Role of Quaking (QKI) protein in the regulation of alternative splicing and circular RNA biogenesis: a computational analysis

by

Thomas Hamilton Lipscomb

Mentor: Dr. Kristin C. Gunsalus

A thesis in fulfillment of the Masters in Biology Degree

Department of Biology

New York University

September 2018

_____

Dr. Kristin C. Gunsalus

Acknowledgements

I am heartily thankful to my advisor, Dr. Kris Gunsalus, whose encouragement, guidance and support during the project provided me with excellent preparation for my career, as well as for providing constructive feedback on my thesis drafts. I also extend my sincere thanks to Dr. Hin Hark Gan for finding the NUMB paper on which my 3D modeling is based, creating the 3D modeling script, insightfully answering my questions, and his great help editing my thesis rough drafts. To Alan Twaddle, thank you for finding and explaining bedtools closest. To Dr. Brian Parker, thank you for showing me how to create a histogram in R Studio. I especially thank Dr. Shenglong Wang and Dr. Mohammed Khalfan for debugging my use of NGSPlot.

Abstract

Both alternative splicing of mRNAs and circular RNA biogenesis are important for basic cellular functions. To gain insight into the genome-wide regulatory potential of QKI-5, I have examined QKI-5 binding preferences using 3D structural modeling and transcriptome-wide QKI-5 binding patterns obtained from PAR-CLIP. Energy calculations predict that QKI-5 binds more favorably than SF1 at 10 out of 16 known SF1 recognition motifs. Analysis of PAR-CLIP data reveals that QKI-5 binding sites are preferentially located in introns flanking all exons but are not enriched in introns flanking either cassette exons or circular RNAs. QKI-5 binding sites are also enriched at UTR-CDS junctions and at transcription start and end sites. These results support a broad regulatory role for QKI-5 in exon skipping by competition with SF1 binding and suggest that QKI has additional roles in mRNA processing or translational regulation across the transcriptome.

Table of Contents

Introduction

Understanding exon skipping and back-splicing

Eukaryotic RNA Polymerase II begins RNA transcription of the precursor mRNA (pre-mRNA) at a gene promoter, then continues through the 5' UTR, through alternating CDS and introns, through the 3' UTR, and ends a few bp or a few thousand bp past the cleavage site that defines the end of the transcript. The primary transcript is cleaved downstream of the CDS and a poly(A) tail is added at the 3' end to guard from degradation, assist in export to the cytoplasm, and the poly(A) tail aids the binding of proteins involved in initiating translation (lumen, 2013b). The transcribed region beyond the cleavage site is then digested by a 5'-exonuclease (lumen, 2013a). The cleavage and polyadenylation site is determined by a PAS (polyA addition signal) sequence, AAUAAA, located around 20 nt upstream, and a GU-rich sequence around 40-60 nt downstream of the PAS site. Co-transcriptionally, the introns of the pre-mRNA are spliced out and a trimethylguanosine cap is added to the 5' end of the transcript to guard from degradation and act as a recognition site in the initiation of translation. A mature mRNA transcript consists of a 5' cap, a 5' UTR, a coding region, a 3' UTR, and a poly(A) tail (lumen, 2013b).

Splicing is regulated by both cis-elements (ESE, ESS, ISS, and ISE) and trans-acting splicing factors (SR proteins, hnRNP, and unknown factors) (Sanford et al., 2003; Wang, Z. and Burge, 2008). Splice signals are spread throughout the intron and include the splice donor site, splice acceptor site, branch point sequence, and other recognition motifs (Figure 1).

Introns are recognized by the splicing machinery (the spliceosome), comprising 52 RNA binding proteins (RBPs) and numerous small RNAs (snRNAs) (Xiaofeng Zhang et al., 2018). Constitutive splicing (Figure 2A) generates mature transcripts that include all exons within a gene. When the introns of the pre-

mRNA are spliced out during constitutive splicing, the U1 snRNP (Figure 2) binds to the splice donor site (Clancy, 2008), SF1 recruits U2AF to the branch point sequence, which recruits the spliceosome (Selenko et al., 2003), then the splice donor site (Figure 1) base pairs with the downstream branch point sequence, forming a lariat. The splice donor site is ligated to the branch point site (Figure 1) by a hydroxyl group at the splice donor site that attacks the phosphodiester bond at the splice acceptor site (Figure 1) (Clancy, 2008). The intron is released as a lariat that is then degraded and the exons are joined (Black, 2016).

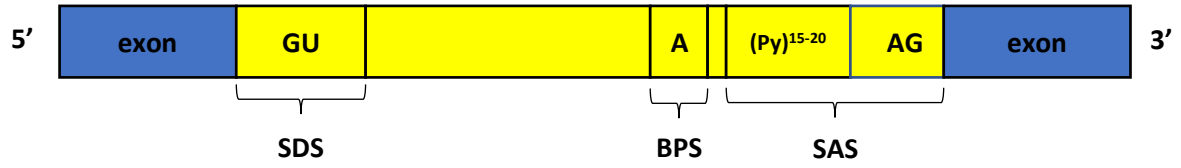

Figure 1: Structure of the intron (intron, yellow; exon, blue). SDS is the Splice Donor Site, BPS is the Branch Point Sequence, and SAS is the Splice Acceptor Site. The almost invariant GU-rich sequence is part of the splice donor site. The A is an adenine nucleotide, which is the branch point sequence. The $(Py)^{15-20}$ is a polypyrimidine tract usually 15–20 base pairs long, about 5–40 base pairs before the 3' end of the intron to be spliced (Lodish et al., 2004). The almost invariant AG sequence is part of the splice acceptor site. Image adapted from (Bhagat, 2013).

Alternative splicing (AS), in which some splice signals are not used, can generate multiple mRNA variants (isoforms) from the same gene that may have different biological functions in the cell. AS is pervasive in metazoans (Skandalis et al., 2010), and between 35% and 59% of human genes are predicted to output alternatively spliced mRNA isoforms (Wu et al., 2002). Shifts of about 15% to 50% of the alternative splicing landscape are common in human diseases, depending on whether only the well-conserved or also the regulatory regions are included (Sveen et al., 2016).

AS is mediated by the spliceosome and regulatory factors that selectively employ a subset of splicing signals to produce diverse mRNA isoforms. The types of linear alternative splicing are skipping of the

cassette exon (Figure 2B), mutually exclusive exons, intron retention, alternative 5' or 3' splice sites (alternative splice donor site and alternative splice acceptor site), alternative promoters, alternative splicing and polyadenylation (see Figure 1 of (Girard et al., 2007)), exon retention, exon extension, and exon truncation (see Figure 3 of (Cha et al., 2008)). Recently it was discovered that back-splicing of RNAs can also occur to generate circular RNAs (circRNAs) (Figure 3), and that these have important roles in gene regulation, particularly in neurons (Chen, L., 2016). Exon-intron circular RNA (EIciRNA) can regulate transcription of their parental genes, and circular RNAs can regulate splicing of their linear cognates (Li, X. et al., 2018). The processing of circular RNAs can affect the splicing of their linear mRNA counterparts (Li, X. et al., 2018). They also can act as miRNA sponges and can act through associated proteins (Li, X. et al., 2018). They can also be translated and are resources for derivation of pseudogenes (Li, X. et al., 2018). This project focuses on computational analysis of exon skipping (Figure 2B) and back-splicing (Figure 3).

In constitutive splicing (Figure 2A), during the first step of spliceosome assembly, the U1 complex interacts with the splice donor site on the intron of the precursor mRNA (pre-mRNA) (E complex) (Shimoyama et al., 2015). Then SF1 binds to the branch point sequence (BPS) (Zong et al., 2014), spliceosome assembly continues, and the spliceosome then removes the intron. In exon skipping, QKI-5 outcompetes SF1 at the splice acceptor site, preventing spliceosome assembly there (for example near the start of Exon 2 in Figure 2B). This forces a downstream acceptor to be used instead, resulting in exclusion of the cassette exon from the mature transcript along with both flanking introns. Although the molecular mechanism is unknown, a plausible hypothesis is that this causes the SR protein to bridge the splice donor and acceptor sites of the exons upstream and downstream of the cassette exon.

