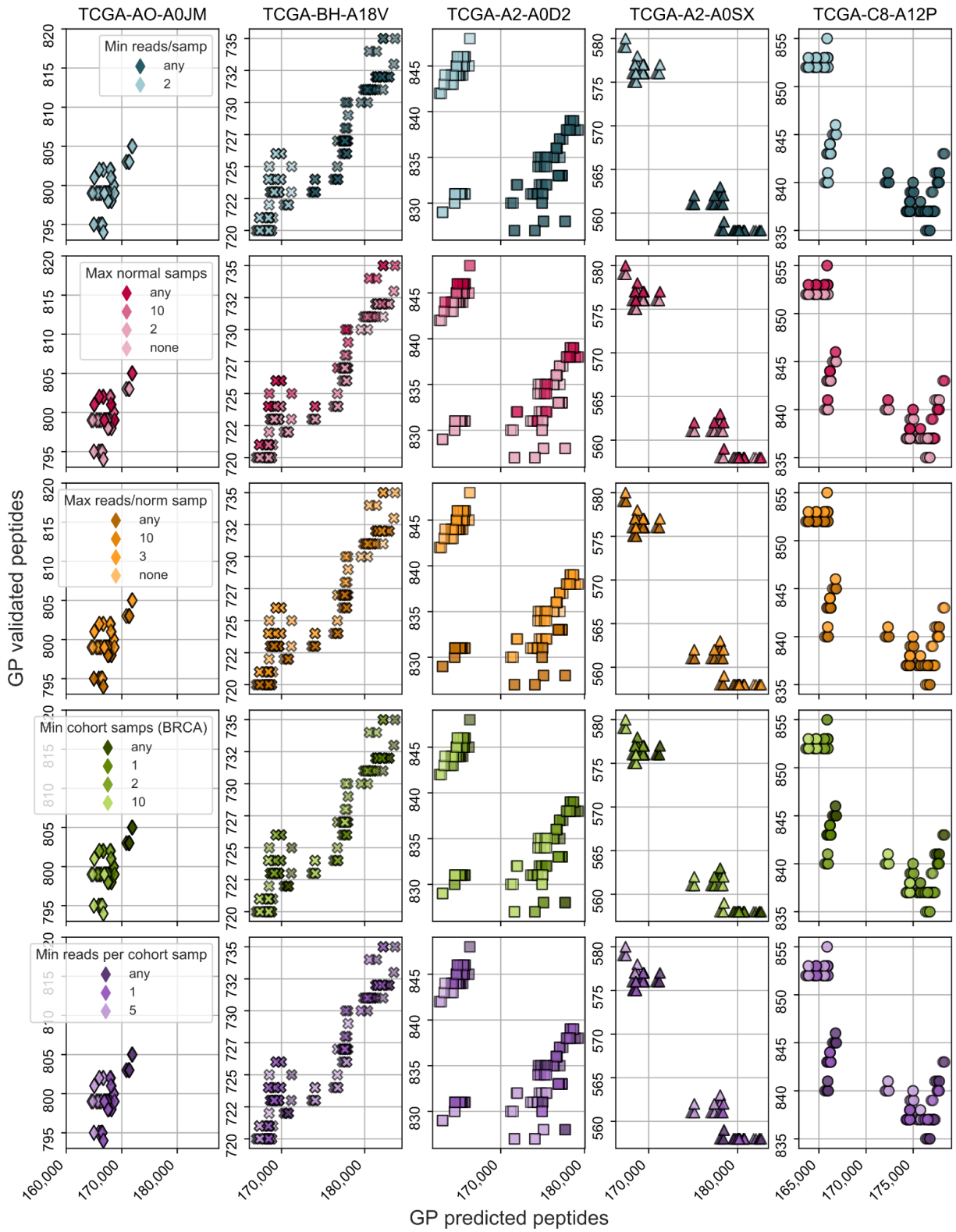
Supplementary Figure 3.7: Effect of JP filters on validated 9-mer count and validation ratios for BRCA samples. For samples **(A)** TCGA-AO-A0JM and **(B)** TCGA-BH-A18V, each vertical axis except the rightmost represents one filter, with its parameter options arranged from most stringent (bottom) to most lenient (top). Each colored line represents one filter experiment, with its path passing through the filters parameters it uses and its color mapped to its validated 9-mer count (top) and validation ratio (bottom) (yellow == low, purple == high). The rightmost axis shows validated 9-mer count (top) and validation ratio (bottom) for each filter experiment, with each filter experiment colored line terminating at its final value. Floating gray boxes show, across experiments passing through the corresponding filter parameter, the mean of the ratio of remaining 9-mers after that filter parameter has been applied to the total initial generated 9-mers for the sample.



Supplementary Figure 3.8: Tryptic peptide counts per sample resulting from the JP and the GP. Left, all tryptic peptides generated by the JP filter experiments (y-axis) and GP filter experiments (x-axis) for each of 5 BRCA samples, indicated by color and shape (legend). Right, the unique set of tryptic peptides generated by the JP filter experiments (y-axis) and GP filter experiments (x-axis) for each of 5 BRCA samples, with point color and shape matched to the left panel.



Supplementary Figure 3.9: Aggregate effect of GP filter parameters on validated junction peptide counts and validation ratios across BRCA samples. Panels contain boxplots of filter experiment results, with each point representing, in the left column, the validated junction peptide count and in the right column, validation ratio (y-axis) for one filter experiment and one sample (within-panel boxplot sets, left to right, TCGA-BH-A18V, TCGA-A2-A0SX, TCGA-A2-A0D2, TCGA-C8-A12P, and TCGA-AO-A0JM). Each panel represents the same data, with boxplots broken out according to parameter options used for each filter type, top to bottom, the minimum expression (normalized read count) in the target sample; the maximum number of normal samples allowed with junction expression; the maximum expression (normalized read count) per normal sample; the minimum number of cancer cohort samples with junction support; and the minimum number of reads required per cohort sample. Within each panel, the box color saturation corresponds to filter leniency (the lightest color represents the most stringent parameter and the darkest color is the most lenient value).



Supplementary Figure 3.10: Effect of GP filter parameters on validated and filtered junction peptide counts across BRCA samples. Panels contain scatter plots of filter experiment results, with each point representing the number of validated junction peptides (y-axis) vs. the number of predicted (filtered)

junction peptides for one filter experiment. Panel columns represent samples, left to right, TCGA-AO-A0JM, TCGA-BH-A18V, TCGA-A2-A0D2, TCGA-A2-A0SX, and TCGA-C8-A12P, with each column panel showing the same data but colored according to parameter options used for each filter type. Rows top to bottom show the minimum expression (normalized read count) in the target sample; the maximum number of normal samples allowed with junction expression; the maximum expression (normalized read count) per normal sample; the minimum number of cancer cohort samples with junction support; and the minimum number of reads required per cohort sample. Within each panel, the point color saturation corresponds to filter leniency (the lightest color represents the most stringent parameter and the darkest color is the most lenient value).

Supplementary Table 3.1: JP annotation and summary across cancer types, for junctions and 9-mers.

	BRCA			OV		
JP sample summary	junctions: mean (min-max)	9mers: mean (min-max)	non-Uniprot 9mers: mean (min-max)	junctions: mean (min-max)	9mers: mean (min-max)	non-Uniprot 9mers: mean (min-max)
total	265,392 (209,648-305,546)	1,186,459 (1,032,360-1,288,043)	242,357 (155,148-294,203)	444,599 (346,163-519,582)	1,715,993 (1,420,563-1,949,306)	688,141 (469,717-883,759)
annotated	166,053 (62.57%) (151,956-178,418)	1,014,348 (85.49%) (939,691-1,078,874)	76,391 (31.52%) (66,278-84,109)	171,488 (38.57%) (157,958-177,873)	1,044,667 (60.88%) (970,020-1,077,875)	80,492 (11.70%) (73,593-83,577)
exon skips	10,516 (3.96%) (5,214-14,854)	58,069 (4.89%) (28,720-82,500)	52,362 (21.61%) (25,778-74,608)	14,452 (3.25%) (10,937-18,078)	79,981 (4.66%) (61,587-100,731)	71,857 (10.44%) (55,408-90,368)
half-annotated	36,326 (13.69%) (20,491-46,123)	109,384 (9.22%) (59,418-137,473)	101,859 (42.03%) (54,973-128,331)	56,580 (12.73%) (44,830-65,843)	187,847 (10.95%) (154,708-224,092)	172,662 (25.09%) (142,593-205,681)
unannotated	52,497 (19.78%) (31,987-68,131)	14,864 (1.25%) (9,932-18,500)	12,739 (5.26%) (8,572-16,110)	202,079 (45.45%) (132,438-260,648)	433,401 (25.26%) (215,340-600,355)	365,814 (53.16%) (180,164-508,748)
GTAG motif	255,905 (96.43%) (204,142-293,153)	1,160,758 (97.83%) (1,012,592-1,260,715)	228,061 (94.10%) (145,234-279,922)	348,072 (78.29%) (287,753-389,213)	1,315,557 (76.66%) (1,178,947-1,403,935)	349,510 (50.79%) (283,108-417,994)
GCAG motif	4,733 (1.78%) (3,165-5,855)	13,005 (1.10%) (10,635-14,694)	4,265 (1.76%) (2,885-5,115)	25,471 (5.73%) (15,730-34,666)	46,384 (2.70%) (29,859-59,713)	31,034 (4.51%) (17,704-41,669)
ATAC motif	384 (0.14%) (288-444)	1,369 (0.12%) (1,248-1,430)	205 (0.08%) (155-246)	1,012 (0.23%) (748-1,301)	2,751 (0.16%) (2,055-3,347)	1,317 (0.19%) (745-1,840)
noncanonical motif	4,370 (1.65%) (2,053-6,111)	12,696 (1.07%) (8,880-17,453)	9,917 (4.09%) (6,919-14,254)	70,044 (15.75%) (35,718-100,083)	367,854 (21.44%) (180,953-512,555)	307,846 (44.74%) (149,533-431,095)

Supplementary Table 3.2: Summary of JP junction translation across cancer types.

JP sample summary	BRCA mean (min-max)	OV mean (min-max)
total junctions	265,392 (209,648-305,546)	444,599 (346,163-519,582)
overlapping a protein coding gene	222,231 (83.74%) (182,638-248,520)	332,660 (74.82%) (271,660-384,046)
fully within a protein coding gene	212,485 (80.06%) (177,240-236,869)	304,940 (68.59%) (249,353-350,418)
5' splice site in an exon CDS	194,550 (73.31%) (166,566-213,802)	275,833 (62.04%) (226,843-314,492)
translated	142,677 (53.76%) (122,619-155,842)	208,121 (46.81%) (170,778-238,681)
total junction-overlapping 9mers	1,186,459 (1,032,360-1,288,043)	1,715,993 (1,420,563-1,949,306)
junction-overlapping 9mers not in Uniprot	242,357 (20.43%) (155,148-294,203)	688,141 (40.10%) (469,717-883,759)

Supplementary Table 3.3: Mutual GP and JP junctions and translation across BRCA samples.

