Understanding recursive splicing in the human genome

Emma Hoppe

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Robert K. Bradley, Chair

Stanley Fields
Steven Henikoff

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Understanding recursive splicing in the human genome

Emma Hoppe

Chair of the Supervisory Committee:
Professor Robert K. Bradley
Basic Sciences & Public Health Sciences Division
Fred Hutchinson Cancer Center

Recursive splicing is a non-canonical splicing mechanism that results in an intron being removed in two or more segments. Identifying recursive splicing presents technical challenges due to the lack of evidence in mRNA and the instability of splicing intermediates and byproducts. Few recursive splice sites have been identified with high confidence in human introns and largely have been located within long introns. Using a stringent approach to map lariat reads, I identified recursive splicing genome-wide, finding evidence for 100 new recursive splice sites in a broader range of intron sizes than previously reported and characterizing a new location for recursive splicing at the distal end of cassette exons. These data demonstrate the unappreciated prevalence of recursive splicing, the potential for finding additional sites as appropriately enriched RNA

sequencing datasets become available, and its possible influence on gene expression through alternative exon isoforms. In addition, I discuss the evidence for intronic and distal exonic recursive splice sites as a mechanism of exon birth.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To my thesis committee, Alice Berger, Stan Fields, Steve Henikoff, and Edith Wang: thank you for your gift of time, advice, and thoughtful questions throughout the years. I greatly appreciate your guidance in both my project and career plans and your understanding as I went through my project's and health's ups and downs. Thank you to Stan Fields and Steve Henikoff for additionally serving on my reading committee.

Thank you to the current and former members of the Bradley Lab. To my mentor Rob Bradley, thank you for your support and mentorship during my time in your lab. I've greatly appreciated your feedback as I grew as a scientist, and feel very blessed to have been a part of the supportive lab environment you've fostered. Thank you to Joey Pangallo and Khrystyna North for helping me find my feet in the lab and troubleshoot my initial project in the lab. Thank you to GuoLiang "Chewie" Chew and Jose Pineda for your help as I made the transition to my computational project. Jose, thank you also for always being a sounding board for my project and life! James Thomas, it was great fun working on the poison exon project paper and sharing a bay with you for a time—you were right Christmas lights really are the perfect light to code by. Dylan Udy, I'm immensely grateful for your support in getting this paper finished up and for your willingness to help with random things throughout the years—you saved the day with numerous Girls Who Code events. Thanks also to Emma De Neef, Austin Gabel, Andrea Belleville, Siegen McKellar, Jake Polaski and Taylor Nicholas for the great conversations about science and life throughout the year.

My most heartfelt thanks to all of my excellent teachers and mentors throughout my life who nurtured my love for learning. In particular, thank you to the excellent teachers and staff of Zoo School, especially Jim Barstow and Sara LeRoy-Toren--you both were instrumental in teaching me to research a topic deeply and consider it from all sides. I greatly appreciated all the discussions and time spent with us outside of class. Thank you to my scientific mentors throughout the years: Dr. Guodong Ren and Dr. Bin Yu; Dr. Jerry Bricker; Dr. Mallory Tuner and Dr. Michael Barry; Dr. Gaurav Pandey; and Dr. SooChin Cho and Dr. Mark Reedy. I greatly appreciated the opportunities and support while I was in your labs and afterward. Thank you also to Dr. Reedy for all of your excellent career advice.

To my incredible family, I couldn't have done this without you. In particular, my parents, Mary & Jim Hoppe, thank you for your unwavering support and understanding. Thanks for easing my time in grad school with all the encouraging words and thoughtful gifts. Thanks for kindling a love of learning in me by taking me to bookstores, museums, and Brightlights camps. Thank you also to my extended Schroeder and Hoppe families—I'm blessed to have your love, support, and inspiration.

Chapter 1. INTRODUCTION

1.1 SPLICE SITES & THE SPLICEOSOME

In metazoans, most nascent transcripts produced by RNA polymerase II contain sequence segments called introns that need to be removed through a process called splicing to produce functional mature RNAs (mRNAs) or long non-coding RNAs (lncRNAs). The segments that remain in the final transcript are called exons. First uncovered in the late 1970s, this concept of split genes is fundamental to our understanding of gene expression and enables the complexity of the higher eukaryotes through alternative splicing (Berget et al., 1977; Chow et al., 1977; reviewed in Roy et al., 2013).

Splicing occurs as two transesterification reactions catalyzed by a ribonucleoprotein (RNP) machine called the spliceosome (Lerner et al., 1980). How the spliceosome recognizes intronic and exonic segments is complex and will be discussed in more detail later but the basic sequence requirements are relatively simple: The upstream exon is bookended by a 5' splice site (5'ss) and the downstream exon is preceded by a branchpoint, polypyrimidine (polyY) tract, and 3' splice site (3'ss). The first splicing reaction cleaves the phosphodiester bond between the upstream exon and the 5' ss and the intron while forming a 2'-5' phosphodiester linkage between the 5' guanosine and the branchpoint nucleotide, typically adenosine (Fig 1.1, panels B* and C). The second step cleaves the phosphodiester bond between the 3'ss and the downstream exon while the two exons are ligated together, releasing the intron as a lariat (Fig 1.1, panels C* and P).

The majority of introns (U2-type) are spliced by what's called the major spliceosome, which is composed of five small nuclear RNAs (U1, U2, U4, U5, and U6). The minimal consensus

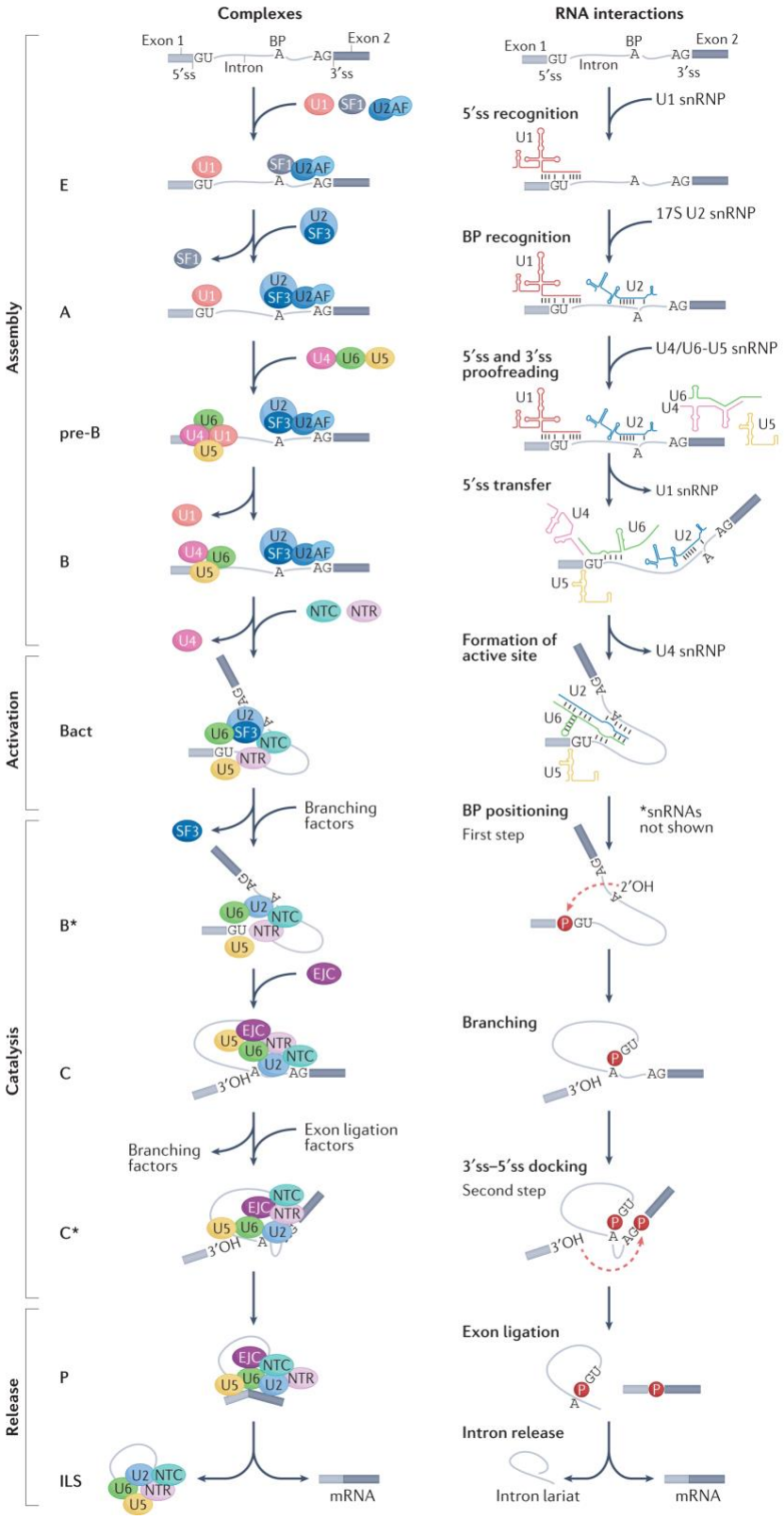

Figure 1.1. Schematic of splicing, adapted from Rogalska, Vivori, and Valcarel (2022). Left column: spliceosomal assembly. Right column: RNA-RNA interactions.

splice sites it recognizes are GU for the 5’ss and AG for the 3’ss. The minor spliceosome accounts for less than 0.5% of introns in most metazoan genomes and largely shares the same 5' and 3' consensus sequences; however, it also splices sites with the non-consensus sequences 5' AT and 3' AC (reviewed in Turunen et al., 2013). As the major spliceosome and U2-type intron sequence make up the vast majority of splice sites in metazoan genomes, I will be focusing on them in the following discussion of splice site selection.

The U1 small nuclear RNP (snRNP) is responsible for recognizing the 5'ss of the intron in the first step of spliceosome assembly through base-pairing interactions with the 5' end of its U1 small nuclear RNA (snRNA) (Fig 1.1, panel pre-B). Therefore, the U1 snRNA sequence contributes heavily to the 5'ss consensus sequence: 5'-AG|GURAGU-3' (Mount et al., 1983). However, most 5'ss do not form perfect base-pairing interactions with the 5' end of the U1 snRNA (reviewed in Roca et al., 2013). Likewise, a massively parallel splicing assay assessing all possible perturbations of the consensus sequence revealed that gene context plays a substantial role as well, suggesting that nearby sequences and associated factors interacting with U1 contribute to 5'ss selection (Wong et al., 2018). Still, adherence to the consensus motif primarily governs splicing efficiency for 5’ss (Wong et al., 2018).

In mammals, the 3'ss is comprised of two parts: a highly conserved YAG located immediately upstream of the exon (position -1 to -3) and the polyY tract (position -5 to -10 or more). Introns with weak or short (~5-10 pyrimidines) polyY tracts require a 3' AG, whereas introns with long polyY tracts may not require an AG (reviewed in Moore, 2000). Recognition of the 3' splice site is due to the cooperative binding of splicing factor 1 (SF1), which binds to the branchpoint, and the U2AF heterodimer of U2AF1, which binds to the polyY tract, and U2AF2, which binds to the YAG portion of the 3'ss (Fig 1.1, panel E). SF1 is then replaced by the U2

snRNP, which includes SF3B1, (Fig 1.1, panel E to A). There are known distance constraints between the branchpoint and the 3' splice site, with a typical range of ~10-50 nt and a median of 25 nt for typical U2-type introns (Pineda, Nicholas, and Bradley, under review). A SELEX experiment performed with the human U2AF heterodimer produced the sequence 5'-UUUYYYYUNYAG|GU-3' (Wu et al., 1999), though the preference for the exonic portion to be GU is much less pronounced when looking at all cassette or constitutive exons (see **Chapter 3,** Fig. 3.1 for sequence logo plots)**,** perhaps because of coding sequence constraints.

After the U1 and U2 for an intron have been paired by the spliceosome (see section "Intron and Exon Definition" for further details), U4, U5, and U6 are recruited; U2-U6 undergo conformational changes that prime the splicing reaction and components of the exon junction complex (EJC) are likewise recruited (Fig 1.1, panels pre-B through B*). Then, the 2' OH of the branchpoint completes the nucleophilic attack on the 5' phosphate of the guanosine, detaching the 5'ss from the upstream exon and forming the intron-downstream exon lariat intermediate (Fig 1.1, panel C). Next, further conformational changes of the U2-U6-U5 complex allow the freed 3' OH to complete the second nucleophilic attack on the 3'ss guanosine, uniting the two exons and freeing the intron lariat. The remaining spliceosome disassociates and can be recycled, and the EJC proteins are deposited on the RNA (Woodward et al., 2016). The intron lariat is readily debranched by DBR1 and readily degraded (Chapman and Boeke, 1991).

1.2 Intron and Exon Definition

For the minority of metazoan introns that are very short (<200 nt), the binding of U1, U2, and associated splicing factors is thought to be enough to commit the intron to splice through the process of intron definition (Fox-Walsh et al., 2005; reviewed in Ke and Chasin, 2011 and Conti, Barelle, and Buratti, 2013). In longer introns, U2 and U1 interact via regulatory proteins (e.g. SR-

proteins) bridging the exon in what's called the exon definition complex (Sterner, Carlo, and Berget, 1996). Exon definition has been hypothesized to operate as a quality control mechanism that prevents splicing at isolated splice site sequences not adjacent to an exon (reviewed in Ke and Chasin, 2010). This simple length threshold model between intron and exon definition, however, has been called into question by more recent work on co-transcriptional splicing that found, at least in Drosophila, longer introns (>1000 nt) were spliced before their downstream exon had been fully transcribed and therefore must have been spliced using intron definition (Prudencio et al., 2022). Therefore, while broadly most metazoan introns flanked by short exons seem to be spliced through exon definition, including potentially many recursive introns, the co-transcriptional nature of splicing may help further elucidate how splice sites are chosen and paired.