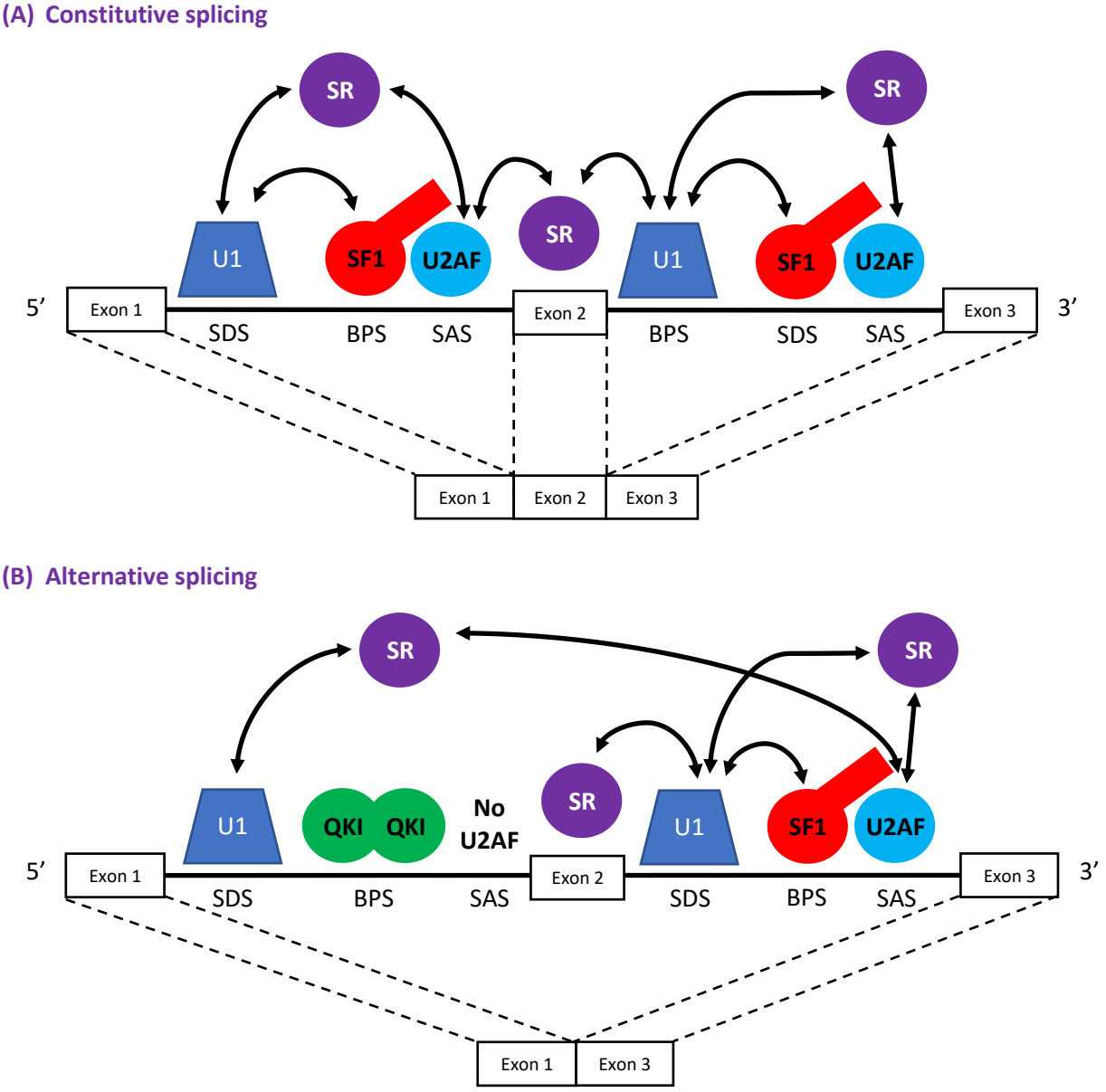

Figure 2: Splicing mechanisms. (A) Constitutive splicing (adapted from (Ohe and Mayeda, 2010)). (B) Proposed mechanism for QKI-5-mediated exon skipping. (A) SR family proteins have an inherent flexibility that allows them to function at several steps in spliceosome assembly and regulation of splice site selection (Sanford et al., 2003). The interaction of the two intronic SRs with U2AF and the U1 spliceosomal factor causes intron bridging (Ohe and Mayeda, 2010), leading to splicing out of the intron. The interaction of SR bound to the exon with U2AF and U1 causes exon definition (Ohe and Mayeda, 2010), where basal splicing machinery is placed between the exons (Ram and Ast, 2007). (B) When QKI-5 outcompetes SF1, U2AF is not recruited because QKI does not have a domain capable of interacting with U2AF (see Figure 4). The current proposed scenario is that SR attaches to the next downstream U2AF, causing exon skipping. Abbreviations: Ser/Arg-rich protein (SR), small nuclear ribonucleoprotein U1, U2 Auxiliary Factor (U2AF), Splicing Factor 1 (SF1), Quaking (QKI), and Branchpoint Sequence (BPS). Arrows indicate protein-protein interactions.

SF1 has a tail domain that recruits U2AF and both play a crucial role in pre-mRNA splice acceptor site recognition for the spliceosome (Shimoyama et al., 2015). Experimental studies have shown that QKI-5 outcompetes SF1 about half the time at a site upstream of exon 12 in NUMB oncogene (Zong et al., 2014). In that case U2AF fails to be recruited because QKI lacks a U2AF binding tail (Figure 4).

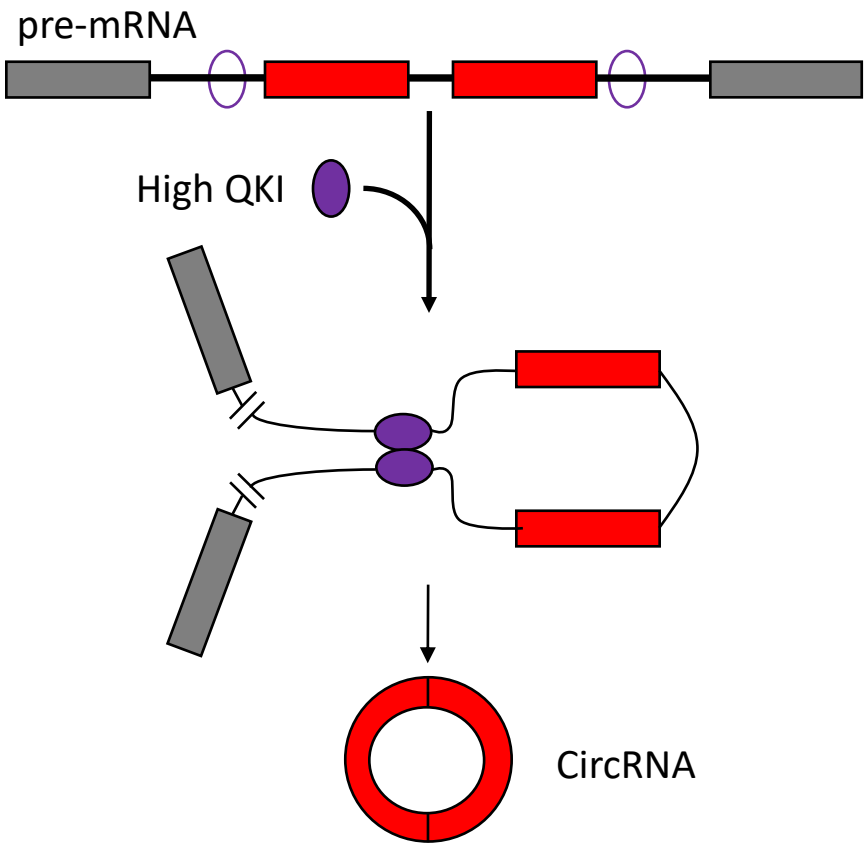

Figure 3: Model of circular RNA biogenesis induced by QKI-5 (adapted from (Conn et al., 2015)). Putative QKI-5-binding sites inserted in the introns flanking the red exons induced de novo circular RNA formation, suggesting that dimerization of QKI-5 bound to flanking introns enables circular RNA formation, most likely by juxtaposition of the start and end of the circular RNA (Conn et al., 2015). Circularized products that contain introns are referred to in the literature as exon-intron circular RNA (EIciRNA); products with only exons are referred to in the literature as circRNA (Shao and Chen, 2016). Shown here is circRNA.