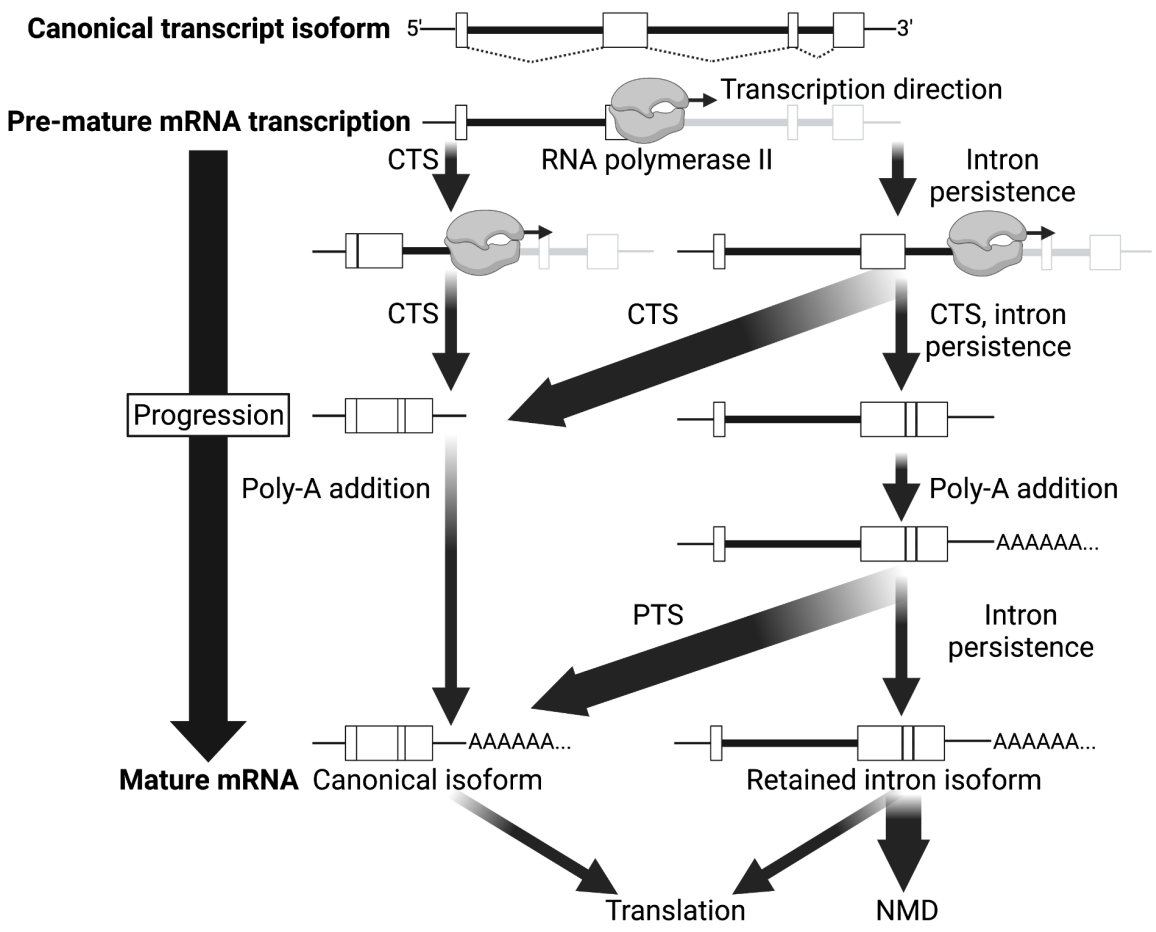
	junction-based pipeline	graph-based pipeline
total junctions	489,995	n/a
mutual junctions	270,785 (55.26%)	n/a
translated junctions	212,001 (43.27%)	713,703 (n/a)
total peptides	326,472	1,376,720
uniquely translated junctions	19,258	520,960
mutually translated junctions (MJ)	192,743	
unique MJ peptides	297,043	520,097
peptide-MJ pairs	307,773	521,765
MJ pairs with match	301,485 (97.96%)	404,899 (77.60%)
exact matches	14,721 (4.78%)	18,384 (3.52%)
contained matches	229,854 (74.68%)	287,134 (55.03%)
overlap matches	56,910 (18.49%)	99,381 (19.05%)

Supplementary Table 3.4: Annotation and motif summary for JP filtered junctions and 9-mers.

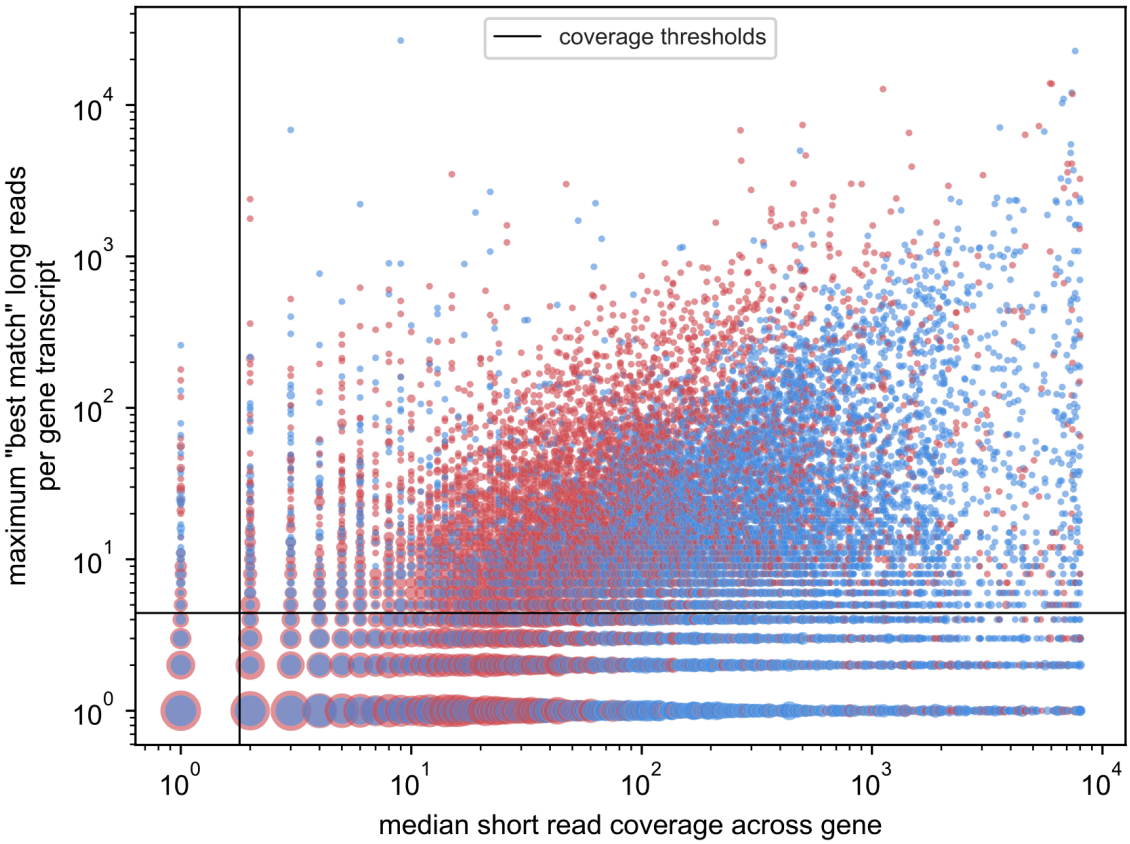
	BRCA			**OV**		
JP filter output summary	junctions: mean (min-max)	9mers: mean (min-max)	non-Uniprot 9mers: mean (min-max)	junctions: mean (min-max)	9mers: mean (min-max)	non-Uniprot 9mers: mean (min-max)
total	4,237 (1,887-5,842)	34,295 (15,207-46,996)	29,183 (13,000-40,148)	57,114 (31,288-77,863)	483,605 (266,447-653,210)	404,558 (220,137-549,079)
annotated	317 (7.48%) (141-467)	2,693 (7.85%) (1,188-3,954)	423 (1.45%) (225-638)	1,395 (2.44%) (780-2,016)	11,904 (2.46%) (6,678-17,155)	1,878 (0.46%) (1,242-2,367)
exon skips	894 (21.10%) (286-1,548)	7,178 (20.93%) (2,289-12,418)	6,415 (21.98%) (2,034-11,124)	2,401 (4.20%) (1,624-3,352)	19,196 (3.97%) (13,000-26,843)	16,890 (4.17%) (11,458-23,594)
half-annotated	2,005 (47.32%) (766-2,924)	15,856 (46.23%) (6,025-23,133)	14,755 (50.56%) (5,585-21,574)	8,014 (14.03%) (5,619-10,490)	65,452 (13.53%) (46,216-85,385)	58,709 (14.51%) (41,613-76,332)
unannotated	1,021 (24.10%) (694-1,423)	8,605 (25.09%) (5,716-11,903)	7,610 (26.08%) (5,160-10,710)	45,305 (79.32%) (22,037-63,179)	388,606 (80.36%) (192,205-539,565)	327,482 (80.95%) (160,861-456,425)
GTAG motif	3,226 (76.14%) (1,199-4,882)	25,892 (75.50%) (9,498-39,123)	21,855 (74.89%) (7,883-33,376)	15,016 (26.29%) (9,781-20,166)	124,197 (25.68%) (81,339-166,141)	102,665 (25.38%) (68,308-138,966)
GCAG motif	116 (2.74%) (47-150)	968 (2.82%) (401-1,284)	783 (2.68%) (296-1,030)	3,555 (6.22%) (1,798-4,841)	30,416 (6.29%) (15,325-41,588)	24,531 (6.06%) (12,325-33,491)
ATAC motif	5 (0.12%) (3-9)	34 (0.10%) (18-73)	31 (0.11%) (15-67)	139 (0.24%) (67-200)	1,206 (0.25%) (608-1,740)	1,015 (0.25%) (520-1,475)
noncanonical motif	890 (21.01%) (638-1,403)	7,410 (21.61%) (5,287-11,571)	6,518 (22.33%) (4,806-10,402)	38,404 (67.24%) (18,613-53,902)	330,492 (68.34%) (162,917-461,620)	276,959 (68.46%) (135,604-388,373)

Supplementary Table 3.5: Annotation and motif summary for JP junctions and 9-mers validated in BRCA.

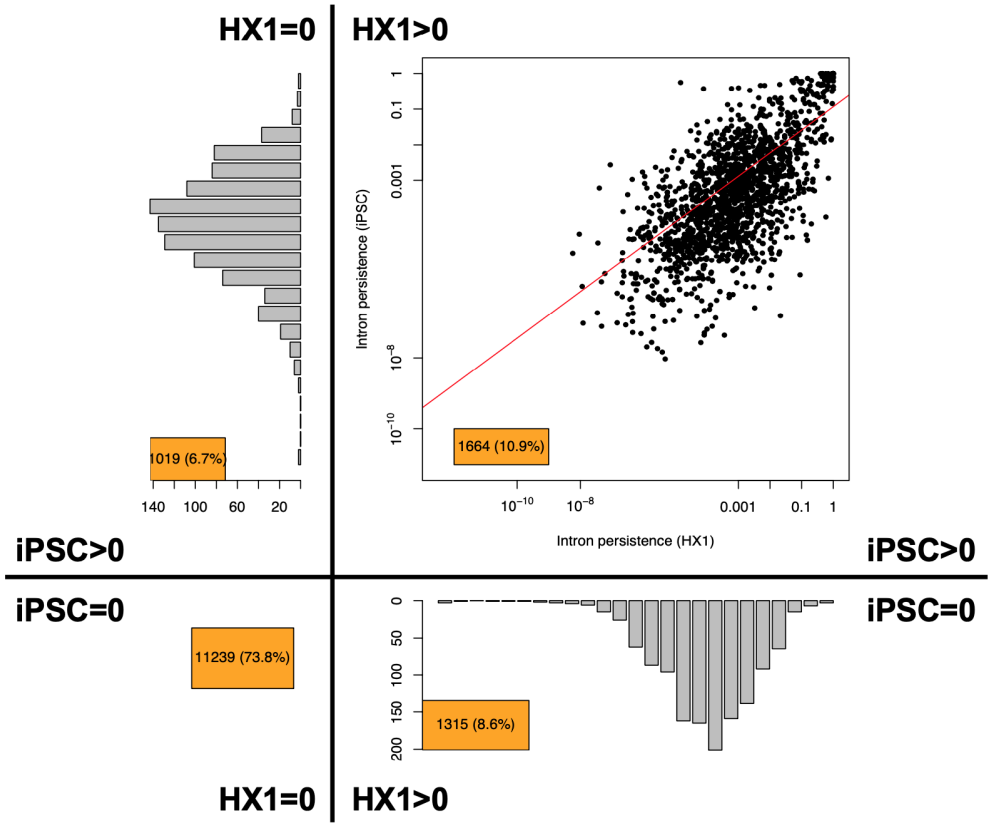
	BRCA		
JP validation summary	validated junctions: mean (min-max)	validated peptides: mean (min-max)	validated 9mers: mean (min-max)
total	70 (26-114)	105 (40-154)	509 (201-805)
annotated	2 (2.87%) (0-3)	3 (2.87%) (0-5)	12 (2.36%) (0-24)
exon skips	16 (22.99%) (3-40)	23 (21.99%) (5-52)	120 (23.58%) (24-286)
half-annotated	33 (47.41%) (9-50)	49 (46.85%) (14-78)	240 (47.15%) (81-354)
unannotated	19 (27.30%) (10-28)	29 (27.72%) (16-44)	138 (27.11%) (77-216)
GTAG motif	48 (68.97%) (17-83)	70 (66.92%) (25-106)	355 (69.74%) (140-589)
GCAG motif	4 (5.75%) (2-8)	7 (6.69%) (4-13)	28 (5.50%) (17-45)
ATAC motif	0 (0-0)	0 (0-0)	0 (0-0)
noncanonical motif	17 (24.43%) (6-25)	28 (26.77%) (8-42)	126 (24.75%) (44-205)