1.3 Co-transcriptional Splicing

When measured by nascent RNA-seq prepared from chromatin fractions, co-transcriptional splicing appears to be the predominant means of splicing, with 75% in budding yeast, up to 84% of introns in human, 87% in Drosophila, though only 45% was detected in mouse liver (Carillo Oesterreich et al., 2010; Ameur et al., 2011; Khodor et al., 2011; Khodor et al., 2012; Tilgner et al., 2012; reviewed in Neugebauer, 2019). Fluorescent microscopy likewise demonstrates the prevalence of the removal of the intron between 15 seconds and 4.5 minutes after the transcription of the 3'ss in human cells, suggesting a tight linkage between transcription and splicing (Huranova et al., 2010; Martin et al., 2013; Schmidt et al., 2011; Coulon et al., 2014; reviewed in Neugebauer, 2019). Complicating matters, direct nanopore sequencing of nascent RNA from human and Drosophila cell lines found that introns tend to be spliced in a defined order that does not necessarily follow transcription and after RNA pol II has transcribed several kilobases (4 kb in humans and 2 kb in Drosophila) of pre-mRNA, suggesting a more distant link (Drexler, Choquet,

and Churchman, 2019). Another recent study using bromouridine (BrU) RNA-seq, 4sU-seq, and its variant TT-seq found that fewer than half of human introns across six cell lines had their introns removed co-transcriptionally (Bedi et al., 2021). Regardless of whether it represents the majority of splicing or a large minority, co-transcriptional splicing, and therefore the many factors that regulate transcription kinetics, also contributes to splicing efficiency and splice site choice in a subset of transcripts (reviewed in Neugebauer, 2019 and Rogalska, Vivori, and Valcarcel, 2022).

1.4 Recursive Splicing

In the late 1990s, it was observed that the proximal end of a small internal exon within the long intron of the Ultrabithorax (Ubx) gene in Drosophila was alternatively used a a 5'ss (Hatton et al., 1998). It was later found that many long introns in Drosophila used alternative exons and non-exonic elements, called either recursive splice sites or ratchet points, to splice out introns in multiple pieces (Burnette et al., 2005). It was hypothesized and later demonstrated that this stepwise removal of introns in pieces helps to maintain splicing accuracy in the long introns in Drosophila (Pai et al., 2018). However, its function in humans remains less clear.

The term recursive splicing is sometimes differentiated from that of intra-splicing in the literature: Recursive splicing in the narrower definition refers to the stepwise, co-transcriptional splicing that removes sub-intronic pieces in transcriptional order whereas intra-splicing (also "nested splicing") refers to the removal of nested sub-intronic pieces (Ott et al., 2003; Burnnette et al., 2005; Suzuki et al., 2013; Radtke et al., 2017). In **Chapter 2**, I present my work using lariats in total RNA sequencing to identify sites of sub-intronic splicing and additional details on previous efforts to characterize sites of recursive splicing. Given our method's inability to detect the order of splicing as it uses the intron lariat and the precedent in the literature (Pai et al., 2018; Zhang et al., 2018; Wan et al., 2021), I will refer to both possibilities here broadly as recursive splicing.

There exists a large gap between the number of recursive sites identified by mapping split-reads across putative recursive junctions and those that have been supported by higher confidence means, such as sawtooth sequencing and lariat sequencing. The mapping of split-reads from nascent RNA-seq, when not explicitly constrained to AGGT sites, produces a majority of non-AGGT sites (>65%; Sibley et al., 2015), which raises questions either about the fidelity of the spliceosome or the RNA-seq amplification steps. Still, despite the potential that many of the apparent sites derive from chimeric cDNA produced during library preparation, it seems likely that some fraction, given the repeated findings across samples or even datasets, also represent genuine recursive intermediates.

Sawtooth sequencing, which utilizes read densities in nascent RNA-seq that follow a sawtooth pattern with peaks at 5'ss and approximately linear decreases across introns, represents a high-confidence metric for identifying recursive splicing, but its throughput is often limited by the need for manual verification even if computational methods are used (Duff et al., 2015; Zhang et al., 2018; Moon and Zhao, 2022). It was developed with the narrower definition of recursive splicing in mind, relying on the assumption that splicing occurs co-transcriptionally and that the intron segments will be spliced readily after RNA pol II transcribes the 3'ss such that 5'-mapping segments exist at a much higher abundance than the 3'-mapping segments, creating the slope between the two splice sites, and the characteristic sawtooth-like pattern between recursive splice (RS) sites. It also assumes that recursive splicing will be the sole or heavily dominant splicing mechanism for that intron, else the sawtooth pattern will not be readily discernable and performs best on long introns where the noise inherent in RNA-seq data can be overcome and not mistaken for signal by algorithms or visual inspection. To narrow the gap between the large number of putative RS sites identified by split-reads and those that have been further supported by other

means up until this point in humans, I sought to identify recursive splicing genome-wide using intron lariats in total RNA-sequencing, the results of which are presented in **Chapter 2.**

While the characteristic sawtooth read patterns suggested that recursive splicing reflected usage of "zero-nucleotide" exons, recent work in humans, mice, and Drosophila has identified that many, if not all, of these RS sites are linked to cryptic or alternatively included RS exons (Sibley et al., 2015; Blazquez et al., 2018; Joseph, Kondo, and Lai, 2018; Moon and Zhao, 2022). As described above, the consensus sequences for both the 3'ss and the 5'ss enforced by U2AF2 and the U1 snRNA, respectively, preference for a minimal recursive splice sites 5'-AG|GU-3'. Because of the presence of these proximal exonic 5' RS sites among cassette exons identified in the literature, I hypothesized that we might also see the usage of distal exonic 3' RS sites in cassette exons. The results from that analysis are presented here in **Chapter 3.**

Chapter 2. RECURSIVE SPLICING DISCOVERY USING LARIATS IN TOTAL RNA SEQUENCING

A version of this chapter and the next has been published on bioRxiv as:

E. R. Hoppe, D. B. Udy, R. K. Bradley, Recursive splicing discovery using lariats in total RNA sequencing. bioRxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.12.22.521701v1

I led this work and my contributions to this paper included conceptualization, investigation/analysis, visualization, and writing (original draft, review, and editing).

2.1 INTRODUCTION

RNA splicing is the process by which introns are removed from precursor mRNAs to form functional mature mRNAs and long non-coding RNAs, a ubiquitous process among eukaryotes essential for proper gene expression (reviewed in Lee and Rio, 2015). Alternative splicing broadly refers to the phenomenon in which the use of different splice sites produces different mature mRNA sequences that often code for different protein isoforms (reviewed in Nilsen and Graveley, 2010; Ule and Blencowe, 2019; Kelemen and Convertini et al., 2013); widespread usage of alternative splicing in the transcriptome greatly expands and diversifies the proteome without increasing genome size to the same degree (Nilsen and Graveley, 2010; Blencowe 2017). The breadth of alternative splicing within a transcriptome correlates with the complexity of the organism (Kim et al., 2004; Kim et al., 2007), with >95% of human genes displaying some form of alternative splicing (Pan et al., 2008; Wang et al., 2008). Such prevalence among higher eukaryotes implies that alternative splicing has an important contribution to the functional complexity of these organisms (Merkin et al., 2012; Barbosa-Morais et al., 2012).

Efficient removal of intronic sequences is essential for splicing; failure to do so leads to mature mRNAs that do not, for example, code for the correct protein sequence. Most splicing reactions are thought to proceed in a single set of two transesterification reactions (Wilkinson et al., 2020; Padgett et al., 1984), leading to the removal of the intron as a single lariat structure that is rapidly degraded (Padgett et al., 1984; Ruskin et al., 1984; Ruskin and Green, 1985). However, some introns are excised as multiple pieces across multiple reactions, a process referred to as recursive splicing (RS) (Hatton et al., 1998; Burnette et al., 2005; Kelly et al., 2015; Pulyakhina et al., 2015; Gazzoli et al., 2016; Gehring and Roignant, 2021; Conboy, 2021). Previous work characterizing recursive splicing has focused on long introns (> 50 kb) (Hatton et al., 1998; Burnette et al., 2005; Sibley et al., 2015), with the rationale that there exists a maximum length for intron excision that preserves the accuracy of splicing (Shepard et al., 2009; Suzuki et al., 2013; Pai et al., 2018; Burnette et al., 2005; Duff et al., 2015; Sibley et al., 2015) and/or maintain proper coordination between the spliceosome and RNA polymerase (Zhang et al., 2018; Pai et al., 2018).

Empirically identifying recursive splice sites is inherently difficult due to: (1) the transient nature of intron-lariat structures (Ruskin and Green, 1985; Mohanta and Chakrabarti, 2021), (2) the dearth of RNA-seq reads derived from lariats due to poly(A)-selection in most sequencing library preparations, and (3) the low probability of lariat-derived sequencing reads traversing the recursive splice site owing to the large size of many introns. Additionally, the frequency at which recursive splice sites are chosen over conventional sites remains unclear. Site preference may be stochastic, depend on intron identity, or regulated by as yet unknown mechanisms (Wan et al., 2021; Radtke et al., 2017). The rarity of recursive splicing relative to conventional splicing may also present challenges. Nevertheless, previous work has utilized novel approaches to discover recursive splice sites. Foundational minigene experiments in Drosophila demonstrated that the

Ubx cassette exon can be excluded via the usage of the cassette exon's 5' terminus as a recursive splice site, which excises the long stretch of intron-exon-intron sequence in multiple pieces (Hatton et al., 1998). Work from the same group used splice site sequence preferences to search for potential recursive splice sites among all annotated introns in the Drosophila genome (Burnette et al., 2005); this analysis yielded 165 potential recursive splice sites, primarily in long introns >10 kb, and a subset were experimentally verified.

The increased prevalence of RNA-seq has led to more high throughput analyses for identifying potential recursive splice sites from sequencing data. Two groups searched for novel recursive splice sites in Drosophila and humans (Duff et al., 2015; Sibley et al., 2015) using total RNA-seq data to identify "saw-tooth" sequence patterns – named for the characteristic coverage patterns in intronic regions due to reads from splicing intermediates: peaks at 5' splice sites followed by approximately linear decreases in signal intensity approaching 3' splice sites. Sibley et al. identified recursive splice sites in human genes for the first time, although restricting the analysis to only genes with introns >150 kb meant the search was not exhaustive. Subsequent studies used 4-thiouridine labeling to sequence RNA shortly after transcription to enrich for nascent RNA and more readily identify recursive splice sites in Drosophila (Pai et al., 2018) and human cells (Zhang et al., 2018). Importantly, the recursive splice site identification in human cells was restricted to introns >5 kb (Zhang et al., 2018), limiting the total number of introns analyzed.

The regulatory capacity of recursive splicing and its ability to modulate gene expression has not been characterized in detail. To assess potential mechanisms, the identification of recursive splice sites in a broader range of introns is necessary. Therefore, we sought to build upon the work of previous studies by searching for novel recursive splice sites in constitutive introns of all sizes in the human transcriptome using an unbiased approach to analyze previous sequencing datasets

potentially enriched for lariat-spanning reads and employing stringent filters to identify the highest confidence recursive splice sites.

2.2 RESULTS

2.2.1 *Global annotation of recursive spicing using total RNA sequencing*

The rarity of lariat spanning reads, specifically from recursive splicing, in traditional RNA-seq datasets makes thorough annotation of such sites difficult. To maximize the likelihood of identifying lariats associated with recursive splicing, we chose to analyze datasets without poly(A)-selection, a preparation method that depletes lariats (see Methods for more details).

Our method for recursive splice site identification differed from those of other groups that restricted their searches to long introns (Zhang et al., 2018; Sibley et al., 2015). We searched all constitutive introns in the human genome for $Y_{5+}N_{0-4}$YAGGT motifs, representing a minimal 3' splice site (short polypyrimidine tract and 3' splice site YAG) joined to a minimal 5' splice site (GT), to identify potential recursive splice sites. For each site, we chose five “decoy” sites at random positions in the same intron and filtered out those likely to be involved in cryptic splicing. Decoy sites are important for calculating the empirical false discovery rate (FDR) in recursive splice site identification. Each putative and decoy recursive splice site was paired with the annotated 5' and 3' splice sites in the respective intron and with other putative RS or decoy sites within the same intron, creating pairs of junctions for aligning with reads.

For empirically identifying recursive splice sites from the putative sites, we aligned RNA-seq reads to our junctions using a “split-read” approach described previously (Mercer et al., 2015; Pineda and Bradley, 2018). Briefly, the first 20 nucleotides of each junction, corresponding to the 5' splice site region, were mapped to prefiltered reads (**Fig 2.1A**). Successfully aligned reads were trimmed to remove the 5' splice site region and mapped to the 3' splice site region (last 250

nucleotides of the junction, **Fig 2.1A**). Reads that aligned successfully to the recursive splice site junctions were further filtered as described later (**Fig 2.3A**). Examples of recursive splice junctions supported by sequencing reads are shown in **Fig 2.1B**, including reads that contain the 5' portion of the recursive splice site (green) and reads in which the recursive splice site is inferred downstream from the branchpoint position found in the read (blue). Additional details on putative recursive splice site determination and split-read mapping are available in the Methods section.

2.2.2 *Lariat sequence achieves low empirical FDR with key filters*

The pipeline for empirically identifying recursive splice sites yielded a substantial number of such sites. However, the rate of false negative detection, based on the number of decoy sites that were called as recursive splice sites, was higher than anticipated (**Fig 2.3A**), indicating that the traditional sequence features used to identify lariat reads are insufficient for identifying genuine recursive splice sites with a low FDR. Therefore, we employed a series of filtering steps to minimize the number of false negatives while still identifying novel recursive splice sites with much higher confidence (**Fig 2.3A**).