Structure and function of mammalian STAR family proteins: QKI and SF1

The three major QKI mRNA isoforms are named for the length of their 3' UTR: QKI-5 (5kb), QKI-6 (6kb) and QKI-7 (7kb) (Wang, Y. et al., 2013). All three protein isoforms have the Signal Transduction and Activation of RNA (STAR) domain (QUA1-KH-QUA2) (Figure 4) and their protein sequence differs only in the C-terminus. The QKI-5 protein is localized to the nucleus, while QKI-6 and QKI-7 are cytoplasmic. This project focuses on QKI-5 because it regulates splicing and RNA metabolism in the nucleus, while QKI-6 and QKI-7 other regulatory functions (Artzt et al., 2010).

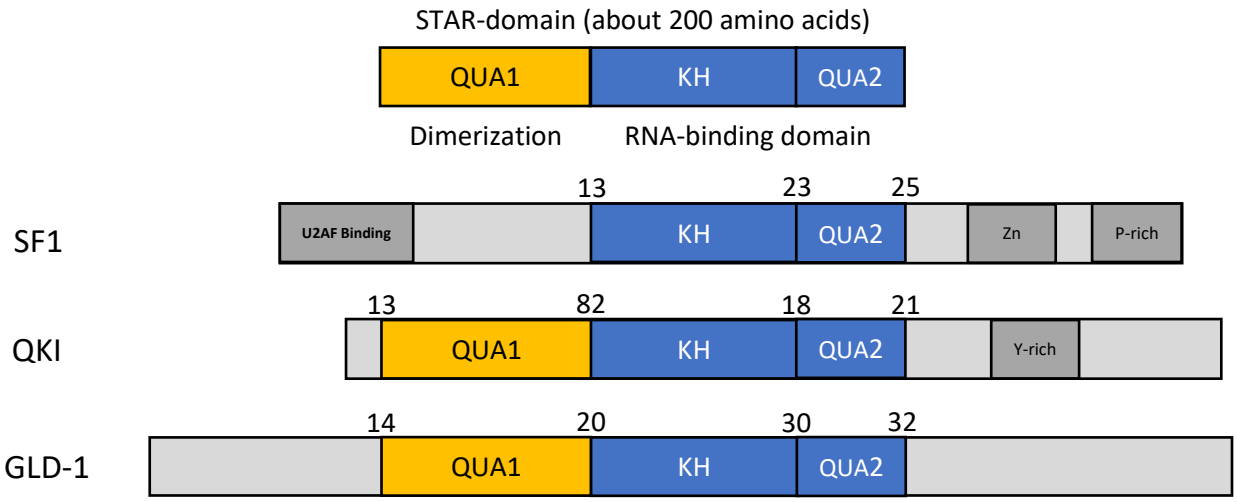

Figure 4: Signal Transduction and Activation of RNA (STAR) proteins SF1, QKI, and GLD-1 adapted from (Artzt et al., 2010). The diagram for QKI is valid for all 3 QKIs (QKI-5, QKI-6, and QKI-7), which differ in the tail region length, the length of which is not specified in the figure. GLD-1 is the C. elegans homolog of the vertebrate protein QKI. All QKI isoforms contain QUA1, KH, QUA2, and the Y-rich domains, but differ in their C-terminus (Artzt et al., 2010). RNA-binding domains in STAR homologs are KH and QUA2. The KH domain recognizes the BPS, the QUA2 domain recognizes the 5' (U/C)AC moiety of the RNA, and the QUA1 domain in GLD-1 and QKI can form homodimers (Feracci et al., 2016). SF1 does not contain a QUA1 domain and so cannot form dimers. SF1 has an N-terminal tail that binds to U2AF in the early spliceosome complex (Selenko et al., 2003).

QKI has biological functions other than its role in regulating splicing: cytoplasmic QKI isoforms (QKI-6 or QKI-7) may stabilize mRNA (Artzt et al., 2010), regulate translation (e.g. QKI-6 can repress translation

(McInnes and Lauriat, 2006)), (isoform unspecified) direct nucleocytoplasmic localization (Artzt et al., 2010), and (isoform unspecified) inhibit microRNA maturation by binding to the primary RNA transcript that generates the microRNA (pri-microRNA) (Morlando et al., 2008) and then sequestering the pri-miRNA in the nucleus to prevent further processing (Darbelli and Richard, 2016). The QKI gene is named for the "quaking" phenotype produced in an experiment that resulted in mice born in a state of demyelination (Li, Z. et al., 2000). Quaking was first observed in a female autosomal recessive mutant mouse from the DBA/2J strain (Sidman et al., 1964). A later experiment found a molecular basis for the quaking phenotype (Li, Z. et al., 2000). The 3' UTR of the myelin basic protein (MBP) mRNA was removed, which prevented QKI from binding to the 3' UTR (Li, Z. et al., 2000). The MBP 3'UTR plays important roles in stabilization and localization of MBP mRNAs to the myelin sheath (Li, Z. et al., 2000). The absence of stabilization and localization caused demyelination, resulting in quaking phenotype in the newborn mice (Li, Z. et al., 2000).

SF1 is another member of the STAR family proteins with significant structural and functional differences from the QKI proteins. SF1 binds to the consensus sequence YNCURAY (Y is C or T, R is A or G, N is any base), determined experimentally by using a chloramphenicol acetyltransferase (CAT) reporter assay to indirectly measure the efficacy of SF1 interaction with branchpoint sequences containing the canonical UACUAAC motif and single nucleotide variants (Peled-Zehavi et al., 2001). The strength of SF1-branchpoint sequence interactions determines how well they promote the activity of CAT, which is measured by how much acetylation of chloramphenicol occurs. Based on the experimentally inferred binding motif YNCURAY, my research selected a truncated 6-nt NCURAY as the SF1 motif for which QKI-5 and SF1 binding affinities were calculated (Figure 13), because my 3D modeling found the first Y to not be near the protein.

In another experimental approach, SELEX analysis (systematic evolution of ligands by exponential enrichment) identified the QKI recognition motif is as ACUAACN$_{1–20}$UAAC (Richard and Galarneau, 2005). The SELEX procedure progressively selects, from a large combinatorial oligonucleotide library, DNA or RNA (in this case RNA) ligands with variable binding (in this case to QKI) affinities and specificities by repeated rounds of enrichment of sequences bound by the protein of interest (Wilson and Szostak, 1999). The SELEX-derived QKI recognition motif comprises two sites (ACUAACN and UAAC) because QKI is a homodimer, and thus contains two KH-QUA2 RNA binding domains, which bind to the first full site (ACUAACN) and, 1-20 nt downstream, a second site that can either be another full site (ACUAACN) or a half site (UAAC) (Beuck et al., 2012). Using the same SELEX data (Richard and Galarneau, 2005), another work proposed a more general QKI homodimer recognition motif as NACUAAY-N(1-20)-UAAY (de Miguel et al., 2015). The X-ray structure shows that the QKI homodimer is capable of recognizing two RNA sequence elements present on separate RNAs or on a single RNA chain with a >10nt linker between the start and the end of the loop (Teplova et al., 2013). The SF1 and QKI recognition motifs (excluding the half-site) have the same length (seven nucleotides) and similar sequences because SF1 and QKI both contain the KH and QUA2 domains. The project I undertook here computationally tests competition between QKI and SF1 at the SF1 binding motif NCURAY, by using 3D modeling of the sixteen possible nucleotide combinations bound to either QKI or SF1 to compute their binding affinities.

The structure of QKI (Figure 5) shows substantial similarity to other STAR proteins (Teplova et al., 2013). The QUA2 domain recognizes the nucleotides at the 5’ end of the BPS RNA, while the KH domain recognizes the nucleotides at the 3’ end of the BPS RNA (Artzt et al., 2010). The QUA1 dimerization domain shows roughly perpendicular helix–turn–helix folds, forming a symmetric dimer, stabilized mainly by hydrophobic interactions (Teplova et al., 2013). The hydrophobic ‘‘zipper’’ residues stabilizes the QUA1 fold, and a hydrogen bond between a conserved Tyr and a Glu side chain, previously shown to

be essential for homodimerization *in vivo*, maintains the three-helix bundle dimer interface. The dimer interface is further stabilized by a salt bridge (Teplova et al., 2013).