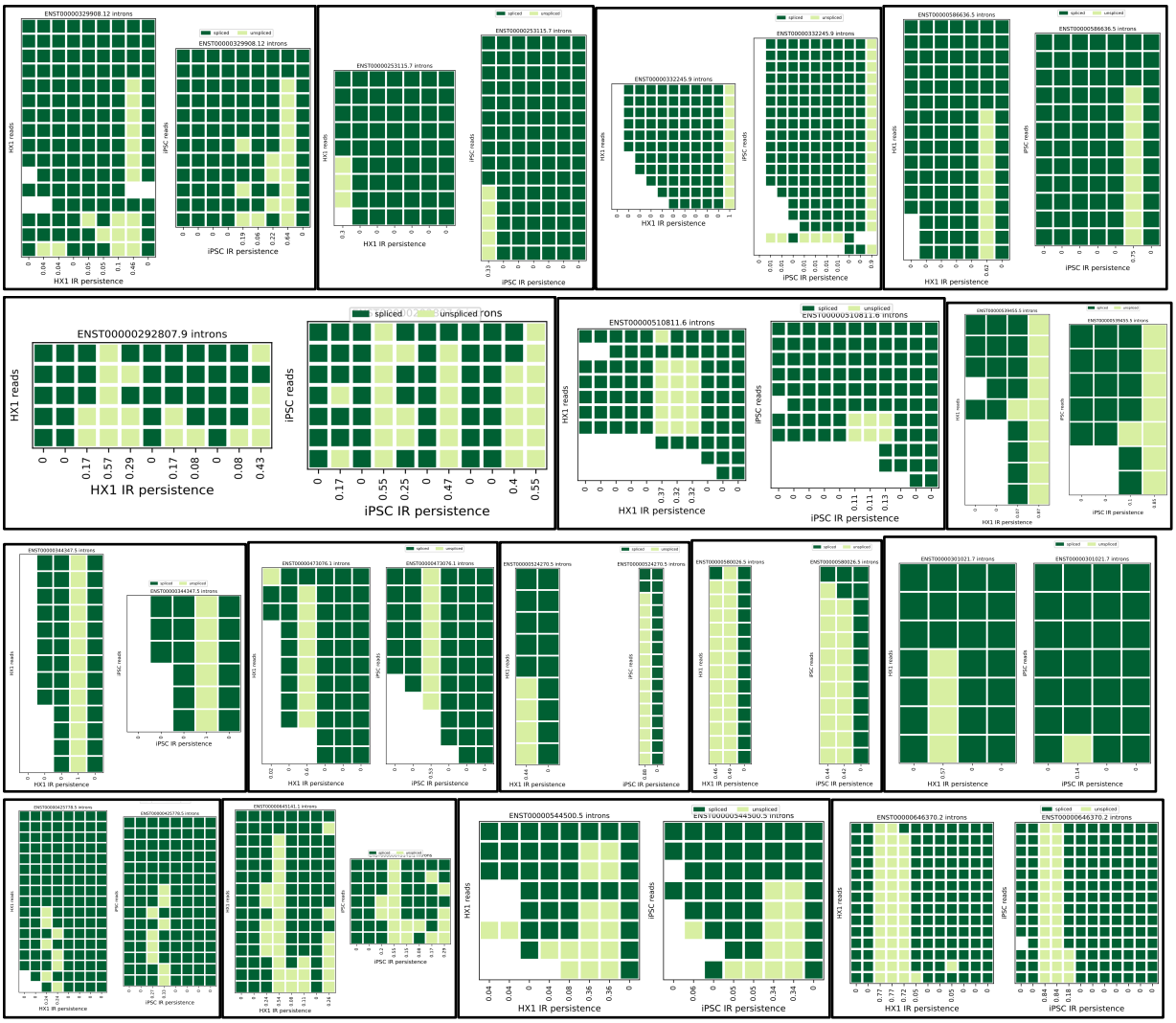
Supplementary Figure S4.1: Progression of transcription and splicing. Diagram depicts successive steps in transcript processing which progress from top to bottom. At the top is shown the presumed canonical transcript isoform with its expected splice pattern, followed by pre-mRNA processing steps, which branch between transcription by RNA polymerase II, co-transcriptional splicing (CTS), intron persistence, and poly(A) addition. At the bottom are the possible mature mRNA endpoints, including results from post-transcriptional mRNA splicing (PTS) and processing, which include translation and nonsense-mediated decay (NMD). Arrows are labeled with the events they represent, where arrow width sizes indicate their expected event frequencies Diagram created with BioRender.com.



Supplementary Figure S4.2: Short- and long-read coverage of genes by sample. The maximum number of long reads assigned to one transcript of each gene (y-axis) vs. the median short-read coverage per base across the entire gene (x-axis) for HX1 (red) and iPSC (blue) samples, in log scale. The vertical line represents the minimum median short-read coverage (2) and the horizontal line represents the minimum total long read coverage per transcript (5) required for a gene to be included in our analysis; genes considered are in the upper left segment of the plot.

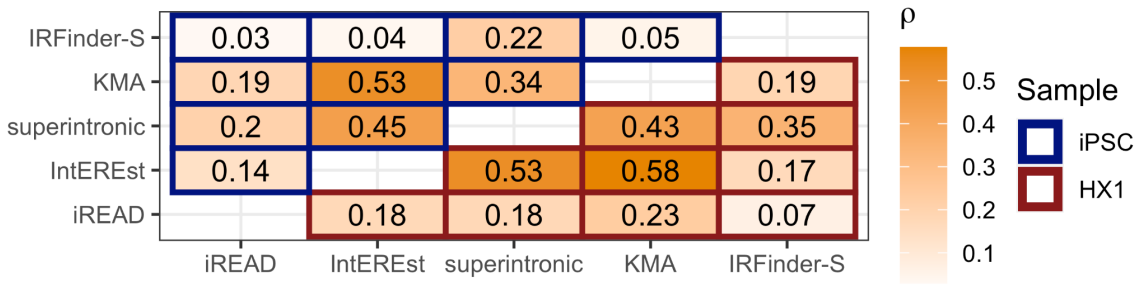


Supplementary Figure S4.3: Distribution of intron persistence values for introns in HX1 and iPSC samples. For introns included in both sample studies, bottom left quadrant represents introns with no persistence across both samples (73.8%), upper left represents introns with persistence in iPSC but not HX1 (6.7%), bottom right represents introns with persistence in HX1 but not iPSC (8.6%), and upper right is a scatterplot of persistences in iPSC (y-axis) vs. HX1 (x-axis) for introns with persistence in both (10.9%).

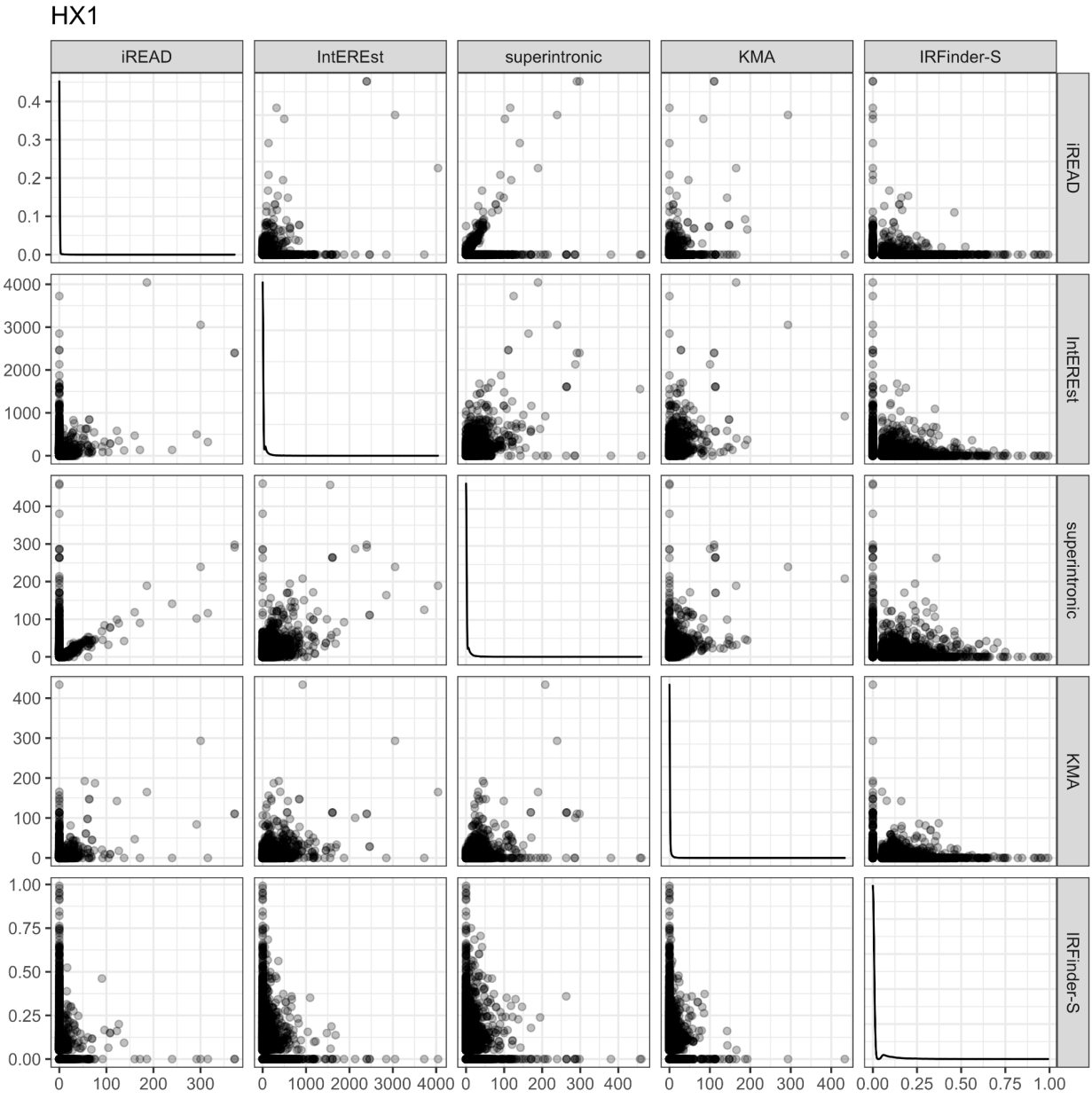


Supplementary Figure S4.4: Splicing similarity between samples. Heatmaps of splicing patterns for a selection of matched transcripts between HX1 and iPSC. Each subplot shows data for one transcript in HX1 (left) and iPSC (right) with rows representing transcript-matched long reads and columns representing transcript introns in 5'-->3' order. Transcripts were selected from a subset with 5–20 matched long reads in each sample and 5–20 introns. Dark green indicates a spliced out intron in a given read, light green indicates a retained intron, and white indicates no coverage of the intron in the read.