The first filtering step required at least one read with a mismatch at the branchpoint site; such a mismatch is created from the low fidelity of reverse transcriptase when traversing the 2'-5' phosphodiester linkage at the branchpoint nucleotide (Vogel et al., 1997; Gao et al., 2008) and is indicative of bona fide lariat sequences. The next filtering step selected only “high confidence” recursive splice junctions that mapped to reads as described in Pineda and Bradley, 2018. Briefly, this required (1) >5% of reads mapping to a specific junction to contain a mismatch at the branchpoint nucleotide and no other mismatches in the 3' splice site portion of the read and (2) the

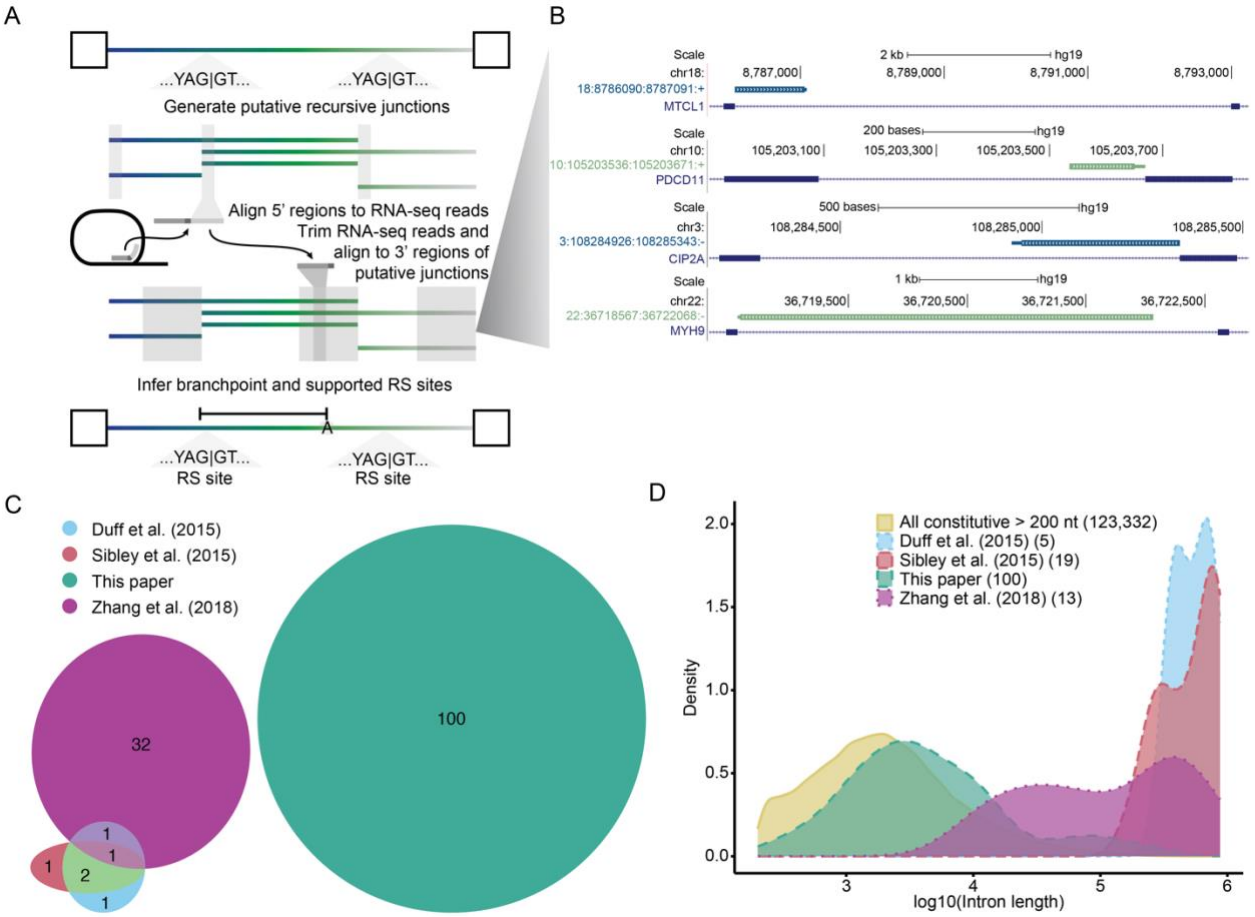

Figure 2.1. Global annotation of recursive spicing using lariats in total RNA sequencing. **(A)** Overview of the recursive lariat mapping algorithm. Intronic sites matching the sequence YAGGT with an upstream polypyrimidine track containing at least five adjacent pyrimidines separated by a span of 0-4 N (see Methods section "Lariat detection" for sequence motif) were identified as putative recursive sites. These sites were mapped against reads from total RNA-seq, after removing genome- and transcriptome-mapping reads, in a stepwise, split fashion to ensure the inverted read expected of lariats and enable us to identify branchpoint mismatches. **(B)** Representative recursive lariats. Thick bar, the region between the mapped 5' splice site and the branchpoint (loop of the lariat); adjacent thin bar, the inferred branchpoint to 3' splice site; navy, conventional 5' ss to RS 3'; green, RS 5' ss to conventional 3' ss. The plot is derived from the custom track generated from the University of California at Santa Cruz (UCSC) Genome Browser (Meyer et al., 2013). **(C)** Overlap of human RS sites identified as high confidence in previous studies, predominantly using sawtooth mapping, and those in the current work. **(D)** Size distributions for introns in which recursive sites were identified across different studies compared with constitutive introns. Counts of supported junctions with intron length annotations in parentheses.

sequences upstream and downstream of the branchpoint be unique to the intron of interest and not found elsewhere in the genome (to abrogate the misannotation of gene duplications as recursive splice sites). This filtering step had the largest impact on reducing the empirical FDR (**Fig 2.3A**, column 3).

An additional filter excluded mapped recursive splice sites within 100 nucleotides of annotated splice sites (**Fig 2.3A**, column 4) as a heuristic to exclude cryptic alternative 5' and 3' splice sites. Additionally, mapped sites were removed if the twenty nucleotides required for mapping the recursive site (downstream of the 5' splice site or upstream of the branchpoint) overlapped with a repeat or low complexity region (**Fig 2.3A**, column 5) to reduce uncertainty in the mapping. Previous work has shown that the branchpoint to 3' splice site is predominantly constrained within a range of 10-60 nucleotides (Taggart et al., 2012; Mercer et al., 2015; Taggart et al., 2017; Pineda and Bradley, 2018). Only mapped recursive splice sites in which the distance between the branchpoint and 3' splice site fell within this range were included for further analysis (**Fig 2.3A**, column 6).

A substantial portion of remaining reads that mapped to decoy random-site junctions were best explained as originating from self-primed amplification—whereby linear reads circularized or took on the appearance of being circularized by ligating up/downstream with themselves or a daughter strand during the PCR amplification step of RNA-seq library preparation—and did not represent reads from bona fide recursive splice sites based on manual inspection. These types of rare potential technical artifacts were removed in the last filtering step by comparing the five nucleotides upstream of the 5' splice site to the four nucleotides upstream of the branchpoint nucleotide plus the branchpoint and excluding reads with exact matches (**Fig 2.3A**). In aggregate,

these filters produced an overall empirical false discovery rate of 2.51% across all of our datasets, with a range of 0-3.9% for the five individual datasets included (**Fig 2.3B**).

2.2.3 *Comparison of high-confidence recursive splice sites with previously identified sites*

The methodological differences in our approach, compared to previous studies (Zhang et al., 2018; Duff et al., 2015; Sibley et al., 2015), led to the identification of previously unknown recursive splice sites. All 100 sites we identified were absent from previously annotated high-confidence sets of recursive splice sites (**Fig 2.1C**), while the previous sets had at least one recursive splice site in common with each of the other previous sets (**Fig 2.1C**). However, our sites did have limited overlap with the lower confidence method of split-read mapping from Sibley et al. (2015), including one site that met their recursive motif (**Fig 2.4**) and nine additional sites (data not shown). Together, this indicates that we have yet to reach saturation of the human recursive splicing annotation. Additionally, the size distribution of the introns from which the recursive splice sites were identified in previous work is much larger than that of our work (**Table 2.1**, **Fig 2.1D**), reflecting the specific selection for large introns in previous analyses (Zhang et al., 2018; Sibley et al., 2015). Our study demonstrates that smaller introns, too, which were largely excluded by previous efforts, are capable of recursive splicing. These comparisons highlight the value of unbiased analysis in all constitutive introns to discover novel recursive splice sites, as well as the need for the continuation of work on this topic, as it remains likely that many sites remain undiscovered.

Table 2.1. Characteristics of recursive splicing identified in this work compared to previous studies.

Sites here include all those provided that meet the filtering described in the methods of the respective papers for the primary analysis, typically not including the motif filtering that may be used elsewhere in the paper (excluding Duff). Intron annotations were used from only the originating paper, but from whichever class of intron they could be obtained. Both site numbers and intron lengths were weighted to accommodate multiple intron/gene annotations if they arose.

		Sites with Motif			Intron Length (kb)
	n	AGGT	YAGGT	+ polyY tract[1]	Median (min, max)
Duff et al. (2015)					
Sawtooth	5	5 *(100%)*	5 *(100%)*	4 *(80%)*	552.3 (370.4, 766.9)
Sibley et al. (2015)					
Sawtooth	19	9 *(47%)*	9 *(47%)*	7 *(37%)*	547.4 (171.8, 874.9)
Split-read (>150 kb introns)	2,520	849 *(34%)*	469 *(19%)*	115 *(5%)*	195.7 (15.3, 1,097.9)[2]
Split-read (1 - 150 kb introns)	57,343	17,589 *(31%)*	10,219 *(18%)*	2,091 *(4%)*	n.p.
Zhang et al. (2018)					
Lariats	21	21 *(100%)*	20 *(95%)*	9 *(43%)*	n.p.
Lariats + Sawtooth	3	3 *(100%)*	3 *(100%)*	2 *(67%)*	370.4 (32.9, 552.3)
Sawtooth	10	10 *(100%)*	10 *(100%)*	4 *(40%)*	100.8 (11.8, 552.3)
Split-read only	330	330 *(100%)*	280 *(85%)*	114 *(35%)*	36.4 (5.1, 1,055.3)
This paper					
Lariats (additional filtering)	100	100 *(100%)*	100 *(100%)*	100 *(100%)*	3.4 (0.2, 247.9)

[1] motif = $Y_{5+}N_{0-4}YAGGT$

[2] n = 483 with annotated introns

n.p. = annotation not provided by authors

2.2.4 *Lariat sequencing identifies recursive splicing in diverse introns*

With the identification of these high-confidence recursive splice sites, we sought to investigate the nature of the introns from which they were derived. For every constitutive intron in the human genome, we modeled the likelihood of an intron yielding an informative lariat read as a function of the intron length and estimated the proportion of reads we would expect to be informative for identifying recursive splice sites (see Methods for additional modeling details). As introns increase in length, the fraction of informative reads decreases (**Fig 2.2A**), as expected since longer introns contain more sequence not immediately adjacent (and thus indistinguishable from intronic sequence originating from genomic DNA or unspliced intermediate RNA reads) to the branchpoint nucleotide. Our modeling also indicates that longer read lengths lead to a higher likelihood of an informative read for any given intron length (**Fig 2.2A**).

We compared the expected intron length distribution to the length distribution of the identified recursive splice site introns (**Fig 2.2B**). Using the proportion of informative lariat reads calculations from **Fig 2.2A**, we calculated the probability-weighted distribution of expected intron lengths (**Fig 2.2B**, gray curve) and the unweighted distribution (**Fig 2.2B**, orange curve) for all constitutive introns. Interestingly, the sizes of the identified recursive splice site introns displayed a bimodal distribution, with peaks mimicking the peaks in the weighted distribution and the unweighted distribution (**Fig 2.2B**, blue curve). Because our method preferentially identifies recursive splice sites from shorter introns the first peak is expected, but the second peak indicates that recursive splicing may indeed be enriched to an extent in larger introns, where the focus of study has centered to date. It is worth noting that we did not control here for gene expression, which is likely to impact lariat discovery. Further, our annotation to date of recursive sites appears not to be fully saturated so the true distribution of intron lengths implicated in recursive splicing

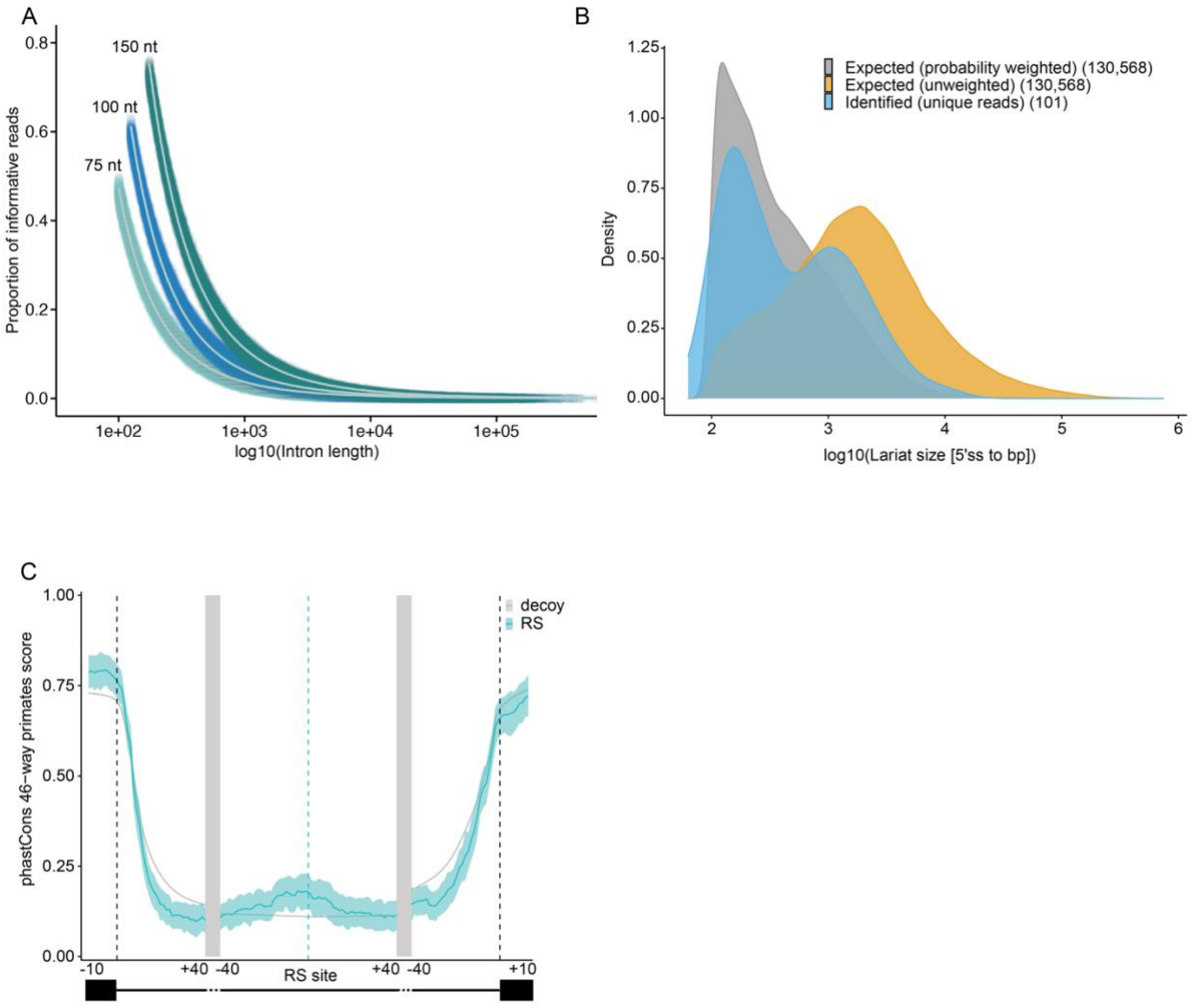

Figure 2.2. Lariats follow expected size distributions and RS sites show conservation enrichment.