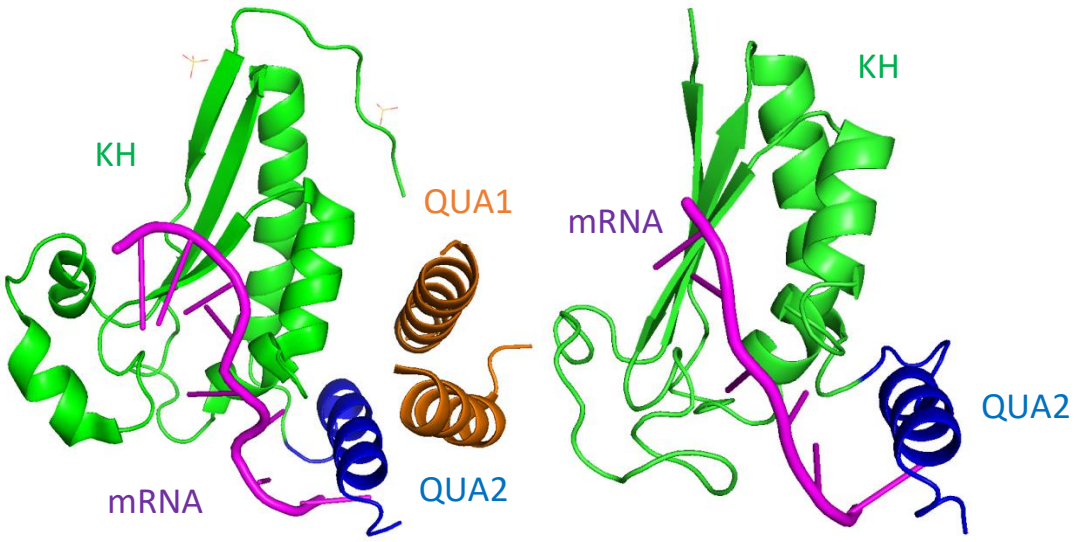

Figure 5: Solved QKI (4JVH, left) and SF1 (1K1G, right) structures (color code: QUA1, orange; KH, green; QUA2, blue: and mRNA (AUUAAC), magenta). See Figure 4 for domain coordinates.

Role of QKI in alternative splicing

QKI is known to regulate AS in several genes. Two examples are exon skipping in NUMB (Figure 6) (Zong et al., 2014) and mutually exclusive exons in macroH2A1 (Novikov et al., 2011). Experimental QKI-5 knockout in mouse cells causes exon 12 to be included more often in NUMB mRNA isoforms, which is associated with increased Notch signaling and increased proliferation (described in more detail below in Figure 6) (Zong et al., 2014). QKI also regulates mutually exclusive exon inclusion of the macroH2A1 pre-mRNA into isoforms macroH2A1.1 and macroH2A1.2 in human cancer cells. macroH2A1 is a variant of H2A, one of the core histone proteins that are responsible for the structure of the nucleosome in the chromosomal fiber in eukaryotes (Costanzi and Pehrson, 1998; NCBI Gene, 2018). QKI enhances expression of macroH2A1.1, which contains all exons except exon 6. In cancer cells, decreased QKI levels

promote generation of macroH2A1.2, which contains all exons except exon 7 (Novikov et al., 2011). Thus, exons 6 and 7 are mutually exclusive in these isoforms. It is also likely that QKI acts a regulator of AS in other cancers (Zong et al., 2014).

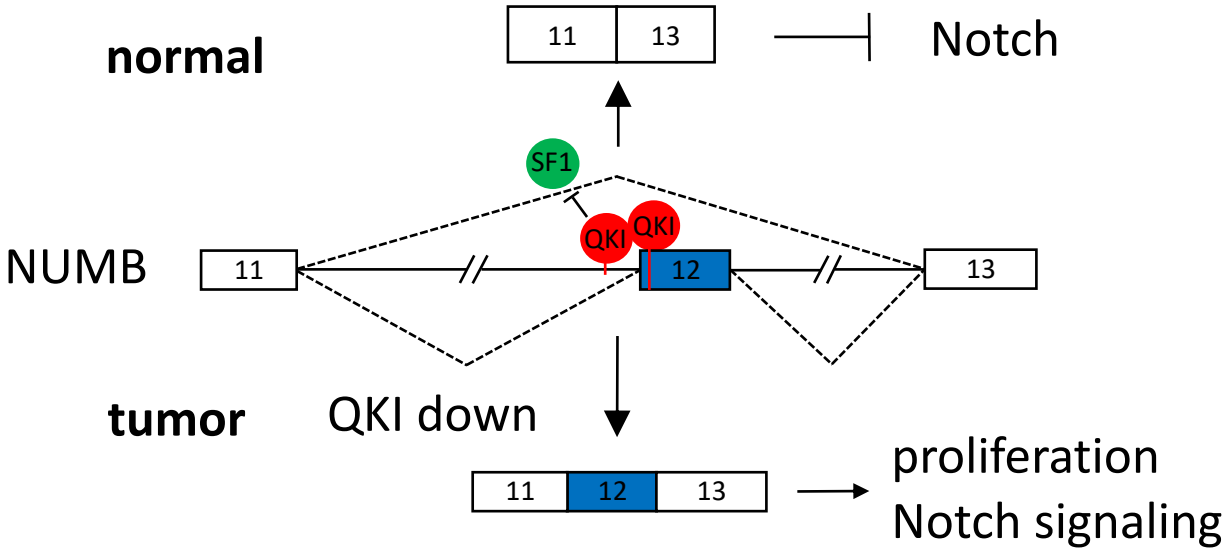

Figure 6: A model for QKI-5-mediated control of exon skipping and its effect on cell proliferation in HeLa and A549 cells adapted from (Zong et al., 2014). In this model, QKI-5 normally outcompetes SF1, leading to exon skipping of exon 12 and inhibition of Notch signaling. In tumors, when QKI levels are reduced or downregulated, so SF1 continues constitutive splicing, exon 12 is included, and Notch signaling leads to increased proliferation.

Role of QKI in circular RNA biogenesis

Circular RNAs are produced from pre-mRNA by back-splicing (Figure 3) (Chen, L., 2016) and were thought to be exceptional curiosities or 'splicing noise' until 2012, when a statistical analysis of RNA-Seq data found that circular RNAs are pervasive in the transcriptome (Salzman, 2016; Sebastian Memczak et al., 2013) . Circular RNAs are ubiquitous in the eukaryotic kingdom and particularly in the genes of metazoans (Salzman, 2016). Approximately 22.5% of the human transcriptome produces circular RNAs (Conn et al., 2015). While they are generally expressed at low levels, circular RNAs show cell type- and tissue-specific expression that drives differences in the transcriptome and thus the proteome of cells

and tissues (Rybak-Wolf et al., 2015). Known circular RNA functions include titrating microRNAs, regulating transcription, and interfering with AS (Chen, L., 2016). However, the function of most circular RNAs is unknown (Chen, L., 2016). Hundreds of circular RNAs are regulated during human Epithelial to Mesenchymal Transition (EMT), and over one-third of abundant circular RNAs are dynamically regulated by QKI-5 (Conn et al., 2015). QKI-5 is regulated during EMT, which suggests that QKI-5 is involved in differentiation (Conn et al., 2015). One reason that EMT matters is that many cancers arise from epithelial cells and need to undergo EMT to become invasive and to metastasize, so targeting EMT through QKI-5 may be a novel cancer treatment (Conn et al., 2015).

Role of QKI in alternative splicing aberrations in cancer

Tumor initiation (tumorigenesis) and tumor progression can be caused by changes in AS (Climente-González et al., 2017). The importance of QKI-5 in cancer has been demonstrated in several ways: First, cancer-associated splicing factors QKI and RBFOX2 have been found to regulate ∼20% of the alternative splicing events associated with the ovarian tumor microenvironment (Brosseau et al., 2014). Second, QKI-5 regulates over one-third of abundant circular RNAs during EMT (Conn et al., 2015). Third, experiments have shown that knockdown of QKI in poorly tumorigenic human Hs683 cell line may form tumors (Chen, A. et al., 2012). Fourth, tumor initiation rates, tumor sizes and lung metastasis rate were increased by knocking down QKI in tumor implanting nude mice model (Lu et al., 2014). Conversely, QKI overexpression prolonged G1 phase and shortened both S and G2/M phase, increasing the time between cell divisions, thereby decreasing proliferation without affecting cell apoptosis (Lu et al., 2014). Fifth, the QKI-5 gene is deleted from the chromosome in cancers such as glioblastoma multiforme (Novikov et al., 2011).