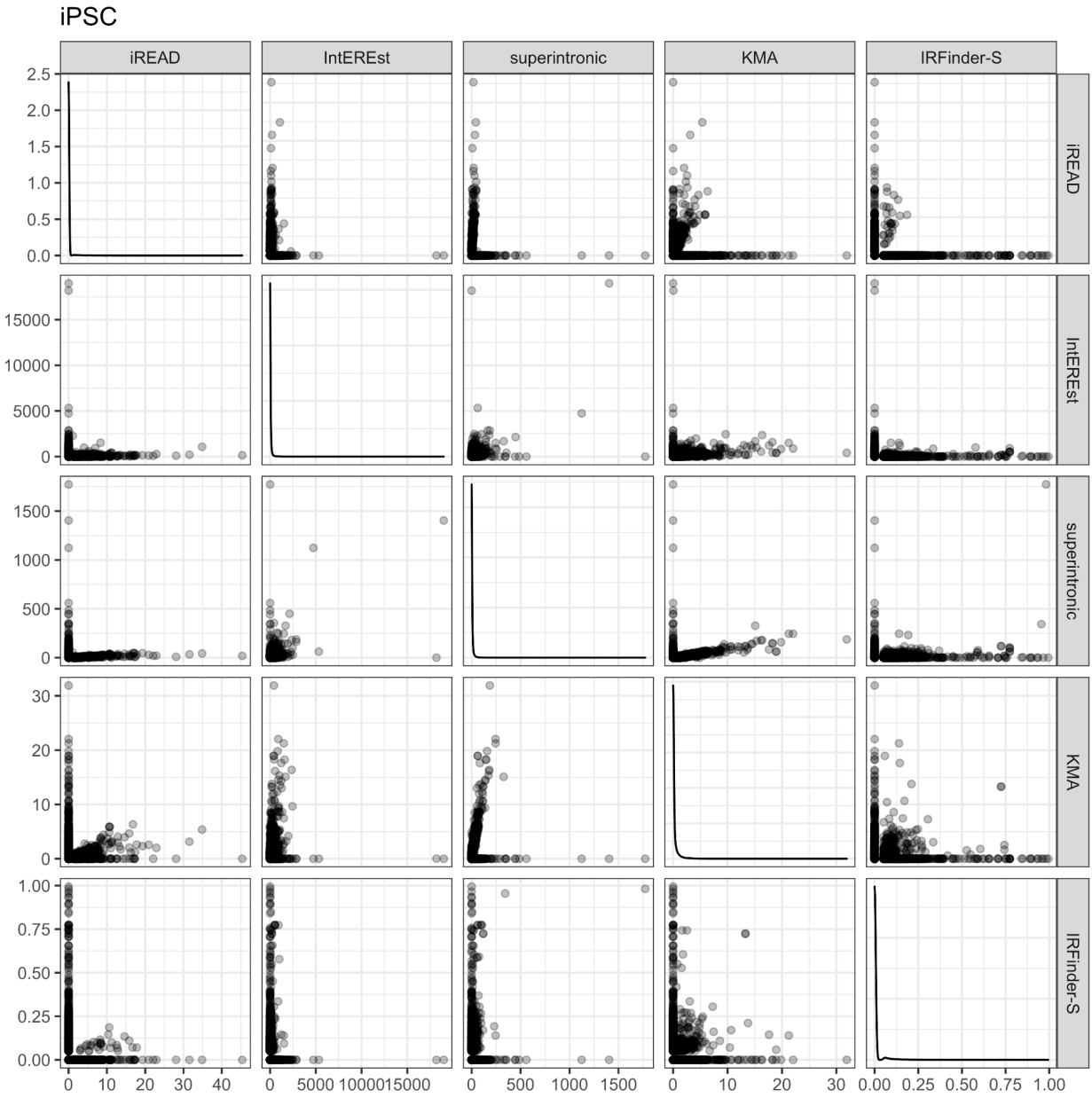
(A)


IRFinder-S
KMA
superintronic
IntEREst
iREAD
0.03
0.04
0.22
0.05
0.19
0.53
0.34
0.19
0.2
0.45
0.43
0.35
0.14
0.53
0.58
0.17
0.18
0.18
0.23
0.07
iREAD
IntEREst
superintronic
KMA
IRFinder-S
ρ
0.5
0.4
0.3
0.2
0.1
Sample
iPSC
HX1

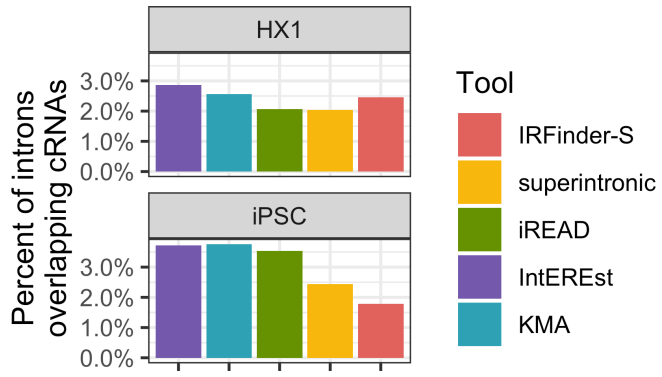
(B)


HX1
iREAD
IntEREst
superintronic
KMA
IRFinder-S
iREAD
IntEREst
superintronic
KMA
IRFinder-S

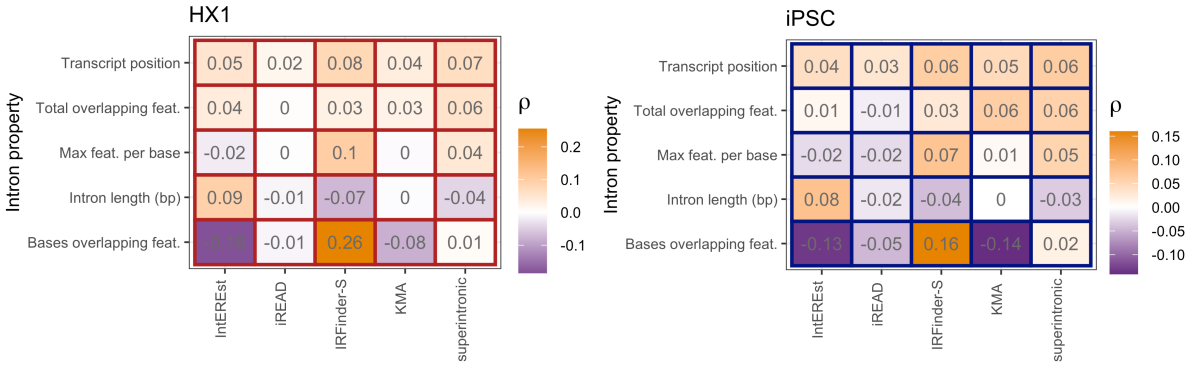
(C)



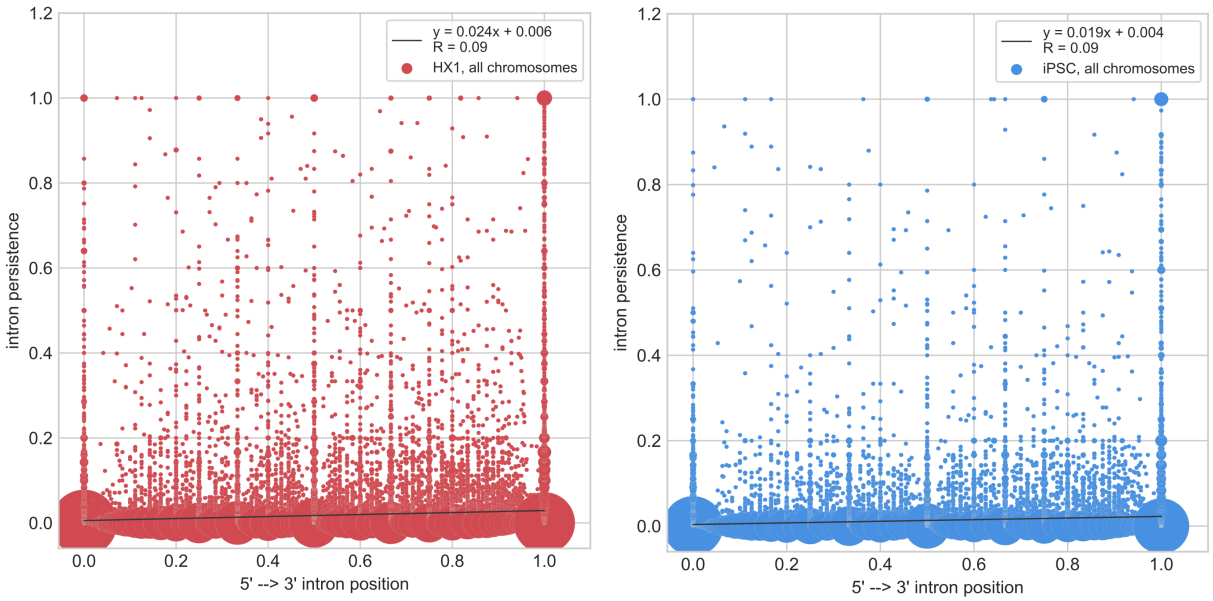
Supplementary Figure S4.5 Correlations between intron expression values output by short-read tools. (A) Pairwise correlations among the intron expression values output by five short-read tools. Each element in this heatmap depicts the correlation in intron expression values (Spearman's test) between the indicated pair of short-read tools, as labeled along the x- and y-axes. Cell text indicates Spearman ρ coefficient, with corresponding color value obtained by the color gradient scale shown (from white to orange). Cell outline color indicates the sample for which inter-tool correlation was assessed (iPSC [top left] and HX1 [bottom right] are outlined in blue and red, respectively). **(B)** Intron expression scatter plots between all SR IR-detection tool pairs (lower and upper triangles of plot grid) and density plots for each of the five individual tools (diagonal plot grid) for HX1. **(C)** Intron expression scatter plots between all SR IR-detection tool pairs (lower and upper triangles of plot grid) and density plots for each of the five individual tools (diagonal plot grid) for iPSC.



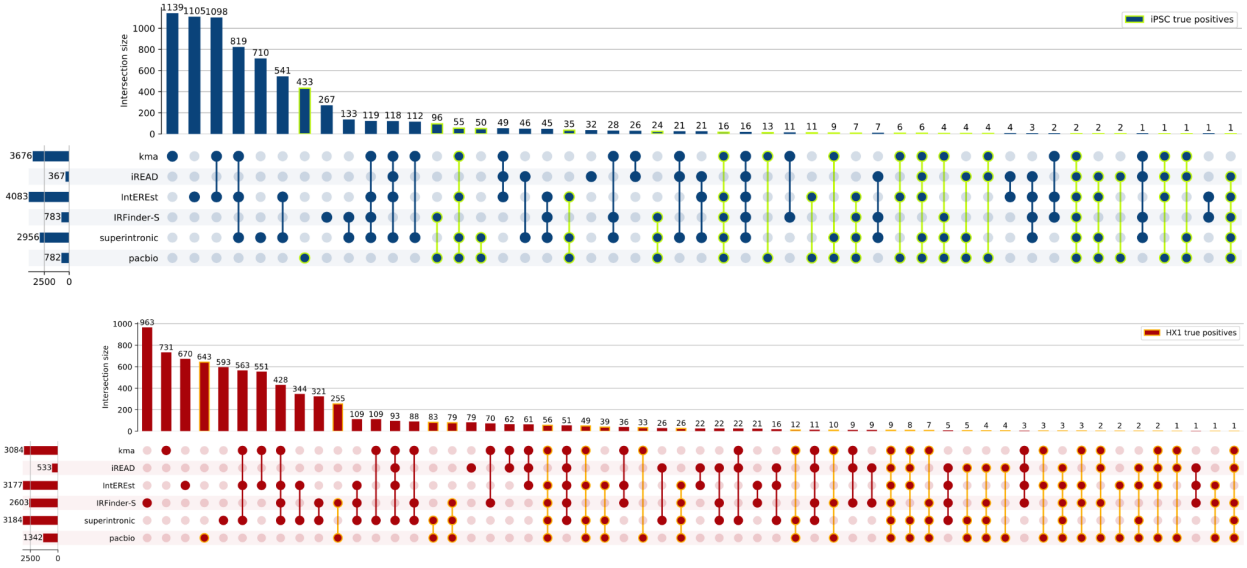
Supplementary Figure S4.6: cRNA overlap of called RIs. Barplots quantify the percent of called RIs (y-axes/bar heights) overlapping cRNAs are shown for 5 short-read tools (x-axes/bar colors, purple = IntEREst, blue = KMA, green = iREAD, yellow = superintronic, red = IRFinder-S), grouped by sample (rows/titles, top HX1, bottom iPSC).