(A) Estimated probabilities of a read of a given length being informative for lariat detection given the mapping constraints used in this work (see Methods section "Modeling informative reads" for parameters of informative reads). 75 nt, 100 nt, and 150 nt indicate the read length parameter used for each set of simulations. **(B)** Comparison of expected lariat sizes with our mapped lariats. The distributions of lariat sizes from constitutive introns, assuming equal expression, with and without weighting according to the probability of identifying the lariat (for 100 nt reads) compared to our recursive lariat reads. **(C)** PhastCons scores for the regions around the RS site and flanking exons for primates. RS sites identified in this analysis, blue; randomly chosen decoy site, grey. Light blue and light grey areas indicate 95% confidence remains to be uncovered. Nonetheless, these data indicate that the recursive splicing exists over a range of intron lengths. intervals from RS and random decoy sites, respectively, calculated using bootstrapping (replicates = 1,000).

remains to be uncovered. Nonetheless, these data indicate that the recursive splicing exists over a range of intron lengths.

To assess the level of evolutionary conservation at these sites, we analyzed the phastCons scores of the upstream and downstream regions of the site and surrounding conventional exons. The mean base-level conservation amongst primates was increased modestly immediately upstream of RS sites compared to randomly chosen decoy sites (**Fig 2.2C**; bootstrap support >95%). We also compared the level of conservation for placental mammals, which was slightly increased in the RS site region (**Fig 2.5A**; bootstrap support >95%), and amongst vertebrates, which was not increased (**Fig 2.5B**). Given the difficulty of alignment in intronic regions, a low level of conservation still may indicate selection at these sites. It also may point to the variability of conservation at such sites, in accordance with the finding that recursive splicing appears to be stochastic at some sites and seemingly regulated at others (Wan et al. 2021).

2.3 DISCUSSION

Since its initial discovery in *Drosophila*, much of the work on recursive splicing has focused on its prevalence in long introns. Here, we present evidence that recursive splicing occurs more broadly–both in small- to medium-length introns and at the distal end of exons—by employing an unbiased lariat sequencing approach to identify recursive splicing genome-wide.

For identifying RS sites from lariat reads, we chose a sequence motif that adhered to the U2 consensus sequence, which seemed likely to capture a substantial number of sites, both U2-type and U12-type, with minimal computational requirements. This analysis yielded 100 RS sites that strongly adhered to the U2 3′/5′ consensus sequence in constitutive introns. We limited the analysis to constitutive introns so there would be somewhat less uncertainty about the origins of the lariat, although alternative introns are a potentially rich source of recursive splicing and would

be of vital interest for future work. Importantly, our method assessed the false discovery rate of lariat sequencing and found it was highly dependent on the sequencing method used. Although poly(A)-selected RNA-seq can contribute meaningfully to conventional lariat counts (Pineda and Bradley, 2018), it was a poor fit for recursive analyses given the high FDR (data not shown). Datasets with more specific selection—such as those that use RNaseR to enrich for non-linear RNA—yielded the lowest FDRs, and poly(A)-minus datasets also produced reasonable results. The filtering steps employed to reduce the FDR identified sequencing artifacts that are called as recursive splice sites, including apparent self-priming amplification that caused circularization of DNA using at least five bases. Unexpected sequencing artifacts such as these justify the use of stringent filters for these analyses and emphasize the importance of calculating empirical FDR.

The filtering steps used to minimize FDR in our analyses revealed multiple aspects of sequencing that affect data quality. First, as expected from our modeling, increased sequencing read lengths lead to a higher proportion of informative reads and extend the size of lariats we could practically identify. As long read sequencing depth increases and short read sequencing length increases, our ability to identify additional recursive sites in introns in the 10-100kb size range will improve. Second, some selection is required for optimal sequencing of lariats. While removing poly(A) sequences (in addition to standard rRNA depletion) appears sufficient for achieving a low FDR, enriching further with RNaseR treatment or enriching for specific introns/lariats with a pull-down would be ideal for targeted RS lariat sequencing experiments. Third, the selection of small RNAs prior to reverse transcription appears to eliminate the bulk of lariats. The statistical size constraints described previously limit the usefulness of these datasets, so we do not recommend the use small RNA datasets for RS site identification. Fourth, patient samples that otherwise met the above criteria aren't useful with this method, regardless of tissue of origin. We hypothesize

that this is due to sample processing and handling protocols that may lead to the degradation of lariats, as has been observed with other classes of RNAs (Dvinge et al., 2014).

Although recursive splicing has been proposed to be important for *Drosophila* developmental timing (Duff et al., 2015; Pai et al., 2018), splicing fidelity (Shepard et al., 2009; Pai et al., 2018; Burnette et al., 2005; Duff et al., 2015; Sibley et al., 2015), and timing between RNA polymerase and the spliceosome (Zhang et al., 2018; Pai et al., 2018) in long introns (>10 kb), it is unclear what purpose recursive splicing serves in shorter human introns. Recent work that tested recursive splice sites using a reporter found that some sites were necessary for proper splicing of the intron (Radke et al., 2017), which, combined with the conservation of recursive splice sites among primates (**Fig 2C**) in our data, implies that recursive splicing is more than just a stochastic process in at least a subset of introns. Previous studies using single-molecule imaging (Wan et al., 2021) and 4-thiouridine labeling (Zhang et al., 2018) show that both RS and conventional splicing can occur in the same intron and the use of either splicing mechanism is often cell-type specific (Zhang et al., 2018), further demonstrating the need for additional work to elucidate the role of recursive splicing in splicing regulation.

2.4 Materials and Methods

2.4.1 *Genome annotations and alternative splicing identification*

All analyses were performed with the hg19 (GRCh37) genome assembly. The transcriptome/splicing annotation used throughout was created by merging the Ensembl v.71.1 gene annotation (Flicek et al., 2013), the UCSC knownGene gene annotation (Meyer et al., 2013), and the MISO v.2.0 isoform annotation (Katz et al., 2010).

2.4.2 *Dataset selection*

A manual search of the literature and datasets available in the Sequence Read Archive (SRA) was conducted to identify nascent human RNAseq datasets lacking polyA-selection and, ideally, enrichment for circular RNAs or lariats via polyA depletion, RNaseR, or inborn mutations in proteins involved in lariat degradation. Chosen runs were downloaded using the SRA toolkit and converted to fastq files [https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software]. Files that failed to download three times were excluded.

2.4.3 *Lariat detection*

Generate putative and decoy junctions. Constitutive introns, defined as those contained in all RefSeq transcripts of each parent gene, were searched for potential recursive splice sites using the U2 consensus 3' splice site motif appended with the minimal 5' splice site: $Y_{5+}N_{0-4}$YAGGT. For each potential recursive splice site within each intron, five decoy sites were selected at random positions in the intron. These decoy sites were then filtered to remove those with an adjacent GT and those within mapping range (250 nt) of a potential cryptic 3' splice site (identified with the motif $Y_{5+}N_{0-4}$YAG), another potential recursive splice site, or an annotated 3' splice site. In total, this method produced 202,395 unique putative recursive splice sites and 121,154 decoy splice sites. Each set of sites was used, independently, in conjunction with the annotated 5' and 3' splice sites to produce all pairwise junctions of potential 5' and 3' splice sites.

Pre-filter reads against genome and transcriptome. As described in Pineda and Bradley (2018), reads with less than 5% ambiguous bases were sequentially mapped against the transcriptome and then the genome (hg19/GRCh37) using Bowtie2. The following parameters were used with each mapping step: bowtie2 -x - -end-to-end –sensitive - -score-min L,0,-0.24 -k 1 - -n-ceil L,0,0.05 -U . Successfully aligned reads were discarded for subsequent steps.

Map reads to identified junction 5′ and 3′ regions. The pre-filtered reads were mapped to the putative recursive splice and decoy junctions described above, as detailed in Pineda and Bradley (2018). Briefly, the first 20 nt of each junction (conventional, putative RS, and decoy) was mapped to a Bowtie index created from the pre-filtered reads. For successful 5' alignments with no mismatches or indels where the reads aligned to a single 5' splice site region, the original read was trimmed from the first nucleotide of the 5' splice site alignment to the end of the read. These trimmed reads were mapped against a Bowtie index generated from the last 250 nt of each junction.

- 5' Mapping: bowtie2 -x –end-to-end –sensitive –k 10000 –no-unal -f -U <FASTA file of 5′ splice site sequences>
- 3' Mapping: bowtie2 -x <index file for 3′ splice sites> –end-to-end –sensitive -k 10 –no-unal -f -U

Infer branchpoint positions and recursive sites from split-read alignments. Using the same methods described in Pineda and Bradley (2018), alignments were restricted to the best-scoring alignment for each read, those with inverted alignments indicative of lariat reads, those that mapped to sites within a single gene, and those with a single mismatch at the branchpoint position. The branchpoint position was defined as the last nucleotide of the trimmed read alignment.

Recursive splice sites were deemed to be supported if they had at least one read that supported their use as either a 5' or 3' splice site that fit the following filtering criteria. In cases where multiple RS sites acting as a 3' splice site were supported by a single 5'-branchpoint mapping, a weight was assigned equal to 1/(number of possible RS sites), i.e. if two nearby RS sites could have produced the 5'-bp mapping, each was assigned a weight of 1/2 in counts.

2.4.4 *Filter hits*

Require criteria for high-confidence lariat detection. Mapped junctions were required to meet the criteria for high confidence outlined in Pineda and Bradley (2018): one or more reads with a mismatch at the branchpoint but no additional mismatches or indels in the 3′ splice site region of the read; ≥5% of mapping reads have a mismatch at the branchpoint but no additional mismatches or indels in the 3′ splice site region of the read; and unique sequences at the 5′ splice site and the region upstream of the branchpoint.

Self-primed. In order to filter out reads that originated from self-primed amplifications during RNA-seq processing, the 5 nt of sequence upstream of and including the identified branchpoint was compared to the 5 nt upstream of the identified 5' splice site, taking into account the strandedness of the sequence and read. Sequences with exact matches were discarded.

Low-complexity regions and simple repeats. The RepeatMasker out file for hg19 was obtained from UCSC (https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.out.gz). If the 20 nt region (minimum mapped region) including and downstream from the 5' splice site or upstream and including the branchpoint overlapped a region annotated by RepeatMasker as a simple repeat or a low complexity region, it was discarded.

Proximity to previously annotated splice sites and annotated exons. Mapped junctions were excluded if the putative recursive splice site fell within 100 nt of an annotated splice site of the same type (5' or 3'). This was performed by first using the resize function from GRanges [Bioconductor version 3.14] fixed to either “start” (5' splice site regions) or “end” (3' splice site regions) on a list of all intronic regions in our annotation in a strand-specific manner. These regions were then overlapped with the putative recursive junctions using overlapsAny from IRanges (Lawrence et al., 2013) such that those with an RS site acting as a 5' splice site were

overlapped with the 5' proximal regions and likewise those with an RS site acting as a 3' splice site were overlapped with the 3' proximal region.

Branchpoint distance from 3' splice site. The distance between the branchpoint and the 3' splice site, calculated as the absolute value of the position of the 3' splice site minus the position of the branchpoint (i.e. the length of the tail not including the branchpoint), was determined for mapped sequences. Those with distances outside the range of 10-60 nt were discarded, based on the distribution of this distance in previous work (Gao et al., 2008; Mercer et al., 2015; Pineda and Bradley, 2018) and a manual review of our initial results using the Mattick et al. (2015) sequencing data as a test set.

2.4.5 *Filtering UpSet plot*

The results (the number of lariat reads, the number of unique lariat reads, the number of unique lariat reads with one mismatch, with at least one mismatch, and the calculated FDR for each read type) from each of the filtering steps were collected manually for each permutation in order to optimize the order and to determine the essentiality of each set. These results were then manually formatted in Excel into a table compatible with the R package ComplexUpset (Krassowski 2020), with which they were then plotted.

2.4.6 *Modeling informative reads*

For each constitutive intron in the human genome, 1000 positions in the portion of the intron that makes up the lariat loop (i.e. from the start of the intron to the estimated branchpoint position) were chosen at random with replacement. The median U2 branchpoint position of 25 nucleotides upstream of the 3' splice site was assumed (Taggart et al., 2017; Pineda, Nicholas, and Bradley, under review). A position was considered able to produce an informative lariat read if it was (1) at least 20 nucleotides upstream of the estimated branchpoint position (minimum branchpoint-

mapping region) and (2) no less than n nucleotides away from the estimated branchpoint plus 20 (minimum 5'-mapping region), where n is a given sequencing read length. Read lengths of 75, 100, and 150 nucleotides were tested with this model to evaluate common sequencing read lengths.