Key Questions

This work aims to answer two key questions regarding AS regulation. First, how well does QKI-5 compete with SF1 at the RNA motif to promote exon skipping? Previous experiments on the NUMB oncogene suggest that QKI-5 outcompetes SF1 by binding more tightly at or near the branch point sequence and prevents spliceosome assembly by blocking SF1 binding (Zong et al., 2014). Second, can one detect overlap of QKI-5 and SF1 binding sites within regulatory regions that control exon skipping and back-splicing that are consistent with the current model? Previous experiments have shown that insertion of QKI-5 binding motifs into flanking introns causes de novo circular RNA formation from otherwise linearly spliced transcripts and that removal of preexisting QKI-5 binding motifs from flanking introns decreases circular RNA formation, but the molecular mechanism has not been demonstrated (Conn et al., 2015). Answering these questions will contribute to understanding the molecular mechanisms of AS.

To answer these two key questions, I employed two complementary computational methods. To answer the first question, I used 3D modeling as an approach to gain molecular-level understanding of SF1 and QKI-5 binding at NUMB intron/exon 12 junction and QKI and SF1 binding affinity at 16 variants of the SF1 motif. Specifically, structural modeling can provide information about competitive binding of these RNA binding proteins as well as the effects of mutations at the binding site. To answer the second question, I used computational transcriptome screening of QKI-5 binding sites from PAR-CLIP data to provide evidence for overlap of QKI-5 with regulatory regions to infer where QKI-5 might regulate exon skipping and circular RNA formation on a genomic scale.

Summary of contributions

This work contributes in several ways to understanding the role of QKI-5 in alternative splicing and circular RNA biogenesis with implications for human disease.

In answer to the first key question above, our structure modeling is consistent with the idea that QKI-5 outcompetes SF1 and mutations in the NUMB exon 12 branch-point sequence decrease QKI-5 and SF1 binding affinity. Also, it was calculated that at the SF1 motif NCURAY QKI-5 outcompetes SF1 in 10 out of 16 known SF1 recognition motifs (Figure 13). In answer to the second key question above, our transcriptome-wide search using known QKI binding sites showed that QKI binds within introns near both splice donor site and splice acceptor site (Figure 14), based on analysis of a dataset that includes all human exons and introns. For the dataset of skipped human exons (which constitute a minor subset of all exons), the QKI sites are also predominantly in the intron side of the intron/exon junctions, consistent with the role of QKI as an AS regulator (Figure 16). For human circular RNAs, one might expect that there would be QKI binding peaks at the start and end of the circular RNA because it was hypothesized by (Conn et al., 2015) that QKI juxtaposes two sites along a pre-mRNA to promote circular RNA formation (Figure 3). I found that there was a QKI binding peak inside the start of the circular RNA as expected, but surprisingly no peak of QKI binding at the end of the circular RNA, due partly to the low numbers of sites found (Figure 20). I found other observed peaks of QKI binding at the 5' Untranslated Region-Coding Sequence (UTR-CDS) junction and the CDS-3' UTR junction. The peak of QKI binding at the 5' UTR-CDS junction may support that QKI regulates translation (e.g. QKI-6 can repress translation (McInnes and Lauriat, 2006)) and the peak of QKI binding at the CDS-3' UTR junction could imply that QKI stabilizes mRNA such as MBP by binding to the 3' UTR (Li, Z. et al., 2000).

Methods

Structural modeling of competing QKI and SF1 binding to NUMB

I used 3D modeling to simulate SF1-QKI-5 competition at the splice acceptor site of exon 12 of the NUMB transcript and to measure SF1-QKI-5 competition at the SF1 motif NCURAY. I used a modeling paradigm previously developed to study miRNA-target interactions (Gan and Gunsalus, 2013; Gan and Gunsalus, 2015) to refine the structure, and then compute the binding free energy between two interacting macromolecules. The binding energy is computed as follows:

$$\Delta G_{\mathrm{total}} = \Delta G_{\mathrm{nonelec}} + \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{entropic}}$$ Equation 1: From (Gan and Gunsalus, 2013).

$\Delta G_{\mathrm{nonelec}}$ is the sum of the change in van der Waals energy and solvation energy upon complex formation; ΔG_{elec} is the change in electrostatic energy, including the effects of ions; and $\Delta G_{\mathrm{entropic}}$ is the change in entropy calculated from the loss of translational, rotational, and vibrational motions.

For input structures, I used the solved x-ray crystal structures of QKI-5 (4JVH) (Teplova et al., 2013) and SF1 (1K1G) (Liu et al., 2001). Both PDB files have QKI-5 and SF1 monomers bound to mRNA. These complexes can provide templates for modeling STAR protein-RNA interactions. To set up molecular simulations, I employed the following *in silico* steps using the Pymol (Schrödinger, 2018) and TINKER (Ponder, 2015) software packages: (1) Convert the eight QKI-5 MSE (selenomethionine) residues, used for crystallography, back to MET (methionine) residues using pymol mutate_bases function. (2) Truncate the two template RNA structures to retain only the nucleotides interacting with the protein (ACUAAC). (3) Convert the trimmed RNA template to the target RNA sequences at the NUMB locus. Specifically, I used mutate_bases in the Pymol command line tool iteratively to mutate the QKI-5 and SF1 intron

binding site from the solved structure ACUAAC to both the four templates of interest (two in the intron and two in exon 12) required for Figure 9 and Figure 10 (QKI-5 and SF1 competition at the splice acceptor junction at the start of exon 12 of NUMB) and also the sixteen templates of interest required for Figure 13 (QKI-5 and SF1 competition at the SF1 motif). The four templates of interest at exon 12 of NUMB are the intronic binding site AUUAAC (WT, mut1) and AUUGAC (mut2) and the QKI-5 exon binding site GCUAAU (WT) and GCGAGU (mut1, mut2) (see Figure 9). (4) Refine the modified complexes with the new target RNA sequences using an energy minimization algorithm with AMBER99 force field (implemented by the minimize command in the TINKER molecular package). The RMS gradient convergence parameter was set to 10^{-5} kcal/mol per Å. (5) Finally, compute the protein-RNA binding affinity using Equation 1, as described above. The simulations in steps 4 and 5 were executed using a shell script written by Dr. Hin Hark Gan to invoke the component software packages and was run on the NYU HPC Linux cluster Prince.

Sources of biological data

	QKI PAR-CLIP	circular RNA	exons	cassette exon	5' UTR	3' UTR	introns
# Transcripts	12024	91534	1140002	74081	293586	291950	957567
Median	32	8266	128	111	112	142	1595
1st Quartile	28	2055	87	74	64	87	505
3rd Quartile	32	24292	194	157	197	309	4605

Table 1: Distribution of region lengths annotated in the human genome (medians, first quartile, and third quartile) for transcripts. QKI binding sites identified by PAR-CLIP were published by (Hafner et al., 2010) and is aligned to human genome 19 (hg19). The circular RNA BED file was published in the circBase database and is aligned to hg19. Exons, introns, 5' UTR, and 3' UTR BED files were obtained using the UCSC Table Browser and are based on human genome 38 (hg38). The cassette exon BED file was published in the HEXEvent database and is aligned to hg38.

Human QKI binding sites in the human transcriptome

QKI target sites in human transcripts have been determined using photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) experiments in HEK295 cells (Hafner et al.,

2010). PAR-CLIP is a biochemical method for identifying the binding sites of RNA-binding proteins. Cells incorporate 4-thiouridine (4SU), a ribonucleoside analogs that is photoreactive, into their RNA, then cells are irradiated by ultraviolet light to cross-link the entire RNA population that binds to the RNA-binding protein. Then the population of RNAs bound to QKI is pulled down by immunoprecipitation. During cDNA generation, preferential base pairing of the 4SU-crosslinked product to a guanine instead of an adenine results in a thymine (T) to cytosine (C) transition in the PCR-amplified sequence, serving as a diagnostic mutation at the site of contact of the RNA-binding protein (Corcoran et al., 2011). The sequence reads for the QKI targets are 20-50 nt long, most being 32 nt long, but can be as much as 200 nt long. The reads contain the QKI binding sites near or at the T to C transition. The QKI PAR-CLIP was applied to the entire cell and QKI binds to both pre-mRNA and mRNA (Hafner et al., 2010). The QKI PAR-CLIP data was downloaded from http://dorina.mdc-berlin.de/regulators.