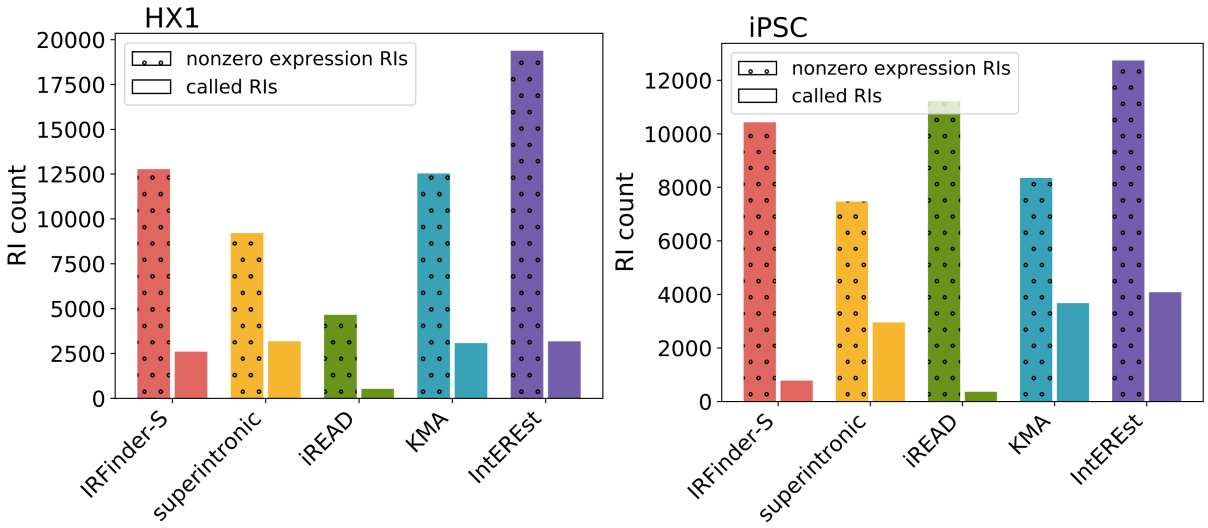
Supplementary Figure S4.7: Correlation of short-read expression with intron properties. Heatmap color fills and text show the Spearman ρ (purple = negative, white = near zero, orange = positive) between intron expression at five short-read tools (columns) and five continuous intron properties (rows), for samples HX1 (left heatmap) and iPSC (right heatmap).



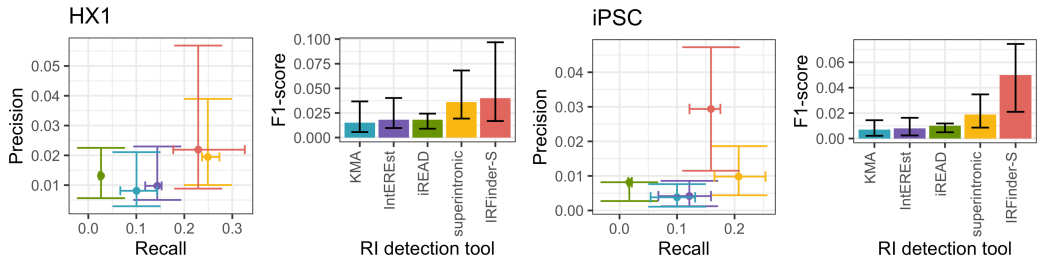
Supplementary Figure S4.8: Association of intron persistence with transcript position. Scatterplots of intron persistence vs. position within a transcript for HX1 (left, red), iPSC (right, blue). Each point represents one or more introns, with point size representing the number of points at each coordinate. Intron position is an intron-count normalized fraction where 0 represents the transcript's 5' end and 1 represents the 3' end. Plotted lines show the linear fit with equations shown in the inset legends.



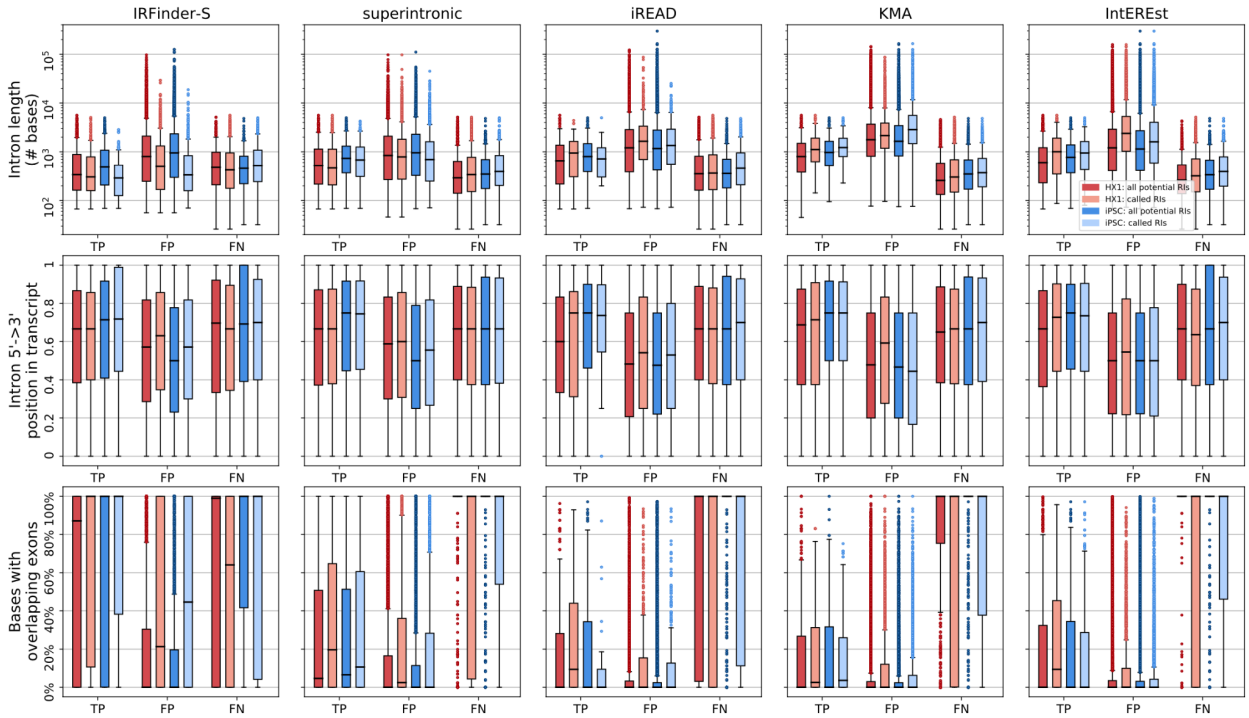
Supplementary Figure S4.9: Set overlaps of persistent introns and called RIs. Upset plots showing overlaps of sets of short-read called RIs and long read persistent introns for iPSC (above, blue) and HX1 (below, red). Sets of true positive persistent introns are highlighted in green for iPSC (above) and orange for HX1 (below).



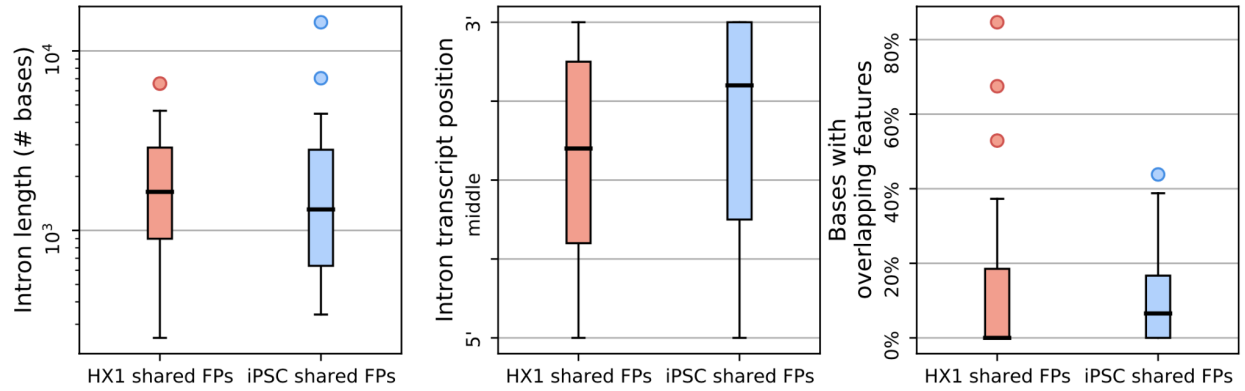
Supplementary Figure S4.10: Potential vs. called RI set sizes across short-read detection tools. For HX1 (left) and iPSC (right), counts of all potential (calculated nonzero expression) RIs (dotted hatch, left for each tool) and called (filtered) RIs (no hatch, right for each tool) for each SR detection tool (IRFinder-S, red; superintronic, yellow; iREAD, green; KMA, blue; IntEREst, purple).



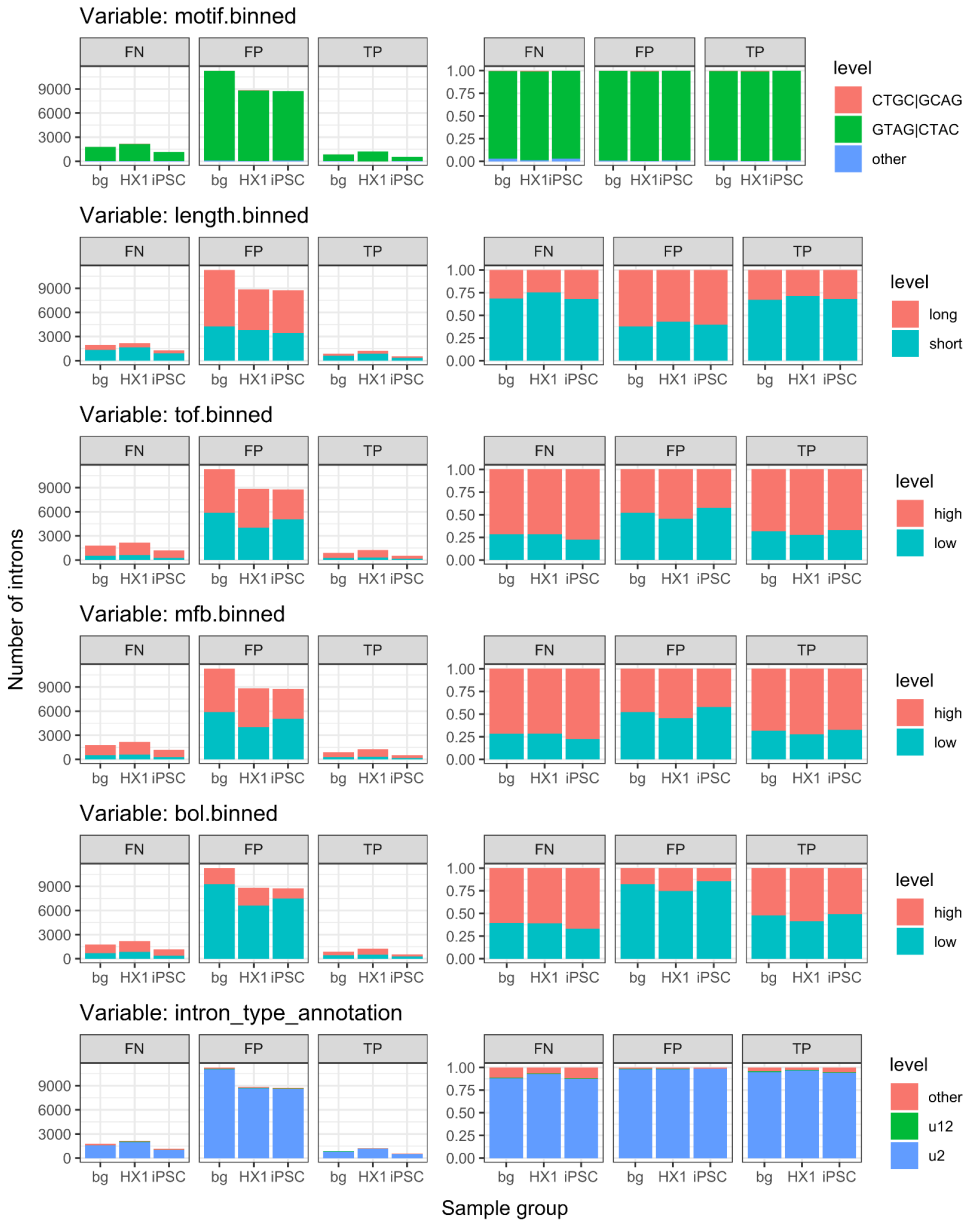
Supplementary Figure S4.11: Performance summaries across persistence cutoffs. Scatter plots of precision (y-axes) vs. recall (x-axes), and barplot of F1-scores (y-axes), for samples HX1 (left plots) and iPSC (right plots). Colors indicate short-read RI detection tools (red = IRFinder-S, yellow = superintronic, green = iREAD, purple = IntEREst, blue = KMA). Centroids and whiskers indicate the measure medians and interquartile ranges across persistence cutoffs varied from 0.1 to 0.9 at 0.1 intervals (Methods).