2.4.7 *phastCons*

Conservation scores for vertebrates (phastCons100way.UCSC.hg19), placental mammals (phastCons46wayPlacental.UCSC.hg19), and primates (phastCons46wayPrimates.UCSC.hg19) were gathered using *GenomicScores* in Bioconductor [version 3.14]. Base levels of conservation were determined by randomly selecting one of the unfiltered, randomly chosen decoy sites for each intron considered in our analysis. The confidence intervals represent one standard deviation, which was determined by bootstrapping (n=100) the mean.

2.4.8 *External site analysis*

Previous work that identified recursive splice sites in human introns genome-wide using sawtooth sequencing or lariat sequencing was queried for tables of which RS sites or RS junctions from which RS sites could be derived. Three works met the criteria: Duff et al., 2015; Sibley et al., 2015; and Zhang et al., 2018.

- Duff et al. (2015) RS sites were acquired directly from the table located at https://static-content.springer.com/esm/art%3A10.1038%2Fnature14475/MediaObjects/41586_2015_BFnature14475_MOESM124_ESM.xlsx**.**
- Sibley et al. (2015) sites were parsed from the table located at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4471124/bin/NIHMS62941-supplement-Table_S1.xlsx.

- For the short-read sites, the RS site was inferred from the junction ID and the annotation of the “first part” and “second part” coordinate and type (whether the first “part” belonged to the exon or intron side) columns using the genomic sequence annotations in the table to manually check a subset of sequences; then, they were filtered according to the table classification as “recursive”. This column appears to be largely determined by adherence to the paper’s described requirement of >11 pyrimidines, a suitable 3'ss motif (annotated), and a suitable 5'ss motif (annotated); however, we were not able to reproduce the exact count reported in the paper so we defaulted to the table’s classification as there may have been an additional factor we overlooked. The junction IDs are shared between these sites and the linear regression sites used for the sawtooth analysis. Thus, we were able to use the intron information provided with the linear regression table to annotate a subset of the short-read sites.
- For the sawtooth linear regression analysis sites, the RS site annotations were populated by joining the table by the shared Junction IDs. Due to the short read annotations containing only Exon-Intron or Intron-Exon annotations, there are no RS-RS sites, which may plausibly exist in the linear regression annotation, so this may be a slight undercount. We attempted to annotate the RS sites based on a comparison of the junction coordinates with the intron coordinates, but due to inconsistencies of unknown origin with the strand annotation relative to the short read sites, were unable to use this method, which would have allowed for the identification of such sites. We defined sawtooth RS sites as those that were

annotated with “Double Significance and improved gradient”, following the paper’s methods.

- Zhang et al. (2018) RS sites were acquired directly from the table from the tables located at https://doi.org/10.1371/journal.pgen.1007579.s011; https://doi.org/10.1371/journal.pgen.1007579.s012; and https://doi.org/10.1371/journal.pgen.1007579.s013. We used the site selection criteria described in the methods (a fold change greater than 2 and a p-value less than 0.01) to filter for sites that met it in at least one of the samples. We did not consider the manual review criteria as it was not parsable from the table’s text and was described as optional in the paper’s methods.

The counts in the table were generated such that they included all provided sites for the given methods that were not supported by a “higher” evidence level. A subset of sites were annotated as belonging to multiple introns; therefore, we weighted based on how many introns they were said to belong to and took the median intron length using the *weighted.median* function from the R package limma (Ritchie et al., 2015). They were then also assessed for the presence of the given motifs at the RS site.

Intersections of our sites with external sites. For the Euler sites and UpSet plots, we included those sites that were annotated as recursive by the authors. The Euler plot was generated using the R package eulerr (Larsson 2021). Sites were intersected using just the chromosome and site information without regard to the strand given the uncertainty about the strand annotations in a subset of the Sibley sites. The UpSet plot was generated using the R package UpSetR (Gehlenborg 2019).

2.4.9 *Plots*

All plots, unless otherwise specified, were generated in R using ggplot2 (Wickham, 2016).

2.4.10 *Tables*

Tables included in the text were generated using the R package gt (Iannone et al., 2022).

2.5 Supplementary Figures and Legends

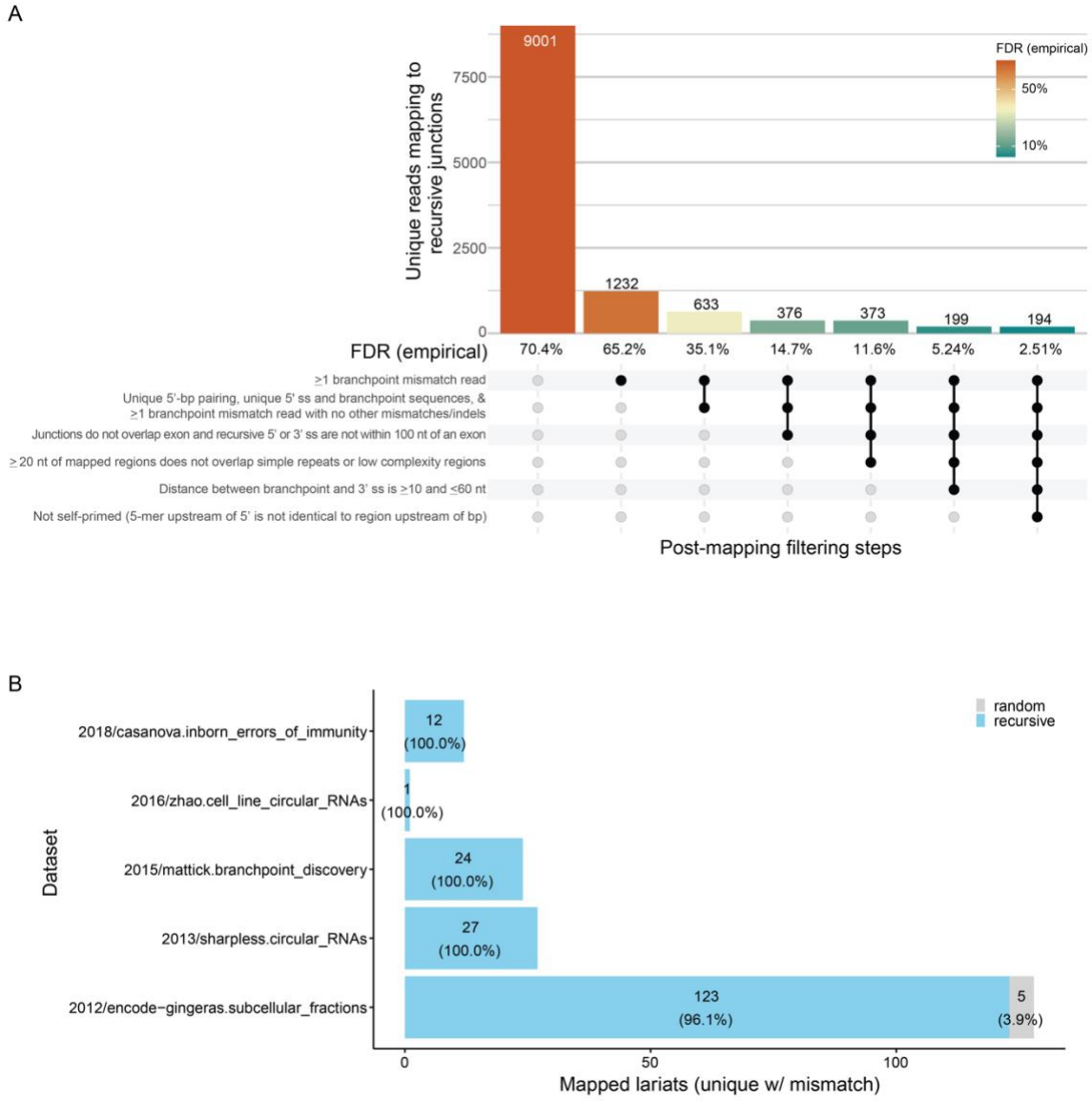

Figure 2.3. Empirical FDR estimation.
(A) Empirical FDR estimations for each consecutive filtering step. See methods for further details. **(B)** By dataset comparison of final FDR estimates. A subset of datasets from Sharpless et al. (2013), Mattick et al., (2015), and Zhao et al. (2016) were treated with RNaseR to enrich for circular RNAs. Mattick et al., (2015) also contained sequences enriched for specific introns using CaptureSeq. Likewise, Casanova et al. (2018) contained patient-derived fibroblasts from individuals with DBR1 mutations predicted to inhibit lariat degradation. The ENCODE data did not specifically enrich for circular RNA but did deplete poly(A)-containing sequences. All datasets involved standard rRNA depletion. For additional enrichment and sequencing information, see Supplementary Table 1 and Methods

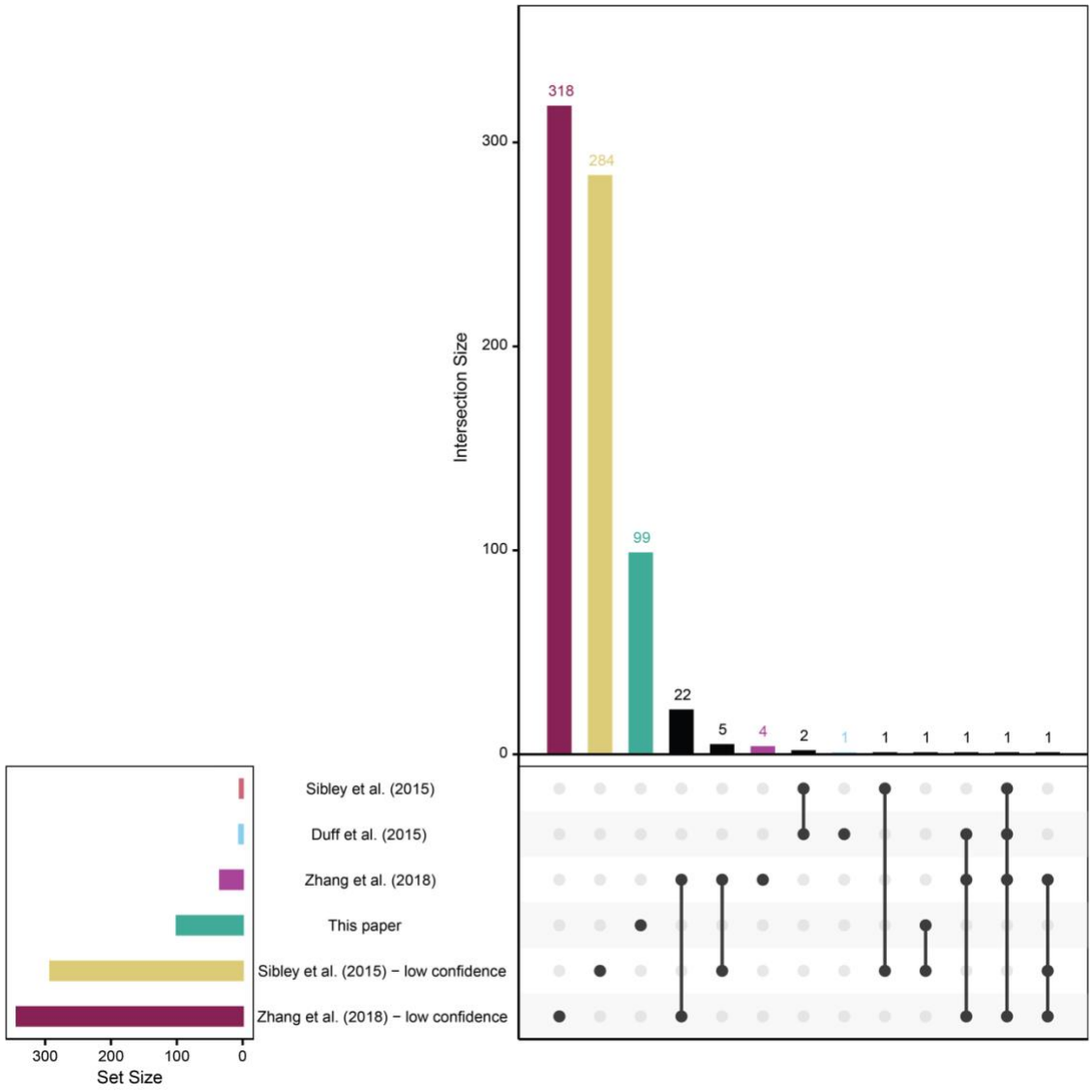

Figure 2.4. Intersection of previously published sites and those from this analysis.
Sites included are those indicated as recursive (typically by the use of a motif filter) in the provided annotation. Sites are stratified by confidence level: high/standard confidence sites for Sibley et al. (2015) and Duff et al. (2015 were identified via the sawtooth approach; high/standard confidence sites for Zhang et al. (2018) were derived using a combination of sawtooth and lariat profiling; and low confidence sites represent those identified by split-read analysis alone.