Human circular RNA transcripts

Human (hg19) circular RNA transcripts (Glažar et al., 2014) were downloaded from circBase http://www.circbase.org/cgi-bin/downloads.cgi. The purpose of finding QKI binding sites near the start and end of circular RNAs was to find more candidate sites where QKI binding to flanking introns aids circular RNA formation, possibly when QKI juxtaposes the start and end of the circular RNA (Figure 3). The circular RNA BED file was generated from experimental measurements of circular RNAs in 33 cell types, including the HEK293 kidney cell line also used in the QKI PAR-CLIP experiment (Hafner et al., 2010).

Human exons, introns, 5' UTRs, and 3' UTRs

Human (hg38) Exons, Introns, 5' UTR, and 3' UTR BED files were downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables) (Tyner et al., 2017). Those BED files were

generated from data provided by the Genome Reference Consortium, not from cell lines but from 11 genomic clone libraries from deceased donors.

Human cassette exons

Cassette exons are skipped during alternative splicing. The cassette exon BED fie was downloaded from HEXEvent (http://hexevent.mmg.uci.edu/cgi-bin/HEXEvent/HEXEventWEB.cgi), which was curated from the UCSC Genome Browser (Tyner et al., 2017). The data was extracted from three tracks in the UCSC Genome Browser: UCSC Genes, Spliced ESTs, and Human mRNAs (Busch and Hertel, 2013). The purpose of finding QKI binding sites near cassette exons is to find more candidate locations that might be associated with exon skipping caused by QKI.

Bioinformatics tools to find QKI binding sites in and near regulatory regions

To determine how QKI binding sites are distributed near transcriptome elements (e.g., intron-exon junctions), I used bedtools closest (Quinlan and Hall, 2010), a command line bioinformatics tool that measures the distance between the elements of two BED files (e.g., QKI binding sites and exons). There are three possible cases for each QKI landmark: within region (assigned zero value), upstream of region (negative distance), and downstream of region (positive distance). The "bedtools closest" utility and subsequent AWK programs were used to find the distance of QKI sites to the nearest adjacent region start or end annotated in the genome (exon, intron, 5'UTR, 3'UTR, transcription start site (TSS), transcription end site (TES)). Then I used R Studio to create a histogram of QKI bound near the start of the region and QKI bound near the end of the region (in Figure 8B that is the splice acceptor junction histogram and the splice donor junction histogram). The NGSPlot function *ngs.plot.r* was used to generate density plots of QKI sites relative to exons (splice acceptor site to splice donor site) and full transcripts (TSS to TES). I did not apply the filtering used for my histograms (Figure 7 and 8) to *ngs.plot.r*

because *ngs.plot.r* output does not make such incorrect measurements, and *ngs.plot.r* is a one-line command from raw data to output graph so I cannot modify its methods in that way.

Raw output from bedtools closest was post-processed using custom scripts to limit analysis to sites within adjoining exons and introns (Figure 7 and 8). For every single exon, bedtools closest finds the nearest QKI binding site. The nearest QKI binding site to Exon 1 is -20 nt upstream. For Exon 2, there are no QKI binding sites in the flanking introns, so bedtools closest looks past the flanking introns to find the nearest QKI and measures -500. -20 is a correct measurement (green) because it is in a flanking intron but -500 is not correct (red), because -500 is not in a flanking intron. Instead (Figure 7B) I used bedtools closest to create 2 files (file 1: all of the QKI binding sites in introns and file 2: all the QKI binding sites in exons) then used my AWK script to measure the distance from the QKI binding site to the nearest part of the intron (file 1) or exon (file 2) that the QKI is inside of (either the start or the end, whichever is closer, e.g. the QKI binding site below is closer to the end of intron 1 so I measure -20 nt from the end of intron 1). -500 is not measured because my AWK script does not go beyond the region that the QKI binding site exists in (e.g. the region that the QKI binding site exists in below is Intron 1).

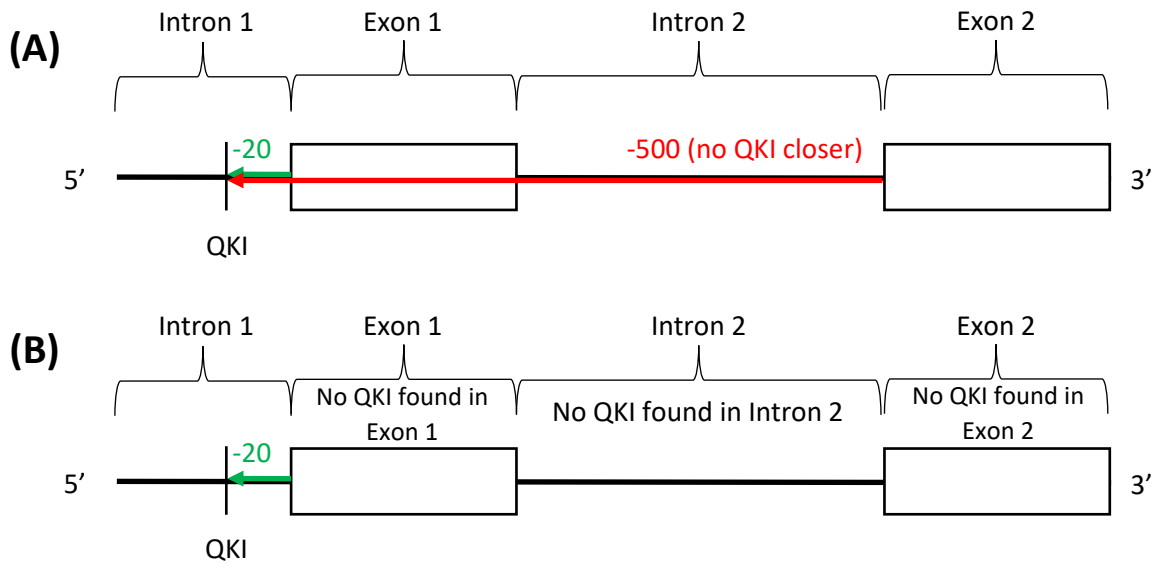

Figure 7: (A) The bedtools closest function can return distances beyond flanking introns (e.g. -500 nt is not within the region of interest with respect to the splice acceptor site for Intron 2). (B) I used bedtools closest to find QKI binding sites within introns and QKI binding sites within exons, then used my AWK script to measure the distance to the closest part of the region (either the start of the region or the end of the region, e.g. the QKI binding site in the figure above is closer to the end of the intron, so -20 nt is measured).

Next, to avoid the problem of depicting intervals that extend beyond adjacent elements of interest (see Figure 8A), I split the QKI binding sites in the exons and introns in half (Figure 8B). File 1 is QKI binding site closer to start of the intron (e.g. +40), file 2 is QKI binding site closer to the end of the intron (eg. -20), file 3 is QKI binding site closer to the start of the exon, and file 4 is QKI binding site closer to the end of the exon). Then I merge file 2 and file 3 to create the splice acceptor junction histogram and merge file 4 and file 1 to create the splice donor junction histogram.