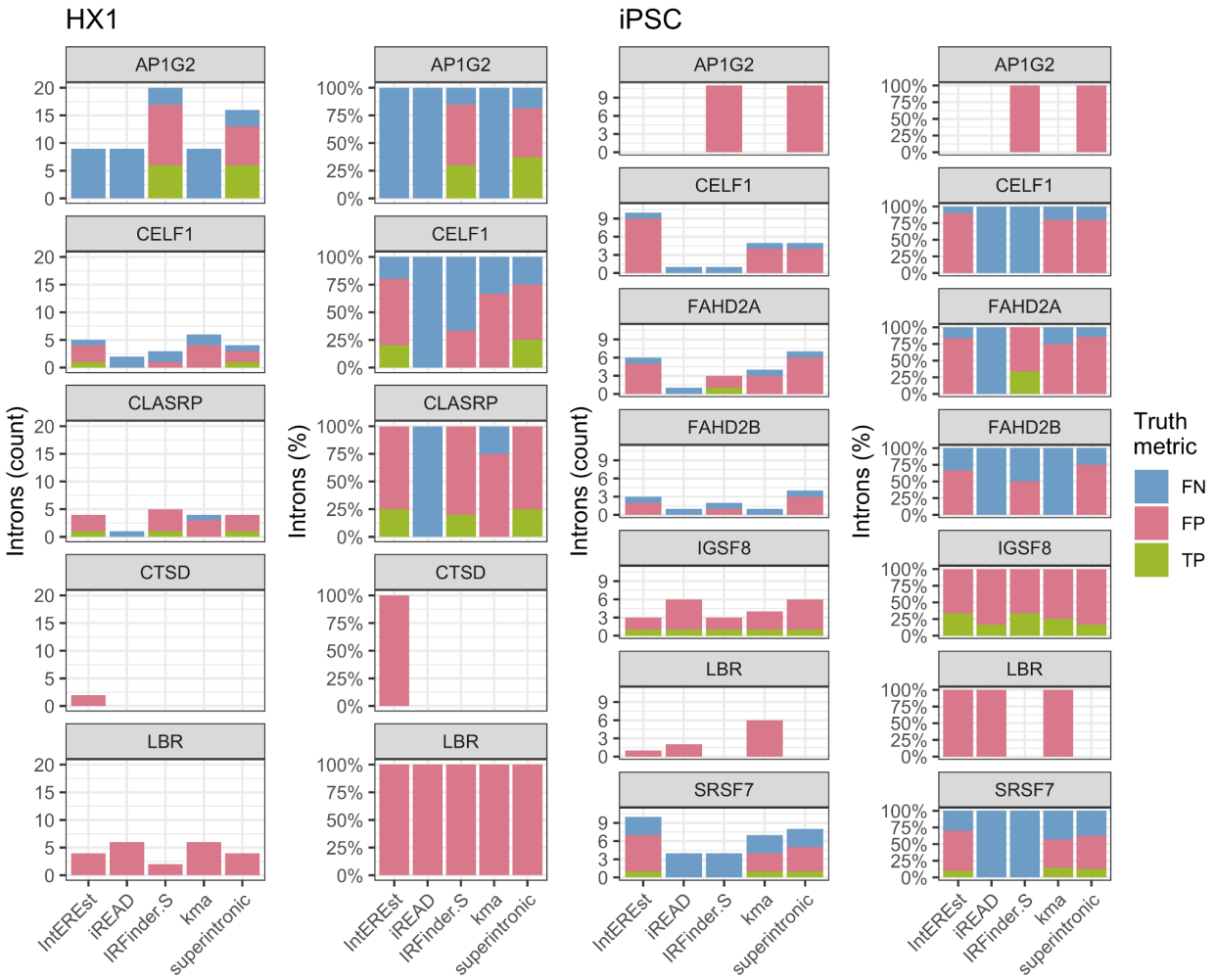
Supplementary Figure S4.12: Intron properties by truth category for potential called RIs, length (top), position along the direction of transcription (0 = 5', 1 = 3') and % of bases with an overlapping annotated exon vs. TP, FP, and FN calls for HX1 and iPSC via potential RIs (darker) and called RIs (lighter), at long read persistence of 0.1 for 5 short read tools.



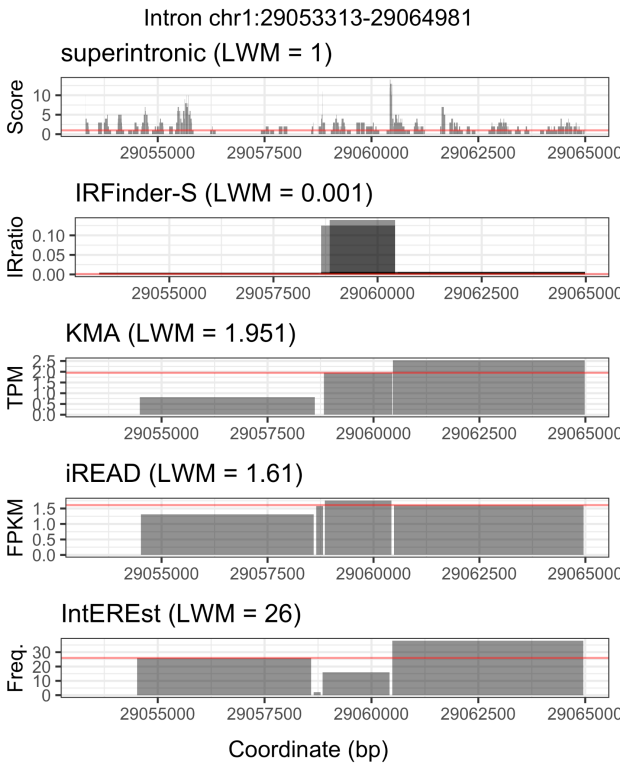
Supplementary Figure S4.13: Distribution of intron properties for shared false positive introns. (FPs) called across all five SR tools. Properties are, left to right, intron length in # of bases (log scale), transcript position, and % of bases with overlapping exons, for, in each panel, HX1 (left, red) and iPSC (right, blue).



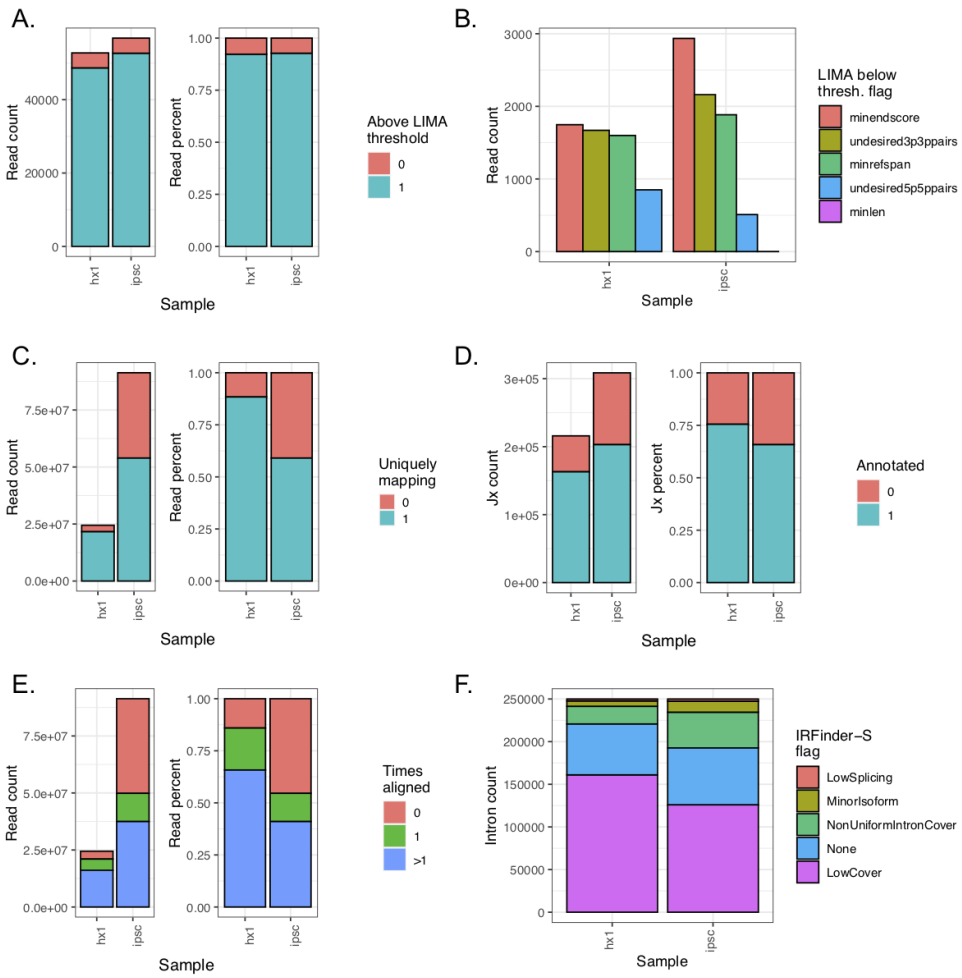
Supplementary Figure S4.14: Distributions of binned intron properties by truth category. Barplots of intron counts (left column) and percentages (right column) across unique levels (fill colors indicated in legends) for binned intron properties (plot titles). Results were binned by sample group types (columns, either HX1, iPSC, or the background of all unique introns) and intron 4+ truth metric categories TP, FP, and FN (ribbon labels, e.g. intron was TP in at least 4 tools for iPSC, etc.). Qualitative properties were binned by the top three most frequent levels (e.g. “intron type annotation” and “motif binned”), and quantitative properties were binned using the 50th quantile cutoff (e.g. “length”, total overlapping features or “tof,” max features per base or “mfb,” and bases overlapped or “bol”).



Supplementary Figure S4.15: Intron abundance by truth category across genes with validated RIs. Barplots show intron counts and percentages (y-axes) grouped by short-read tool (x-axes), gene (titles), for samples HX1 (left two plot columns) and iPSC (right two plot columns). Bar color fills indicate the short-read tool-specific truth category (green = TP, pink = FP, blue = FN).



Supplementary Figure S4.16: Example length-weighted median expression (LWM) at intron chr1:29053313-29064981. Intron expression (y-axes) is shown for genomic coordinates (x-axes), where expressed regions are represented by semi-transparent black rectangles which overlap the target intron. Results are grouped by each of the five short-read tools studied, and the LWM value calculated for each tool is shown in the plot titles and horizontal red lines.