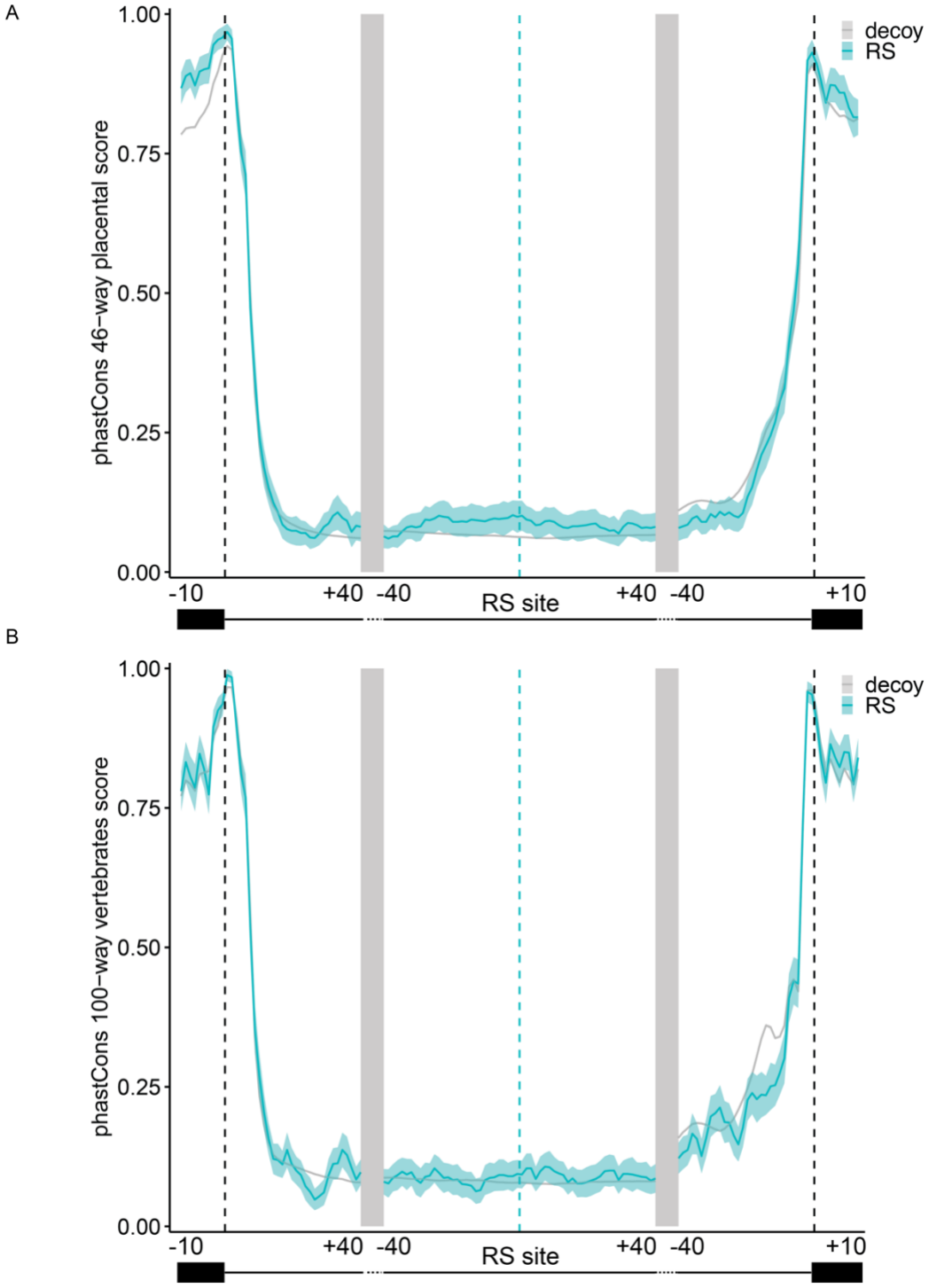

Figure 2.5. Conservation scores for RS sites and flanking exons for placental mammals and vertebrates.

Conservation PhastCons scores for the regions around the RS site and flanking exons for **(A)** placental mammals and **(B)** all vertebrates. RS sites identified in this analysis, blue; randomly chosen decoy site, grey. Light blue and light grey areas indicate 95% confidence intervals for RS and decoy sites, respectively, calculated using bootstrapping (replicates = 1,000).

2.6 Additional Notes

This section serves as an addendum and is not included in the cited manuscript:

After this project was conceptualized, a few groups performed similar lariat analyses. Zhang et al. (2018), which was included above, used a similar inverted, split-read mapping strategy to ours with a couple of key differences: They eliminated only those reads that mapped to the genome, not also those that mapped to the transcriptome. They used a slightly longer 23 nt mapping region on each side, compared to our 20 nt. They allowed branchpoints <100 nt from the 3'ss and allowed but did not require a mismatch. They did attempt to account for repetitive sequences through a remapping technique. Similarly, Radtkey et al. (2017) allowed branchpoints <100 nt from the 3'ss, and seems to have allowed but not required a mismatch at the branchpoint is otherwise vague about their mapping criteria; the sites were not provided with the pre-print and thus were not included in this analysis. Finally, Wan et al. (2021) also performed a split-read analysis by splitting the reads into "pieces of various lengths" and keeping the longest unique alignments of the 3' ends of the reads. While they did require inverted mapping, they did not appear to allow for or require a branchpoint mismatch nor was additional filtering detailed in the methods. I obtained the processed mapping data from the authors and performed the same quality control tests I had done while building my pipeline; unfortunately, I had trouble finding sequences based on the provided annotation that were reminiscent of the genuine lariats reads I identified in my own analysis. Therefore, I did not include these results in the comparison in the above work. Given the more stringent nature of our analysis and the new sites found, we felt that our method represents a substantial contribution to recursive splicing discovery.

Given the lab's prior experience identifying lariats from conventional splicing using bulk RNA sequencing, we were very hopeful we could use standard poly(A)-selected datasets even if

recursive splicing turned out to be a factor or two rarer than conventional splicing. I initially set up filters on the lariat-selected dataset from Mercer et al. (2015), and achieved an empirical FDR of 0% with the lariat characteristics filter (branchpoint mismatch, inverted sequence), the high confidence filter from the lab's previous analysis (unique 5'-bp sequence and pairing, and a read with at least one branchpoint mismatch and no other mismatch/indel), and the branchpoint to 3' splice site distance filter. And while the samples with RNaseR treatment and CaptureSeq produced more lariats or a slightly lower FDR (without the branchpoint filter), there were still many lariats identified from the unenriched samples. Because of this, I attempted to run the pipeline on the TCGA dataset, but the results were disappointing. I used a subset of filtering steps (I had not yet implemented the repetitive regions or self-primed hits), and the FDR was 31% across all TCGA datasets compared to about ~10% without those for the enriched datasets for that set of filters). Even requiring that all branchpoints be the classic adenosine and that we have multiple unique lariats supporting each site (possibly more stringent than our current set) reduced the FDR to ~16% with filters subset, supporting 36 RS sites.

Given this, I searched the Sequence Read Archive for datasets that mentioned circular RNAs, RNaseR, lariats, and other keywords and indicated they might be suitable for our purposes. I also included a sampling of other total RNA-seq datasets that did not select for poly(A) to see if this alone might be sufficient. Unfortunately, only the enriched datasets performed well, with the others contributing only a handful RS lariats, often with FDRs in excess of 10-20% with our initial filter subset, as observed with the TCGA data. The primary human tissue samples seemed to perform especially poorly, which made us wonder if sample processing differences might lead to a degradation of true lariats, effectively enriching for the RNA species that map as decoy hits. This is something that would have been of interest to investigate more

conclusively had there been time as enriching for lariats could aid greatly in producing both discovery RNA-seq datasets and could potentially inform methods for handing samples for confirmatory analyses like TOPO-cloning, which is known to be very finicky.

Chapter 3. DISTAL EXONIC RECURSIVE SPLICE SITES

A version of this chapter, with the exception of the Background section, and the previous has been published on bioRxiv as:

E. R. Hoppe, D. B. Udy, R. K. Bradley, Recursive splicing discovery using lariats in total RNA sequencing. bioRxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.12.22.521701v1

I led this work and my contributions to this paper included conceptualization, investigation/analysis, visualization, and writing (original draft, review, and editing).

3.1 RESULTS

3.1.1 *Distal exonic RS sites contribute to exon exclusion*

Blazquez et al. (2018) described RS-exons where after the upstream intron is spliced out, the reconstituted 5' splice site in the exon is used for the removal of both the exon–here called proximal recursive exons–and the downstream intron. Our results here demonstrate there is likewise exon exclusion using a distal exonic 3' splice site. The choice of this distal recursive splice site enables the upstream intron and exon to be removed in one splice reaction, followed by the splicing of the conventional downstream intron (**Fig 3A**).

We initially observed potential support for this mechanism in sequence logo plots of cassette exons. The final dinucleotide of exons, both cassette and constitutive, is commonly AG (**Fig 3B**; 56.2% cassette, 55.6% constitutive), forming a minimal recursive splice motif. We hypothesized that a subset of these sites had sufficient pyrimidines upstream to be recognized as a 3' splice site to enable the use of this distal exonic site. We calculated the Maximum Entropy (MaxEntScan) scores (Yeo et al. 2004) to assess how closely they adhered to the sequence of

constitutive introns; 10.2% of cassette exons maintain a MaxEntScan score of at least 4.07 (5th percentile of constitutive 3' splice sites) suggesting the potential for recognition as a 3' splice site, with 1.62% of cassette exons (780) having a MaxEntScan score at least as high as the 50th percentile of constitutive 3' splice sites. Given the ability of AG-dependent introns to be spliced with very short or weak polyY tracts as long as they maintained a strong YAG 3' splice site (reviewed in Moore 2000) and the potential constraints of on polyY tract length in coding exons, we moved forward looking at sites with a YAG as the last tri-nucleotide of the exon.

To investigate whether there was evidence for the use of such sites as distal recursive exons, we analyzed the rate of inclusion of cassette exons that ended in this minimal YAG compared to all others, including those that ended in AG. At the 50th percentile for inclusion across the median isoform inclusion rate of the 16 BodyMap tissues, exons that ended in YAG were included 23% less than those that ended in another tri-nucleotide (**Fig 3C**). The overall distribution of the median p.s.i. values of YAG-ending cassette exons were likewise significantly left-shifted (Kolmogorov–Smirnov test, p = 6.51 x 10^-37). These data suggest that the presence of YAG in the 3' splice site of a cassette exon is a strong signal for its exclusion.

Given the evidence for the use of distal recursive exons, we performed a modified version of the intronic recursive splicing analysis above. We determined that the critical filter for distal recursive exons was to require that the branchpoint fall within the body of the exons. This enabled the exclusion of alternative 3' splice sites or longer than typical branchpoint to 3' splice site lengths. In total, we found evidence of 10 distal recursive exons (**Table S2**), which each passed a manual review for sequence and mapping irregularities that had been filtered for in the larger intronic recursive site set. A selection of mapped lariats is shown in **Fig 3D**. The MaxEnt scores for these lariat-supported distal exonic RS sites ranged from -2.62 to 11.61 with a median 7.34 (**Table S2**),

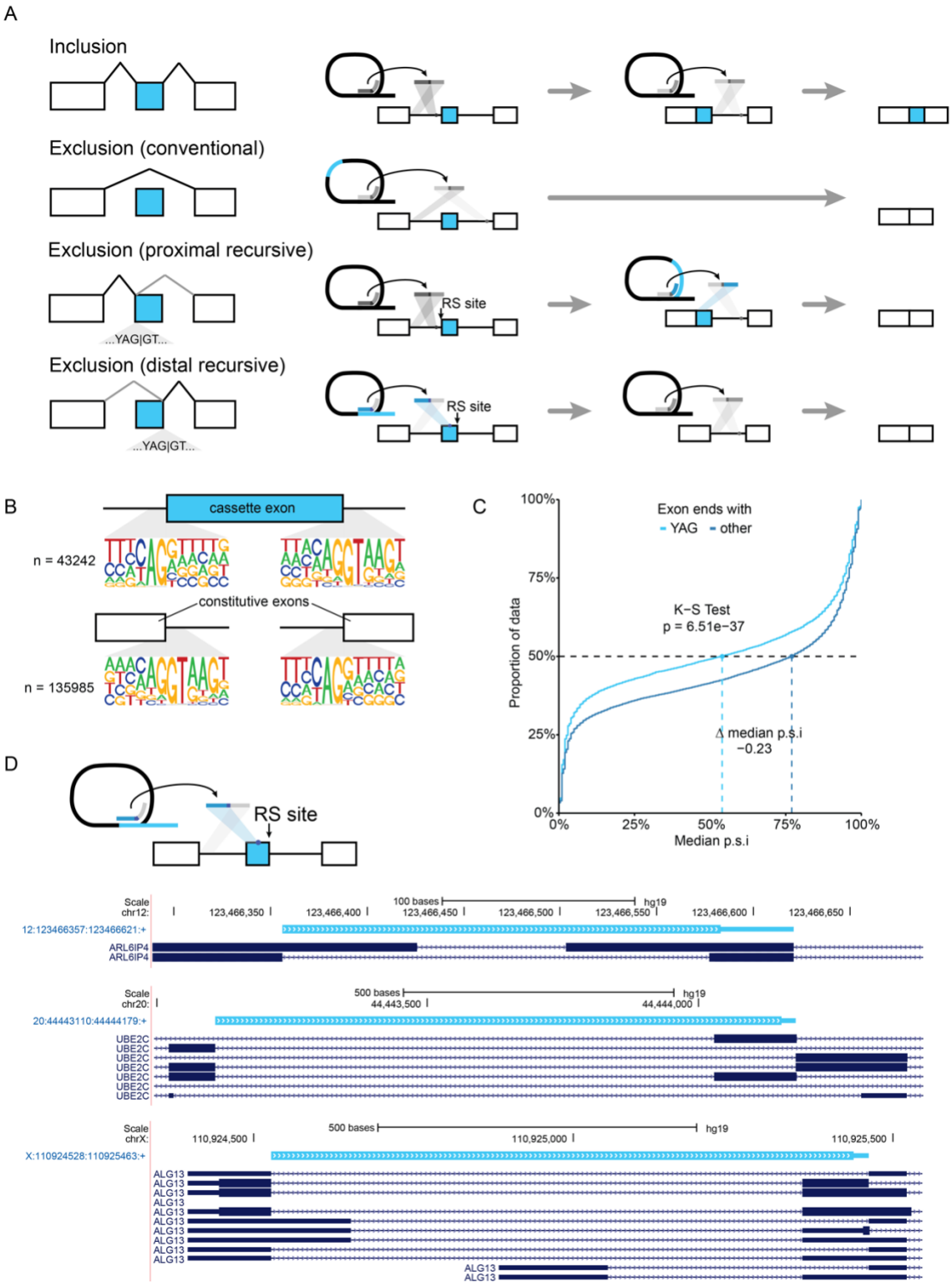

Figure 3.1. Distal exonic recursive splicing may contribute to exon exclusion.
(A) Schematic of exon inclusion or exclusion using conventional or exonic recursive splice sites. **(B)** Sequence logo plots of cassette exons aligned with constitutive exons showing strong enrichment for AG|GT at the distal end of exons. **(C)** Proportion of data with median percent

spliced in (p.s.i.) values across BodyMap tissues of cassette exons ending in YAG or any other trinucleotide. K-S test, Kolmogorov–Smirnov test. **(D)** Representative distal exonic recursive lariats. Thick bar, the region between the mapped 5' splice site and the branchpoint (loop of the lariat); adjacent thin bar, the inferred branchpoint to 3' splice site. The plot is derived from a custom track from the University of California at Santa Cruz (UCSC) Genome Browser (Meyer et al., 2013) and contains all unique, annotated isoforms within the given window.

indicating that they largely, but not exclusively, adhere well to the U2 consensus 3' ss sequence. Some sites are supported by multiple datasets and showed a diversity of branchpoints; for instance, the 9th exon in CAPN10 (hg19 2:241536098:241536359:+) has four unique branchpoints, one of which is supported by lariats from two datasets (**Table S2**).