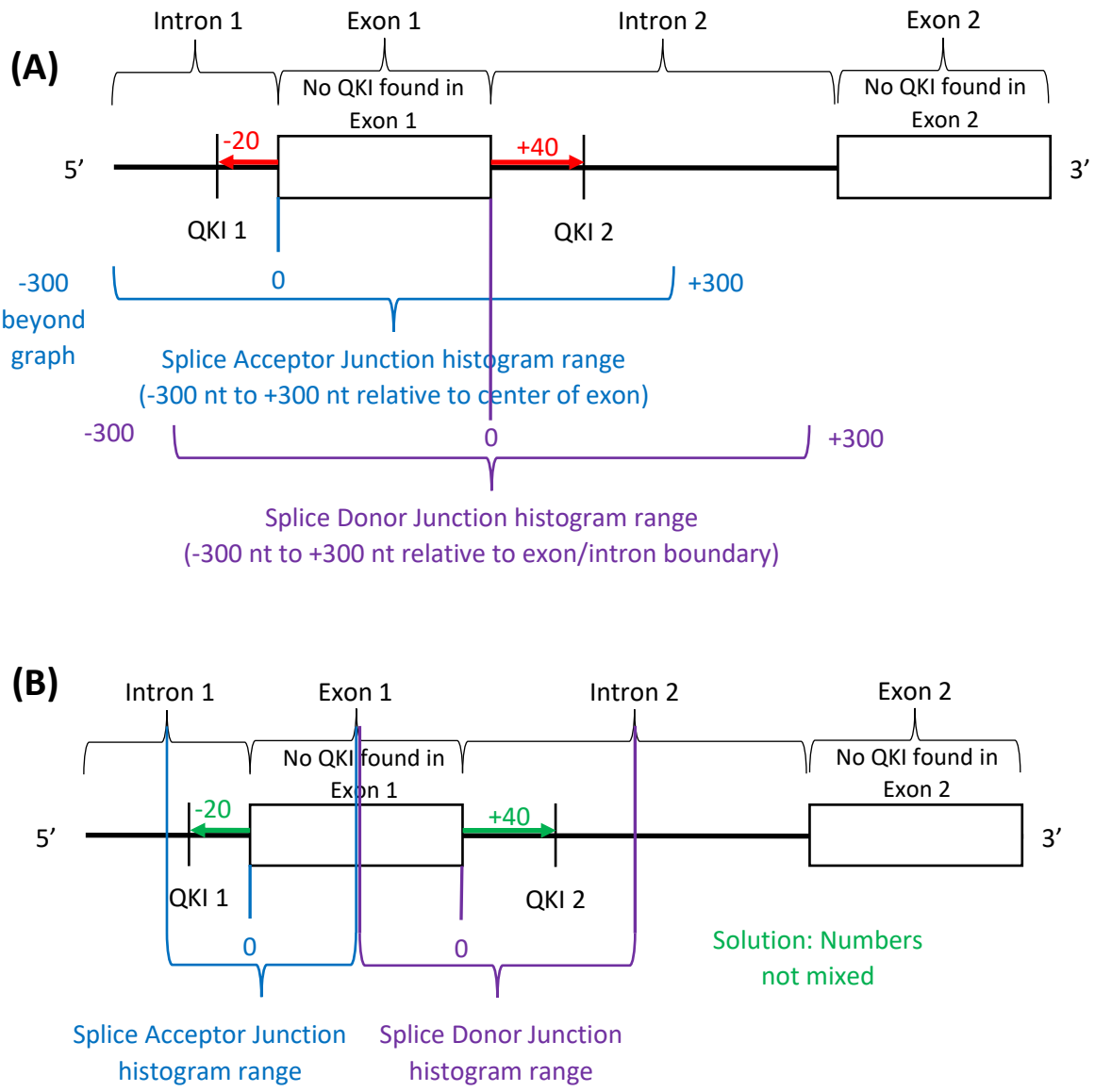

Figure 8: (A) Short regions can cause the junction histograms to overlap, e.g. the short exon above (B). The solution is to split all regions in half and input data to the junction histograms if it is in the proper region (see the blue region and the purple region above).

Results

SF1 and QKI-5 competition near the branchpoint to regulate alternative splicing of NUMB

This work focuses on the NUMB gene because previous *in vivo* research shows that competition between QKI and SF1 determine whether exon skipping occurs at exon 12 of NUMB and point mutations to that NUMB binding site decrease QKI and SF1 binding affinity. Focusing on known binding sites enabled me to evaluate the accuracy of my 3D modeling and also offered the possibility to gain structural insight into the protein-RNA complexes. As described in the Methods section, I used a modeling program previously developed to study miRNA-target interactions (Gan and Gunsalus, 2013). For input structures, I used the solved QKI (4JVH) (Teplova et al., 2013) and SF1 (1K1G) (Liu et al., 2001) x-ray crystal structures. I modeled the protein-RNA complexes for the RNA target sites as underlined in Figure 9.

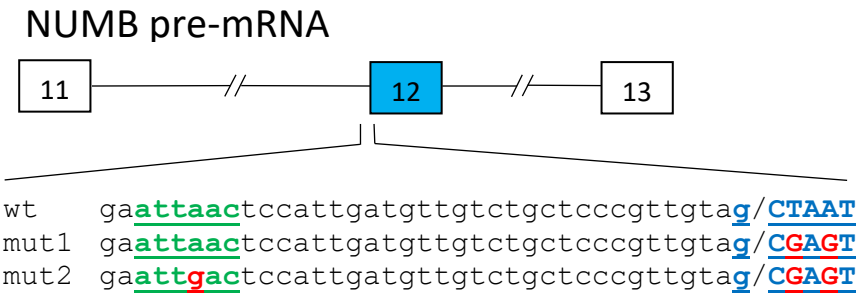

Figure 9: DNA sequences of *in vivo* NUMB wildtype and mutant constructs adapted from (Zong et al., 2014). Intron sequences are in lowercase letters, exon sequences in uppercase letters, and slashes indicate exon/intron junctions. Putative *in vivo* QKI/SF1 binding sites are bold and underlined. Color scheme: mutations, red; intronic binding sites, green; exonic binding sites, blue. QKI binds to both the intron and the exon, whereas SF1 binds to only the intron.

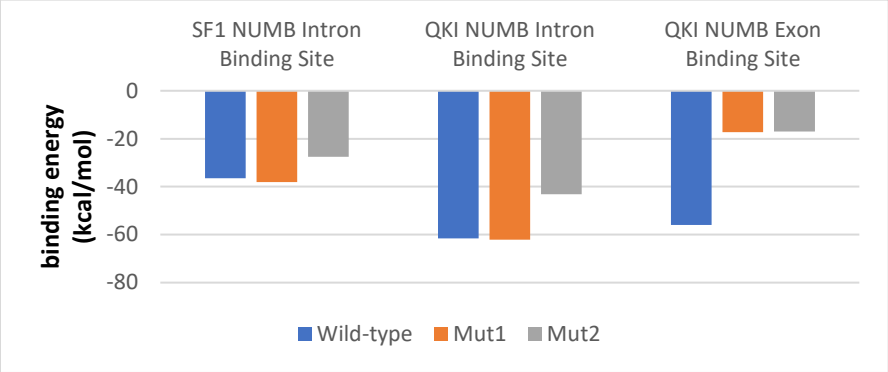

Figure 10: Computed binding affinities for QKI and SF1 binding to WT, mut1, and mut2 RNA targets in Figure 9. Stronger affinity corresponds to more negative values (kcal/mol). QKI outcompetes SF1 binding to the WT intronic splice acceptor site or BPS and the mutations decrease QKI and SF1 binding affinity.

Comparison of QKI and SF1 binding energies from my 3D modeling of both WT and mutated RNA (Figure 10) shows that the pattern of protein-RNA affinity agrees with current experimental data (Zong et al., 2014). SF1 binding was not evaluated at the exon binding site because SF1 always binds to introns. The WT and Mut1 intron binding sites are identical (AUUACC) and show the same binding affinity. The mutated intron binding site AUUGAC (Mut2) shows weaker QKI and SF1 binding affinity compared with WT and Mut1 sequences (Mut2 SF1: -27.6 kcal/mol, Mut2 QKI: -43.1 kcal/mol vs. WT SF1: -36.5 kcal/mol, WT QKI: -61.6 kcal/mol). The mutated exon binding sites GCGAGU (Mut1 and Mut2) show weaker QKI binding affinity than the WT exon binding site GCUAAU (Mut1 QKI: -17.3 kcal/mol vs. WT QKI: -56.0 kcal/mol). Assuming QKI bind to intronic and exonic sites, the combined affinities for WT (-117.6 kcal/mol) are stronger than the combined affinities for Mut1 (-79.3 kcal/mol), which is stronger than the combined affinities for Mut2 (-60.1 kcal/mol). This agrees with previous experimentally determined QKI-5 dissociation constants (256±39 nM for WT, 748±62 nM for Mut1, and >1000 nM for Mut2) (Zong et al., 2014).