Supplementary Figure S4.17: Processing and alignment quality control. Results (y-axes, fill colors) across long-read (A-B) and short-read (C-F) data runs for the samples HX1 and iPSC (x-axes) as follows. (a) LIMA quality among long-reads. Barplot y-axes quantify long-read counts (left) and percentages (right) relative to the quality threshold (blue = above, pink = below), where medians across all runs are shown for iPSC. (b) LIMA flags among long-reads. Barplot y-axes quantify long-reads, where bar colors and x-axes indicate one of the five quality flags (magenta = below minimum length or "minlen", blue = undesired 5-prime 5-prime pairs as "undesired5p5ppairs", green = below reference span as "minrefspan", yellow = undesired 3-prime 3-prime pairs as "undesired3p3ppairs", and pink = below minimum end score as "minendscore"). (c) Unique mapping among STAR-aligned short-reads. Barplot y-axes quantify short-read counts (left) and percentages (right) by mappability (blue/1 = uniquely mapping, pink/0 = not uniquely mapping). (d) Annotation among STAR-aligned short-reads. Barplot y- axes as in (c) with color indicating annotation (blue/1 = annotated, pink/0 = not annotated). (e) Alignment counts among bowtie2-aligned short-reads. Barplot y-axes as in (c), where bar colors show alignment counts (blue = > 1 times, green = 1 time, pink = 0 times). (f) IRFinder-S flag quantities. Barplot y-axes as in (c), where bar colors show flag (magenta = low coverage as "LowCover", blue = none, green = non-uniform intron coverage as "NonUniformIntronCover", yellow = minor isoform presence as "MinorIsoform", pink = low splicing as "LowSplicing").

Supplementary Table S4.1. Sample & run availability by platform.

Sample	iPSC	HX1
Type	Induced pluripotent stem cell line	Whole blood (non-cancer)
Biosample ID	SAMN07611993	SAMN04251426
SRA Study ID	SRP098984	SRP065930
Long read platform	PacBio Iso-Seq RSII	PacBio Iso-Seq RSII
Size fractionated	No	Yes
Iso-Seq runs	27	46
Aligned long reads	839,558	945,180
Short-read platform	Illumina NextSeq 500	Illumina HiSeq 2000
Short read runs	1	1
Aligned short reads (% uniquely aligned)	91,330,785 (59%)	24,463,210 (88%)

Supplementary Table S4.2 Performance metrics for called RIs across persistence thresholds.

sample	short read detection tool (RIs detected)		long read persistence threshold: >0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
HX1	long read RI count		4585	1342	675	458	334	249	181	135	106	72
	iREAD (533)	precision	0.317	0.077	0.032	0.023	0.019	0.013	0.009	0.006	0.002	0.002
		recall	0.037	0.031	0.025	0.026	0.030	0.028	0.028	0.022	0.009	0.014
		f-score	0.066	0.044	0.028	0.024	0.023	0.018	0.014	0.009	0.003	0.003
	IntEREst (3,177)	precision	0.265	0.064	0.037	0.023	0.016	0.010	0.008	0.005	0.003	0.002
		recall	0.184	0.151	0.172	0.162	0.153	0.124	0.144	0.119	0.094	0.083
		f-score	0.217	0.090	0.060	0.041	0.029	0.018	0.015	0.010	0.006	0.004
	superintronic (3,184)	precision	0.444	0.120	0.063	0.040	0.028	0.019	0.014	0.010	0.008	0.004
		recall	0.308	0.285	0.295	0.277	0.269	0.249	0.249	0.237	0.226	0.194
		f-score	0.364	0.169	0.103	0.070	0.051	0.036	0.027	0.019	0.015	0.009
	kma (3,084)	precision	0.258	0.064	0.034	0.021	0.014	0.008	0.006	0.003	0.002	0.001
		recall	0.174	0.146	0.154	0.144	0.126	0.100	0.099	0.067	0.066	0.056
		f-score	0.208	0.089	0.055	0.037	0.025	0.015	0.011	0.006	0.004	0.003
	IRFinder-S (2,603)	precision	0.509	0.175	0.086	0.057	0.036	0.022	0.012	0.009	0.007	0.003
		recall	0.289	0.340	0.330	0.323	0.281	0.229	0.177	0.170	0.179	0.125
		f-score	0.368	0.231	0.136	0.097	0.064	0.040	0.023	0.017	0.014	0.007
iPSC	long read RI count		3194	782	327	212	176	140	109	75	61	40
	iREAD (367)	precision	0.237	0.063	0.014	0.008	0.008	0.008	0.005	0.003	0.003	0.003
		recall	0.027	0.029	0.015	0.014	0.017	0.021	0.018	0.013	0.016	0.025
		f-score	0.049	0.040	0.014	0.010	0.011	0.012	0.008	0.005	0.005	0.005
	IntEREst (4,083)	precision	0.179	0.035	0.013	0.009	0.006	0.004	0.002	0.001	0.001	0.000
		recall	0.229	0.182	0.159	0.165	0.136	0.121	0.083	0.067	0.066	0.025
		f-score	0.201	0.058	0.024	0.016	0.011	0.008	0.004	0.002	0.002	0.000
	superintronic (2,956)	precision	0.344	0.073	0.028	0.019	0.013	0.010	0.006	0.004	0.003	0.002
		recall	0.318	0.276	0.254	0.259	0.222	0.207	0.165	0.173	0.148	0.125
		f-score	0.331	0.116	0.051	0.035	0.025	0.019	0.012	0.009	0.006	0.003
	kma (3,676)	precision	0.148	0.032	0.012	0.008	0.005	0.004	0.002	0.001	0.001	0.000
		recall	0.171	0.150	0.131	0.132	0.102	0.100	0.083	0.053	0.049	0.025
		f-score	0.159	0.052	0.021	0.014	0.009	0.007	0.005	0.002	0.002	0.001
	IRFinder-S (783)	precision	0.542	0.192	0.079	0.047	0.036	0.029	0.019	0.011	0.008	0.004
		recall	0.133	0.192	0.190	0.175	0.159	0.164	0.138	0.120	0.098	0.075
		f-score	0.213	0.192	0.112	0.074	0.058	0.050	0.034	0.021	0.014	0.007

* Green indicates the highest value per threshold and sample for each metric.

Supplementary Table S4.3A. Counts and performance metrics of called iPSC RIs.

thresh	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
longread_intron_count	3194	782	327	212	176	140	109	75	61	40
all_IRFinder-S_RIs	783	783	783	783	783	783	783	783	783	783
IRFinder-S_true_positives	424	150	62	37	28	23	15	9	6	3
IRFinder-S_false_positives	359	633	721	746	755	760	768	774	777	780
IRFinder-S_false_negatives	2770	632	265	175	148	117	94	66	55	37
IRFinder-S_precision	0.5415	0.1916	0.0792	0.0473	0.0358	0.0294	0.0192	0.0115	0.0077	0.0038
IRFinder-S_recall	0.1327	0.1918	0.1896	0.1745	0.1591	0.1643	0.1376	0.1200	0.0984	0.0750
IRFinder-S_fscore	0.2132	0.1917	0.1117	0.0744	0.0584	0.0498	0.0336	0.0210	0.0142	0.0073
all_superintronic_RIs	2956	2956	2956	2956	2956	2956	2956	2956	2956	2956
superintronic_true_positives	1017	216	83	55	39	29	18	13	9	5
superintronic_false_positives	1939	2740	2873	2901	2917	2927	2938	2943	2947	2951
superintronic_false_negatives	2177	566	244	157	137	111	91	62	52	35
superintronic_precision	0.3440	0.0731	0.0281	0.0186	0.0132	0.0098	0.0061	0.0044	0.0030	0.0017
superintronic_recall	0.3184	0.2762	0.2538	0.2594	0.2216	0.2071	0.1651	0.1733	0.1475	0.1250
superintronic_fscore	0.3307	0.1156	0.0506	0.0347	0.0249	0.0187	0.0117	0.0086	0.0060	0.0033
all_iREAD_RIs	367	367	367	367	367	367	367	367	367	367
iREAD_true_positives	87	23	5	3	3	3	2	1	1	1
iREAD_false_positives	280	344	362	364	364	364	365	366	366	366
iREAD_false_negatives	3107	759	322	209	173	137	107	74	60	39
iREAD_precision	0.2371	0.0627	0.0136	0.0082	0.0082	0.0082	0.0054	0.0027	0.0027	0.0027
iREAD_recall	0.0272	0.0294	0.0153	0.0142	0.0170	0.0214	0.0183	0.0133	0.0164	0.0250
iREAD_fscore	0.0489	0.0400	0.0144	0.0104	0.0110	0.0118	0.0084	0.0045	0.0047	0.0049
all_kma_RIs	3676	3676	3676	3676	3676	3676	3676	3676	3676	3676
kma_true_positives	545	117	43	28	18	14	9	4	3	1
kma_false_positives	3131	3559	3633	3648	3658	3662	3667	3672	3673	3675
kma_false_negatives	2649	665	284	184	158	126	100	71	58	39
kma_precision	0.1483	0.0318	0.0117	0.0076	0.0049	0.0038	0.0024	0.0011	0.0008	0.0003
kma_recall	0.1706	0.1496	0.1315	0.1321	0.1023	0.1000	0.0826	0.0533	0.0492	0.0250
KMA_fscore	0.1587	0.0525	0.0215	0.0144	0.0093	0.0073	0.0048	0.0021	0.0016	0.0005
all_IntEREst_RIs	4083	4083	4083	4083	4083	4083	4083	4083	4083	4083
IntEREst_true_positives	731	142	52	35	24	17	9	5	4	1
IntEREst_false_positives	3352	3941	4031	4048	4059	4066	4074	4078	4079	4082
IntEREst_false_negatives	2463	640	275	177	152	123	100	70	57	39
IntEREst_precision	0.1790	0.0348	0.0127	0.0086	0.0059	0.0042	0.0022	0.0012	0.0010	0.0002
IntEREst_recall	0.2289	0.1816	0.1590	0.1651	0.1364	0.1214	0.0826	0.0667	0.0656	0.0250
IntEREst_fscore	0.2009	0.0584	0.0236	0.0163	0.0113	0.0081	0.0043	0.0024	0.0019	0.0005

Supplementary Table S4.3B. Counts and performance metrics of called HX1 RIs.