In further support of the usage of distal exonic RS sites, two of the sites identified in this analysis are located at the junction between two mutually exclusive adjacent exons (**Fig 3D**), known as a dual specificity site (Zhang et al., 2007). The maintenance of dual specificity sites in the genome demonstrates the existence and usage of distal exonic 3' splice sites, and suggests that they also may rely upon exon definition of a downstream exon. Although relatively rare in the genome, dual specificity sites demonstrate that both the distal and proximal exonic recursive splice sites may be used alternatively to include or exclude adjacent exons. What genomic signals are associated with the use of dual specificity sites and exonic recursive sites more broadly remains unclear, but they present an intriguing possibility as an evolutionary intermediate for exon birth and death.

3.2 DISCUSSION

A provocative hypothesis, supported by our finding that the motif suggestive of distal recursive exons is enriched among evolutionarily young exons, is that RS sites contribute to exon birth and death. Such sites could provide variation in splicing—without substantial fitness costs—until

they are paired and selected as distal/proximal RS sites flanking a new exon or eliminated. Our analysis here identified lariats supporting recursive usage of the 3' splice site of dual specificity sites/distal exonic recursive splice site within coding exons (**Fig 3D**). However, we did observe a bias towards the support of distal recursive exons in UTR regions (data not shown), in which case the protein-coding sequence would be unchanged. Zhang et al. reported observing multiple protein products from genes with dual specificity sites (Zhang et al., 2007); the different proteins corresponded to mRNA isoforms differentially spliced using the dual specificity sites, evidence for the presence of functional 3' splice sites within the coding sequence. Together with the work on proximal recursive exons (Blazquez et al., 2018), these data indicate that splicing happens within exonic sequences to some degree. Whether the same mechanisms of exon junction complex removal that promote the usage of proximal recursive splice sites (Blazquez et al., 2018) likewise promote the usage of distal recursive exons remains to be determined

3.3 METHODS

3.3.1 *Lariat detection*

The same mapping (but not filtering) steps were followed as in Chapter 2 except that internal cassette exons and their upstream introns were used to build all possible junctions of paired annotated 5' splice sites and potential distal exonic 3' splice sites, regardless of sequence. Exons were considered internal if they were not first or last across all RefSeq transcripts that included them (across all gene names) in the RefSeq annotation.

3.3.2 *Exon inclusion analysis*

Exon inclusion was estimated across the 16 tissues in the Body Map 2.0 database as described previously (Dvinge et al., 2014). A Kolmogorov–Smirnov test was performed using the function

ks.test [stats R package version 4.3.0] with a null hypothesis of "not greater" for the median percent spliced in (psi) across all Body Map tissues comparing internal cassette exons where the last three nucleotides were YAG were compared to all other internal cassette exons.

3.3.3 *Maximum entropy scores*

MaxEntScan scores as defined in Yeo et al. (2004) were calculated using an R function that called the Perl scripts available for download from http://hollywood.mit.edu/burgelab/maxent/download/ for the 3' splice sites. Traditionally, MaxEntScan scores for conventional 3' splice sites require the 20 bases of the intron and 3 bases of the exon; here, for calculating the score for distal exonic RS sites, we used the 20 bases upstream and including the last base of the putative distal RS exon and the three bases of the downstream intron.

3.3.4 *Sequence logo plots*

Sequence logo plots were generated using an adapted version of the seqLogo R package (Bembom and Ivanek, 2022).

3.3.5 *Exon age categorization*

Exon age annotations were obtained from Corvelo and Eyras (2008) at https://static-content.springer.com/esm/art%3A10.1186%2Fgb-2008-9-9-r141/MediaObjects/13059_2008_2003_MOESM2_ESM.zip, which categorized exons using the conservation of sequences 20 nt upstream and 20 nt downstream the 5' and 3' splice sites of annotated exons. They used a comparison of five species (human, mouse, cow, chicken, and tetradon) so primate-specific exons are those that were in humans, but not mouse or cow (and therefore, may be human-specific or primate-specific); mammalian-specific exons were those in humans, mice, and cows. Vertebrate and older exons were those observed in all five species. The

reported hg18 coordinates were converted to hg19 using the liftOver utility provided by the UCSC Genome Browser (Meyer et al., 2013) using default settings (minimum ratio of bases that must remap = 0.95; allow multiple output regions; minimum hit size = 0; minimum chain size = 0). We performed the binomial exact test (alpha = 0.05, alternative = "greater") using the *binomial.test* function from the R package stats (R Core Team, 2022).

3.3.6 *Plots*

All plots, unless otherwise specified, were generated in R using ggplot2 (Wickham, 2016).

3.4 Supplemental Figures

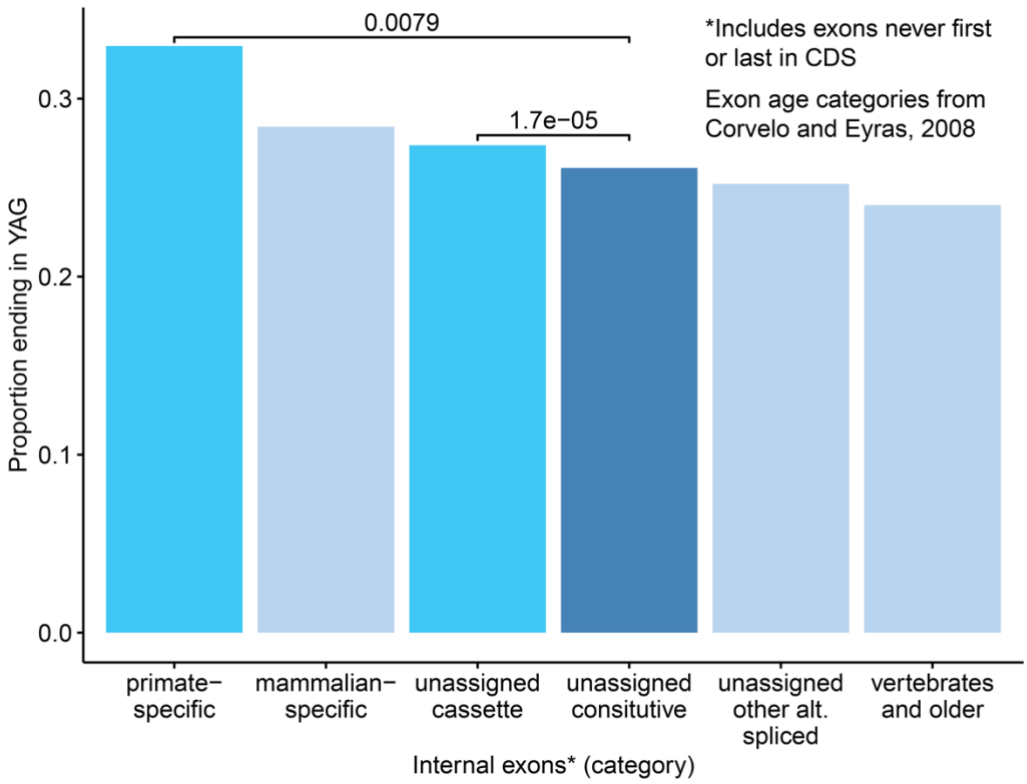

Figure 3.2. Comparisons of the proportion of exons that end in YAG among different exon age classifications and splicing classifications for exons that were not age-classified.
Details on how Corvelo and Eyras (2008) defined each category may be found in the Methods (section "Exon age categorization"). P-values were calculated using the binomial exact test.

Chapter 4. DISCUSSION

The goal of my thesis work was to better understand where splicing happens in the human genome, specifically to better characterize where recursive splicing in the broad sense occurs. In **Chapter** 2, I demonstrated that recursive lariats could be identified in enriched total RNA sequencing identified with a low empirical false discovery rate and found 100 new sites of recursive splicing in a broader range of intron sizes than previously described. In **Chapter 3**, I identified distal exonic RS sites as a new location for recursive splicing, which appear to contribute to the exclusion of cassette exons and may represent one means of exon birth, given their enrichment in younger exons.

Given the limited overlap with previous annotations, it is likely that we are far from reaching saturation with our human recursive splicing annotation. As mentioned in the introduction, analysis of split-read sequencing from just a few datasets finds hundreds of RS sites in humans that have the expected RS motif (Sibley et al., 2015; Zhang et al., 2018), but there has been sawtooth support for fewer than 20 (Duff et al., 2015; Sibley et al., 2015; Zhang et al., 2018). Lariat sequencing provides a useful tool to close that gap—and potentially expand beyond the hundreds of sites with sufficient sequencing data—and complements sawtooth sequencing by preferentially finding RS sites in smaller introns. Yet, the set of sites identified by lariat sequencing and sawtooth sequencing/split-read sequencing is relatively disjoint, indicating that we have a long way to go for the genome to be fully annotated. Here, I have presented both a set of filters that enable the identification of a set of high-confidence recursive lariats from enriched sequencing data as well as a framework for assessing empirical FDR for evaluating additional datasets as well. I'm hopeful this will be built upon to form a more comprehensive annotation of where recursive

splicing occurs in the human genome and further elucidate what factors are essential for regulating it.

It is interesting to speculate, however, about why sawtooth sequencing finds so few splice sites in humans when it can find hundreds in Drosophila (Duff et al., 2015; Joseph, Kondo, and Lai, 2018; Pai et al., 2018). The obvious hypothesis is that there is simply less recursive splicing in humans. There is some support for this. Blazquez et al. (2018) suggests that differences in the EJC of Drosophila and humans may play a role: less efficient repression of proximal recursive exonic 5'ss by the EJC in Drosophila compared to humans may lead to a lower preference for the inclusion of RS exons and higher preference for recursive splicing relative to humans in Drosophila. Likewise, in invertebrates, like Drosophila, small (50-500 nt) introns predominate, whereas in vertebrates, average introns lengths exceed 1-2 kb (Deutsch and Long, 1999; Pai et al., 2017). Perhaps in invertebrates with some long introns like Drosophila, recursive splicing is relied upon as a regulated method of removing the occasional long intron using a spliceosome largely evolved for shorter introns, leading to the observation that recursive splicing tends to subdivide Drosophila introns into recursive segments of regular length (Pai et al., 2018; Joseph, Kondo, and Lai, 2018), while higher vertebrates lacked the selection pressures necessary to fix recursive splicing as the dominant splicing outcome in a similar proportion of introns. Radtke et al. (2017), Wan et al. (2021), and our results suggest that many human introns may often use both recursive and conventional splicing; use of both splicing outcomes is likely to mask the identification of these introns using sawtooth sequencing. Further, depending on how inclusive of a definition of recursive splicing you choose, it may be worth considering sites that are spliced post-transcriptionally, which would be unlikely to show a sawtooth pattern even if they were spliced in pieces. This would be of interest for alternative introns and the last intron of genes, which are less

likely to be spliced co-transcriptionally (Ameur et al., 2011; Khodor et al., 2011; Schmidt et al., 2011; Tilgner et al., 2012). For these reasons, lariat sequencing represents an excellent complementary method for assessing recursive splicing, as it does not need to make assumptions about the order or timing of splicing.

Overall, based on our results, recursive splicing does appear to be much less common than conventional splicing in the human genome. This is perhaps not surprising given the amount of high-confidence RS sites annotations of there were prior to this work, but going into this project, we underestimated, as we imagine other groups did, the impact that rare sequences would have on the ability to detect recursive lariats, especially give the lab's success with conventional lariats in Pineda and Bradley (2018). I am hopeful our filtering and FDR estimation framework will prove helpful. In our results, I found that recursive lariats were approximately 2.5-3 fold less common than conventional lariats in our enriched datasets. Likewise, introns with recursive lariats in our analysis almost always had conventional lariats as well when I checked, though I did not perform a comprehensive analysis. This is consistent, however, with Wan et al. (2021), which proposed a stochastic model for recursive splicing and found that repressing RS sites with an ASO led to a lower splicing efficiency for the recursive intron studied, but cells were generally able to choose another RS site to use. Likewise, Radtke et al. (2017) performed minigene experiments and found a mixture of cases where the putative RS sites were deleterious, neutral, or required for splicing efficiency—surprisingly, the intron where the RS sites were essential was only 1 kb, demonstrating its importance in smaller introns as well. In sum, recursive splicing in humans seems to be largely a supplement for conventional splicing, perhaps used stochastically as a means of recovering transcripts after choosing an intronic splice site. Still, it may be the primary means for splicing in

a subset of introns, including those identified in sawtooth sequencing, or under certain regulated conditions.

To my knowledge, our work is the first to recognize the potential link between dual-specificity sites and recursive splicing. I am hopeful our paper may spark an interest in pursuing this connection, as I suspect these sites may serve as useful models for increasing the field's understanding of splice site choice. Unfortunately, the original annotation of the sites from Zhang et al. (2007) appears to have been lost as they were privately hosted; however, it would be worthwhile to replicate their analysis regardless with a more modern splicing annotation. Alas, it was not within the scope of the work I was able to accomplish here. Given that the U1 snRNA and U2AF2 both preference minimal AG|GU recursive splice sites in their consensus sequences, it is not surprising that dual-specificity sites and RS sites, both exonic and intronic, in general, exist. Our results indicate that there is conservation at RS sites among primates both for intronic sites and enrichment for sequences indicative of distal RS exons in younger exons. However, it remains to be fully elucidated how the spliceosome chooses to use putative recursive splice sites, how their use or disuse is selected, and through what specific mechanism they may contribute to the selection of new sequences for exons.

BIBLIOGRAPHY

Ameur A, Zaghlool A, Halvardson J, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol. 2011;18(12):1435-1440. Published 2011 Nov 6. doi:10.1038/nsmb.2143

Aslanzadeh V, Huang Y, Sanguinetti G, Beggs JD. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast [published correction appears in Genome Res. 2018 Apr;28(4):606.2. Genome Res. 2018;28(2):203-213. doi:10.1101/gr.225615.117

Barbosa-Morais NL, Irimia M, Pan Q, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*. 2012;338(6114):1587-1593. doi:10.1126/science.1230612

Bembom O, Ivanek R. *seqLogo: Sequence logos for DNA sequence alignments*. 2022. R package version 1.64.0.

Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A. 1977;74(8):3171-3175. doi:10.1073/pnas.74.8.3171

Blazquez L, Emmett W, Faraway R, et al. Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Mol Cell*. 2018;72(3):496-509.e9. doi:10.1016/j.molcel.2018.09.033

Blencowe BJ. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends Biochem Sci*. 2017;42(6):407-408. doi:10.1016/j.tibs.2017.04.

Burnette JM, Miyamoto-Sato E, Schaub MA, Conklin J, Lopez AJ. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*. 2005;170(2):661-674. doi:10.1534/genetics.104.039701

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell. 2010;40(4):571-581. doi:10.1016/j.molcel.2010.11.004

Chapman KB, Boeke JD. Isolation and characterization of the gene encoding yeast debranching enzyme. Cell. 1991 May 3;65(3):483-92. doi: 10.1016/0092-8674(91)90466-c.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 1977;12(1):1-8. doi:10.1016/0092-8674(77)90180-5

Conboy JG. Unannotated splicing regulatory elements in deep intron space. *Wiley Interdiscip Rev RNA*. 2021;12(5):e1656. doi:10.1002/wrna.1656

Corvelo A, Eyras E. Exon creation and establishment in human genes. *Genome Biol*. 2008;9(9):R141. doi:10.1186/gb-2008-9-9-r141

Coulon A, Ferguson ML, de Turris V, Palangat M, Chow CC, Larson DR. Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife. 2014;3:e03939. Published 2014 Oct 1. doi:10.7554/eLife.03939

Duff MO, Olson S, Wei X, et al. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature*. 2015;521(7552):376-379. doi:10.1038/nature14475

Dvinge H, Ries RE, Ilagan JO, Stirewalt DL, Meshinchi S, Bradley RK. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci U S A*. 2014;111(47):16802-16807. doi:10.1073/pnas.1413374111

Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, Diener K, Jones K, Fu XD, Bentley DL. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. 2014

Dec 1;28(23):2663-76. doi: 10.1101/gad.252106.114. PMID: 25452276; PMCID: PMC4248296.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc Natl Acad Sci U S A. 2005;102(45):16176-16181. doi:10.1073/pnas.0508489102

Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. Nucleic Acids Res. 2013;41:D48-D55. doi:10.1093/nar/gks1236

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

Gazzoli I, Pulyakhina I, Verwey NE, et al. Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biol*. 2016;13(3):290-305. doi:10.1080/15476286.2015.1125074

Gehlenborg N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. 2019. R package version 1.4.0, https://CRAN.R-project.org/package=UpSetR

Gehring NH, Roignant JY. Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends Genet*. 2021;37(4):355-372. doi:10.1016/j.tig.2020.10.008

Hatton AR, Subramaniam V, Lopez AJ. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Mol Cell*. 1998;2(6):787-796. doi:10.1016/s1097-2765(00)80293-2

Huranová M, Ivani I, Benda A, et al. The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. J Cell Biol. 2010;191(1):75-86. doi:10.1083/jcb.201004030

Iannone R, Cheng J, Schloerke B, Hughes E. *gt: Easily Create Presentation-Ready Display Tables*. 2022. R package version 0.7.0, https://CRAN.R-project.org/package=gt

Joseph B, Kondo S, Lai EC. Short cryptic exons mediate recursive splicing in Drosophila. Nat Struct Mol Biol. 2018;25(5):365-371. doi:10.1038/s41594-018-0052-6

Kassambara A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. 2022. R package version 0.5.0, https://CRAN.R-project.org/package=ggpubr

Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010;7(12):1009-1015. doi:10.1038/nmeth.1528

Ke S, Chasin LA. Context-dependent splicing regulation: exon definition, co-occurring motif pairs and tissue specificity. RNA Biol. 2011;8(3):384-388. doi:10.4161/rna.8.3.14458

Kelemen O, Convertini P, Zhang Z, et al. Function of alternative splicing. *Gene*. 2013;514(1):1-30. doi:10.1016/j.gene.2012.07.083

Kelly S, Georgomanolis T, Zirkel A, et al. Splicing of many human genes involves sites embedded within introns. *Nucleic Acids Res*. 2015;43(9):4721-4732. doi:10.1093/nar/gkv386

Khodor YL, Rodriguez J, Abruzzi KC, Tang CH, Marr MT 2nd, Rosbash M. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. 2011;25(23):2502-2512. doi:10.1101/gad.178962.111

Khodor YL, Menet JS, Tolan M, Rosbash M. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. RNA. 2012;18(12):2174-2186. doi:10.1261/rna.034090.112

Kim H, Klein R, Majewski J, Ott J. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*. 2004;36(9):915-917. doi:10.1038/ng0904-915

Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*. 2007;35(1):125-131. doi:10.1093/nar/gkl924

Larsson J. *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*. 2021. R package version 6.1.1, https://CRAN.R-project.org/package=eulerr.

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8): e1003118. doi:10.1371/journal.pcbi.1003118

Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem*. 2015;84:291-323. doi:10.1146/annurev-biochem-060614-034316

Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. Are snRNPs involved in splicing?. Nature. 1980;283(5743):220-224. doi:10.1038/283220a0

Martin RM, Rino J, Carvalho C, Kirchhausen T, Carmo-Fonseca M. Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. Cell Rep. 2013;4(6):1144-1155. doi:10.1016/j.celrep.2013.08.013

Matera AG, Wang Z. A day in the life of the spliceosome [published correction appears in Nat Rev Mol Cell Biol. 2014 Apr;15(4):294]. Nat Rev Mol Cell Biol. 2014;15(2):108-121. doi:10.1038/nrm3742

Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*. 2015;22(1):5-7. doi:10.1038/nsmb.2942

Mercer TR, Clark MB, Andersen SB, et al. Genome-wide discovery of human splicing branchpoints. *Genome Res*. 2015;25(2):290-303. doi:10.1101/gr.182899.114

Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*. 2012;338(6114):1593-1599. doi:10.1126/science.1228186

Meyer LR, Zweig AS, Hinrichs AS, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*. 2013;41:D64-D69. doi:10.1093/nar/gks1048

Mohanta A, Chakrabarti K. Dbr1 functions in mRNA processing, intron turnover and human diseases. *Biochimie*. 2021;180:134-142. doi:10.1016/j.biochi.2020.10.003

Moon S, Zhao YT. Recursive splicing is a rare event in the mouse brain. PLoS One. 2022;17(1):e0263082. Published 2022 Jan 28. doi:10.1371/journal.pone.0263082

Moore MJ. Intron recognition comes of AGe. *Nat Struct Biol*. 2000 Jan;7(1):14-6. doi: 10.1038/71207

Mount SM, Pettersson I, Hinterberger M, Karmas A, Steitz JA. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. Cell. 1983;33(2):509-518. doi:10.1016/0092-8674(83)90432-4

Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010;463(7280):457-463. doi:10.1038/nature08909

Oesterreich FC, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell. 2016;165(2):372-381. doi:10.1016/j.cell.2016.02.045

Padgett RA, Konarska MM, Grabowski PJ, Hardy SF, Sharp PA. Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science*. 1984;225(4665):898-903. doi:10.1126/science.6206566

Pai AA, Paggi JM, Yan P, Adelman K, Burge CB. Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLoS Genet*. 2018;14(8):e1007588. Published 2018 Aug 27. doi:10.1371/journal.pgen.1007588

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing [published correction appears in Nat Genet. 2009 Jun;41(6):762]. *Nat Genet*. 2008;40(12):1413-1415. doi:10.1038/ng.259

Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev*. 2018;32(7-8):577-591. doi:10.1101/gad.312058.118

Pineda JMB, Nicholas TR, Bradley RK. Most human genes express intron-derived circular RNAs. *Nat. Struct. Mol. Biol*. Under review.

Pulyakhina I, Gazzoli I, Hoen PA, et al. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing [published correction appears in Nucleic Acids Res. 2015 Dec 15;43(22):11068. den Dunnen, Johan [Corrected to den Dunnen, Johan T]]. *Nucleic Acids Res*. 2015;43(12):e80. doi:10.1093/nar/gkv242

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radtke M, Srndic I, Schroeder R. Genome-wide identification of intrasplicing events in the human transcriptome and hints to their regulatory potential. *bioRxiv*. 2017. doi:10.1101/159350

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47. doi:10.1093/nar/gkv007

Roca X, Krainer AR, Eperon IC. Pick one, but be quick: 5' splice sites and the problems of too many choices. Genes Dev. 2013;27(2):129-144. doi:10.1101/gad.209759.112

Rogalska ME, Vivori C, Valcárcel J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects [published online ahead of print, 2022 Dec 16]. Nat Rev Genet. 2022;10.1038/s41576-022-00556-8. doi:10.1038/s41576-022-00556-8

Roy B, Haupt LM, Griffiths LR. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. Curr Genomics. 2013;14(3):182-194. doi:10.2174/1389202911314030004

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

Ruskin B, Green MR. An RNA processing activity that debranches RNA lariats. *Science*. 1985;229(4709):135-140. doi:10.1126/science.2990042

Schmidt U, Basyuk E, Robert MC, et al. Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. J Cell Biol. 2011;193(5):819-829. doi:10.1083/jcb.201009012

Shepard S, McCreary M, Fedorov A. The peculiarities of large intron splicing in animals. *PLoS One*. 2009;4(11):e7853. Published 2009 Nov 16. doi:10.1371/journal.pone.0007853

Sibley CR, Emmett W, Blazquez L, et al. Recursive splicing in long vertebrate genes. *Nature*. 2015;521(7552):371-375. doi:10.1038/nature14466

Sterner DA, Carlo T, Berget SM. Architectural limits on split genes. Proc Natl Acad Sci U S A. 1996;93(26):15081-15085. doi:10.1073/pnas.93.26.15081

Suzuki H, Kameyama T, Ohe K, Tsukahara T, Mayeda A. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. *FEBS Lett*. 2013;587(6):555-561. doi:10.1016/j.febslet.2013.01.057

Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol*. 2012;19(7):719-721. Published 2012 Jun 17. doi:10.1038/nsmb.2327

Taggart AJ, Lin CL, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res*. 2017;27(4):639-649. doi:10.1101/gr.202820.115

Tilgner H, Knowles DG, Johnson R, 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. 2012;22(9):1616-1625. doi:10.1101/gr.134445.111

Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. Wiley Interdiscip Rev RNA. 2013;4(1):61-76. doi:10.1002/wrna.1141

Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell*. 2019;76(2):329-345. doi:10.1016/j.molcel.2019.09.017

Vogel J, Hess WR, Börner T. Precise branch point mapping and quantification of splicing intermediates. *Nucleic Acids Res*. 1997;25(10):2030-2031. doi:10.1093/nar/25.10.2030

Wan Y, Anastasakis DG, Rodriguez J, et al. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*. 2021;184(11):2878-2895.e20. doi:10.1016/j.cell.2021.04.012

Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476. doi:10.1038/nature07509

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016

Wilkinson ME, Charenton C, Nagai K. RNA Splicing by the Spliceosome. *Annu Rev Biochem*. 2020;89:359-388. doi:10.1146/annurev-biochem-091719-064225

Woodward LA, Mabin JW, Gangras P, Singh G. The exon junction complex: a lifelong guardian of mRNA fate. Wiley Interdiscip Rev RNA. 2017;8(3):10.1002/wrna.1411. doi:10.1002/wrna.1411

Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature. 1999;402(6763):832-835. doi:10.1038/45590

Wong MS, Kinney JB, Krainer AR. Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. Mol Cell. 2018;71(6):1012-1026.e3. doi:10.1016/j.molcel.2018.07.033

Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-94. doi: 10.1089/1066527041410418.

Zhang C, Hastings ML, Krainer AR, Zhang MQ. Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *Proc Natl Acad Sci U S A*. 2007;104(38):15028-15033. doi:10.1073/pnas.0703773104

Zhang XO, Fu Y, Mou H, Xue W, Weng Z. The temporal landscape of recursive splicing during Pol II transcription elongation in human cells. *PLoS Genet*. 2018;14(8):e1007579. doi:10.1371/journal.pgen.1007579

VITA

Emma R. Hoppe was born in Lincoln, NE. From 2011 to 2015, she attended Creighton University in Omaha, where she graduated *summa cum laude* and with the Honors distinction with the degree of Bachelor of Science in Biology. Emma began her Ph.D. in Genome Sciences at the University of Washington in September 2016. Her graduate work with Rob Bradley at the Fred Hutchinson Cancer Center focuses on deepening our understanding of splicing mechanisms. During her time at the University of Washington and Fred Hutchinson, she was awarded Mentor of the Month for her work with the Girls Who Code club at Fred Hutchinson. She also served on the Genome Sciences Curriculum Committee as the student representative.