thresh	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
longread_intron_count	4585	1342	675	458	334	249	181	135	106	72
all_IRFinder-S_RIs	2603	2603	2603	2603	2603	2603	2603	2603	2603	2603
IRFinder-S_true_positives	1324	456	223	148	94	57	32	23	19	9
IRFinder-S_false_positives	1279	2147	2380	2455	2509	2546	2571	2580	2584	2594
IRFinder-S_false_negatives	3261	886	452	310	240	192	149	112	87	63
IRFinder-S_precision	0.5086	0.1752	0.0857	0.0569	0.0361	0.0219	0.0123	0.0088	0.0073	0.0035
IRFinder-S_recall	0.2888	0.3398	0.3304	0.3231	0.2814	0.2289	0.1768	0.1704	0.1792	0.1250
IRFinder-S_fscore	0.3684	0.2312	0.1361	0.0967	0.0640	0.0400	0.0230	0.0168	0.0140	0.0067
all_superintronic_RIs	3184	3184	3184	3184	3184	3184	3184	3184	3184	3184
superintronic_true_positives	1413	383	199	127	90	62	45	32	24	14
superintronic_false_positives	1771	2801	2985	3057	3094	3122	3139	3152	3160	3170
superintronic_false_negatives	3172	959	476	331	244	187	136	103	82	58
superintronic_precision	0.4438	0.1203	0.0625	0.0399	0.0283	0.0195	0.0141	0.0101	0.0075	0.0044
superintronic_recall	0.3082	0.2854	0.2948	0.2773	0.2695	0.2490	0.2486	0.2370	0.2264	0.1944
superintronic_fscore	0.3638	0.1692	0.1031	0.0697	0.0512	0.0361	0.0267	0.0193	0.0146	0.0086
all_iREAD_RIs	533	533	533	533	533	533	533	533	533	533
iREAD_true_positives	169	41	17	12	10	7	5	3	1	1
iREAD_false_positives	364	492	516	521	523	526	528	530	532	532
iREAD_false_negatives	4416	1301	658	446	324	242	176	132	105	71
iREAD_precision	0.3171	0.0769	0.0319	0.0225	0.0188	0.0131	0.0094	0.0056	0.0019	0.0019
iREAD_recall	0.0369	0.0306	0.0252	0.0262	0.0299	0.0281	0.0276	0.0222	0.0094	0.0139
iREAD_fscore	0.0660	0.0437	0.0281	0.0242	0.0231	0.0179	0.0140	0.0090	0.0031	0.0033
all_kma_RIs	3084	3084	3084	3084	3084	3084	3084	3084	3084	3084
kma_true_positives	796	196	104	66	42	25	18	9	7	4
kma_false_positives	2288	2888	2980	3018	3042	3059	3066	3075	3077	3080
kma_false_negatives	3789	1146	571	392	292	224	163	126	99	68
kma_precision	0.2581	0.0636	0.0337	0.0214	0.0136	0.0081	0.0058	0.0029	0.0023	0.0013
kma_recall	0.1736	0.1461	0.1541	0.1441	0.1257	0.1004	0.0994	0.0667	0.0660	0.0556
KMA_fscore	0.2076	0.0886	0.0553	0.0373	0.0246	0.0150	0.0110	0.0056	0.0044	0.0025
all_IntEREst_RIs	3177	3177	3177	3177	3177	3177	3177	3177	3177	3177
IntEREst_true_positives	843	203	116	74	51	31	26	16	10	6
IntEREst_false_positives	2334	2974	3061	3103	3126	3146	3151	3161	3167	3171
IntEREst_false_negatives	3742	1139	559	384	283	218	155	119	96	66
IntEREst_precision	0.2653	0.0639	0.0365	0.0233	0.0161	0.0098	0.0082	0.0050	0.0031	0.0019
IntEREst_recall	0.1839	0.1513	0.1719	0.1616	0.1527	0.1245	0.1436	0.1185	0.0943	0.0833
IntEREst_fscore	0.2172	0.0898	0.0602	0.0407	0.0291	0.0181	0.0155	0.0097	0.0061	0.0037

Supplementary Table S4.3C. Counts and performance metrics of all potential iPSC RIs.

thresh	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
longread_intron_count	3194	782	327	212	176	140	109	75	61	40
all_IRFinder-S_RIs	10434	10434	10434	10434	10434	10434	10434	10434	10434	10434
IRFinder-S_true_positives	2147	501	199	122	99	77	59	40	33	18
IRFinder-S_false_positives	8287	9933	10235	10312	10335	10357	10375	10394	10401	10416
IRFinder-S_false_negatives	1047	281	128	90	77	63	50	35	28	22
IRFinder-S_precision	0.2058	0.0480	0.0191	0.0117	0.0095	0.0074	0.0057	0.0038	0.0032	0.0017
IRFinder-S_recall	0.6722	0.6407	0.6086	0.5755	0.5625	0.5500	0.5413	0.5333	0.5410	0.4500
IRFinder-S_fscore	0.3151	0.0893	0.0370	0.0229	0.0187	0.0146	0.0112	0.0076	0.0063	0.0034
all_superintronic_RIs	7465	7465	7465	7465	7465	7465	7465	7465	7465	7465
superintronic_true_positives	1499	296	99	60	43	32	20	13	9	5
superintronic_false_positives	5966	7169	7366	7405	7422	7433	7445	7452	7456	7460
superintronic_false_negatives	1695	486	228	152	133	108	89	62	52	35
superintronic_precision	0.2008	0.0397	0.0133	0.0080	0.0058	0.0043	0.0027	0.0017	0.0012	0.0007
superintronic_recall	0.4693	0.3785	0.3028	0.2830	0.2443	0.2286	0.1835	0.1733	0.1475	0.1250
superintronic_fscore	0.2813	0.0718	0.0254	0.0156	0.0113	0.0084	0.0053	0.0034	0.0024	0.0013
all_iREAD_RIs	11225	11225	11225	11225	11225	11225	11225	11225	11225	11225
iREAD_true_positives	1373	250	74	44	33	22	13	7	6	2
iREAD_false_positives	9852	10975	11151	11181	11192	11203	11212	11218	11219	11223
iREAD_false_negatives	1821	532	253	168	143	118	96	68	55	38
iREAD_precision	0.1223	0.0223	0.0066	0.0039	0.0029	0.0020	0.0012	0.0006	0.0005	0.0002
iREAD_recall	0.4299	0.3197	0.2263	0.2075	0.1875	0.1571	0.1193	0.0933	0.0984	0.0500
iREAD_fscore	0.1904	0.0416	0.0128	0.0077	0.0058	0.0039	0.0023	0.0012	0.0011	0.0004
all_kma_RIs	8352	8352	8352	8352	8352	8352	8352	8352	8352	8352
kma_true_positives	1023	202	66	39	28	22	13	6	5	2
kma_false_positives	7329	8150	8286	8313	8324	8330	8339	8346	8347	8350
kma_false_negatives	2171	580	261	173	148	118	96	69	56	38
kma_precision	0.1225	0.0242	0.0079	0.0047	0.0034	0.0026	0.0016	0.0007	0.0006	0.0002
kma_recall	0.3203	0.2583	0.2018	0.1840	0.1591	0.1571	0.1193	0.0800	0.0820	0.0500
KMA_fscore	0.1772	0.0442	0.0152	0.0091	0.0066	0.0052	0.0031	0.0014	0.0012	0.0005
all_IntEREst_RIs	12743	12743	12743	12743	12743	12743	12743	12743	12743	12743
IntEREst_true_positives	1608	301	84	48	34	22	13	7	6	2
IntEREst_false_positives	11135	12442	12659	12695	12709	12721	12730	12736	12737	12741
IntEREst_false_negatives	1586	481	243	164	142	118	96	68	55	38
IntEREst_precision	0.1262	0.0236	0.0066	0.0038	0.0027	0.0017	0.0010	0.0005	0.0005	0.0002
IntEREst_recall	0.5034	0.3849	0.2569	0.2264	0.1932	0.1571	0.1193	0.0933	0.0984	0.0500
IntEREst_fscore	0.2018	0.0445	0.0129	0.0074	0.0053	0.0034	0.0020	0.0011	0.0009	0.0003

Supplementary Table S4.3D. Counts and performance metrics of all potential HX1 RIs.

thresh	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
longread_intron_count	4585	1342	675	458	334	249	181	135	106	72
all_IRFinder-S_RIs	12772	12772	12772	12772	12772	12772	12772	12772	12772	12772
IRFinder-S_true_positives	3479	967	468	312	219	153	99	72	58	36
IRFinder-S_false_positives	9293	11805	12304	12460	12553	12619	12673	12700	12714	12736
IRFinder-S_false_negatives	1106	375	207	146	115	96	82	63	48	36
IRFinder-S_precision	0.2724	0.0757	0.0366	0.0244	0.0171	0.0120	0.0078	0.0056	0.0045	0.0028
IRFinder-S_recall	0.7588	0.7206	0.6933	0.6812	0.6557	0.6145	0.5470	0.5333	0.5472	0.5000
IRFinder-S_fscore	0.4009	0.1370	0.0696	0.0472	0.0334	0.0235	0.0153	0.0112	0.0090	0.0056
all_superintronic_RIs	9208	9208	9208	9208	9208	9208	9208	9208	9208	9208
superintronic_true_positives	2536	613	282	178	124	83	57	42	33	21
superintronic_false_positives	6672	8595	8926	9030	9084	9125	9151	9166	9175	9187
superintronic_false_negatives	2049	729	393	280	210	166	124	93	73	51
superintronic_precision	0.2754	0.0666	0.0306	0.0193	0.0135	0.0090	0.0062	0.0046	0.0036	0.0023
superintronic_recall	0.5531	0.4568	0.4178	0.3886	0.3713	0.3333	0.3149	0.3111	0.3113	0.2917
superintronic_fscore	0.3677	0.1162	0.0571	0.0368	0.0260	0.0176	0.0121	0.0090	0.0071	0.0045
all_iREAD_RIs	4657	4657	4657	4657	4657	4657	4657	4657	4657	4657
iREAD_true_positives	736	179	82	48	26	18	12	4	2	1
iREAD_false_positives	3921	4478	4575	4609	4631	4639	4645	4653	4655	4656
iREAD_false_negatives	3849	1163	593	410	308	231	169	131	104	71
iREAD_precision	0.1580	0.0384	0.0176	0.0103	0.0056	0.0039	0.0026	0.0009	0.0004	0.0002
iREAD_recall	0.1605	0.1334	0.1215	0.1048	0.0778	0.0723	0.0663	0.0296	0.0189	0.0139
iREAD_fscore	0.1593	0.0597	0.0308	0.0188	0.0104	0.0073	0.0050	0.0017	0.0008	0.0004
all_kma_RIs	12533	12533	12533	12533	12533	12533	12533	12533	12533	12533
kma_true_positives	1779	430	191	117	77	48	32	22	18	10
kma_false_positives	10754	12103	12342	12416	12456	12485	12501	12511	12515	12523
kma_false_negatives	2806	912	484	341	257	201	149	113	88	62
kma_precision	0.1419	0.0343	0.0152	0.0093	0.0061	0.0038	0.0026	0.0018	0.0014	0.0008
kma_recall	0.3880	0.3204	0.2830	0.2555	0.2305	0.1928	0.1768	0.1630	0.1698	0.1389
KMA_fscore	0.2079	0.0620	0.0289	0.0180	0.0120	0.0075	0.0050	0.0035	0.0028	0.0016
all_IntEREst_RIs	19384	19384	19384	19384	19384	19384	19384	19384	19384	19384
IntEREst_true_positives	2876	670	291	176	115	74	51	35	28	17
IntEREst_false_positives	16508	18714	19093	19208	19269	19310	19333	19349	19356	19367
IntEREst_false_negatives	1709	672	384	282	219	175	130	100	78	55
IntEREst_precision	0.1484	0.0346	0.0150	0.0091	0.0059	0.0038	0.0026	0.0018	0.0014	0.0009
IntEREst_recall	0.6273	0.4993	0.4311	0.3843	0.3443	0.2972	0.2818	0.2593	0.2642	0.2361
IntEREst_fscore	0.2400	0.0647	0.0290	0.0177	0.0117	0.0075	0.0052	0.0036	0.0029	0.0017

Supplementary Table S4.4 Properties and sources of experimentally validated RIs studied.

Gene name	Source	Intron coordinates	Discovery assay	Validation assay	Disease or cell type association	Coverage in samples	Sample intron persistence
AP1G2	Jeong 2021[279]	chr14:23565702-23565815 (intron 5)	short-read RNA-seq	RT-PCR	mesenchymal stem cell	HX1, iPSC	0.06, 0
CELF1	Li 2021[281]	chr11:47478953-47482694 chr11:47478941-47482694	short-read RNA-seq	Nanostring	Alzheimer's disease	HX1, iPSC	0, 0
CLASRP	Li 2021[281]	chr19:45069249-45070021	short-read RNA-seq	Nanostring	Alzheimer's disease	HX1	0
CTSD	Wong 2013[83]	chr11:1755029-1757323 (intron 5)	short-read RNA-seq	RT-PCR, RNA-seq	granulocyte	HX1	0
FAHD2A	Li 2021[281]	chr2:95412765-95412894	short-read RNA-seq	Nanostring	Alzheimer's disease	iPSC	0.06
FAHD2B	Li 2021[281]	chr2:97083818-97083947	short-read RNA-seq	Nanostring	Alzheimer's disease	iPSC	0.05
IGSF8	Li 2021[281]	chr1:160094172-160094868	short-read RNA-seq	Nanostring	Alzheimer's disease	iPSC	0.02
LBR	Wong 2013[83]	chr1:225410417-225411336 (intron 9)	short-read RNA-seq	RT-PCR, RNA-seq	granulocyte	HX1, iPSC	0.02, 0.01
SRSF7	Lejeune 2001[280]	chr2:38748654-38749528 (intron 3)	*in vitro* splicing assays	Northern blot	--	iPSC	0.17