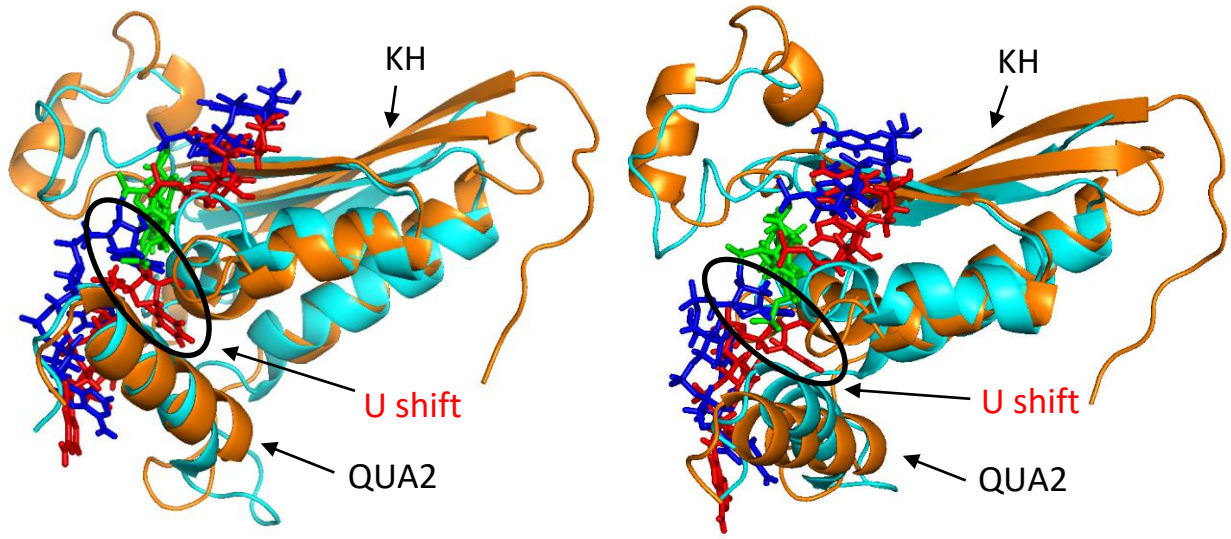

Figure 11: Structural overlap of QKI and SF1 complexes at WT (AUUAAC, left) and mutated Mut2 (AUUGAC, right) branch point sequence (intronic binding site). Color scheme: QKI, orange; RNA bound to QKI, red; SF1, cyan; RNA bound to SF1, blue; nucleotide of interest, green. The QUA2 domain is the bottom alpha helix; the rest is the KH domain. The mutated RNA nucleotide occurs near the KH domain. The U shift seems to be the biggest change in the RNA conformation between QKI and SF1 and might explain why QKI was calculated to have higher affinity to the RNA than SF1 (Figure 10).

Next, I examined QKI-RNA and SF1-RNA complexes to understand the molecular basis of these differences in binding free energy. The greater affinity of RNA for QKI compared to SF1 at the BPS can be explained based on the structural features of their protein-RNA complexes (Figure 11). The third nucleotide (U) of the RNA bound to QKI projects deep into the pocket between QUA2 and KH domains, providing a tighter conformational fit that might increase binding affinity, but the third nucleotide (U) of the RNA bound to SF1 is not in that pocket. This can be seen in Figure 11, where the red RNA (the RNA bound to QKI) has a U inside the pocket and the blue RNA (the RNA bound to SF1) does not. Consistent with these structural differences, my calculations show that QKI has greater binding affinity for WT and Mut2 intronic RNA (WT: -61.6 kcal/mol; Mut2: -43.1 kcal/mol) than SF1 (WT: -36.5 kcal/mol; Mut2: -27.6 kcal/mol) (Figure 10). In addition, my calculations indicate that the intronic mutated binding site (Figure 11), which has a purine-purine mutation A to G, lowers the binding affinity of both QKI and SF1 (WT has greater binding affinity than Mut2) (Figure 10).

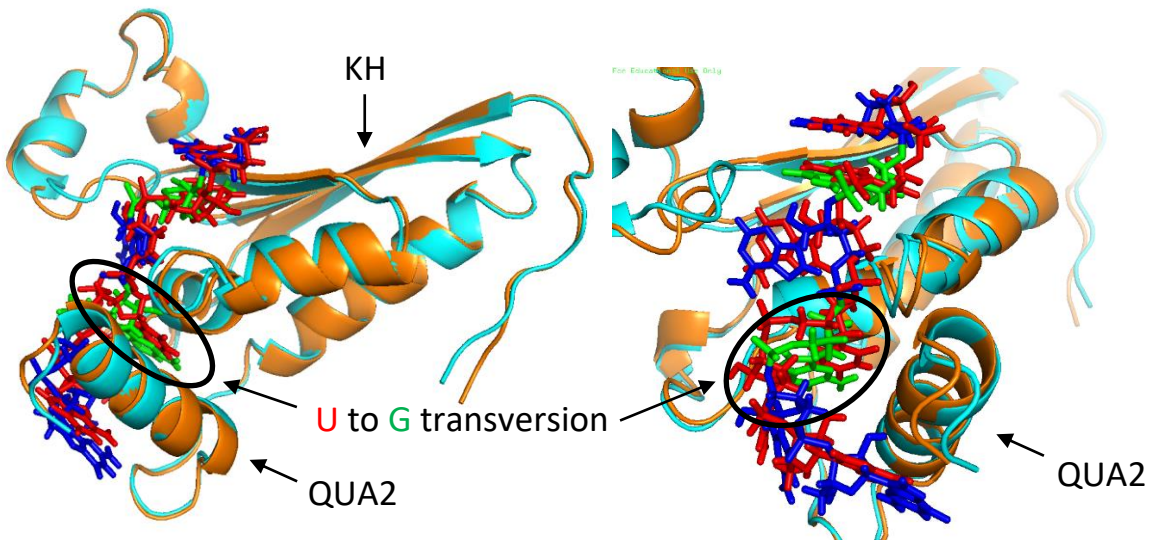

Figure 12: Comparison of QKI bound at exon 12 of NUMB WT or mutated pre-mRNA. The right image rotates and zooms in to display the U to G transversion. Color scheme: QKI(orange)/WT RNA (GCUAAU) complex; QKI(cyan)/Mut2 RNA (GCGAGU) complex, with mutations highlighted in green. The QUA2 domain is the alpha helix at the lower left. The KH domain includes the beta sheets and other helices. The double mutation, containing both the U to G transversion and the A to G transition, distorts the backbone of the RNA, yielding a decrease in binding affinity.

It might be expected that U to G transversion extending into the binding pocket would change the binding affinity, however it does not according to my calculations. Whereas the individual mutations U to G transversion (GCGAAU -63.5 kcal/mol) and A to G transition (GCUAGU -42.5 kcal/mol) are both close to the wild type (GCUAAU -56.0 kcal/mol), only the double mutation GCGAGU (-17.3 kcal/mol) decreases calculated binding affinity of QKI for the pre-mRNA. Structural modeling revealed significant distortion of the RNA backbone in the complex containing the double mutation, providing an explanation for the observed differences in binding affinity (Figure 12).

Another way of assessing my binding energy calculations is to compare them with experimentally measured proportions of NUMB exon 12 inclusion and exclusion for WT, Mut1, or Mut2 RNA sequences in mature transcripts (Zong et al., 2014). The measured NUMB exon 12 inclusion for WT is 60.6±5.1%, Mut1 is 72.9±6.8%, and Mut2 is 34.5±1.8%. Mut1 has an exonic mutation, which decreases QKI binding

but not SF1 binding (because SF1 does not bind to the exon whereas QKI does). This increases the inclusion of exon 12, presumably because QKI is less effective. Mut2 has both intronic and exonic mutations, which decreases both QKI and SF1 binding at the intron and decreases QKI binding at the exon. Mut2 causes less exon 12 inclusion than does the WT sequence, presumably because the intronic mutation prevents SF1 from binding, blocking spliceosome assembly leading to exon skipping. In summary, Mut1 causes less exon skipping than WT due to reduced QKI binding and Mut2 causes more exon skipping than WT due to reduced SF1 binding. These interpretations agree with my calculations (Figure 10).

SF1 and QKI-5 competition on NCURAY SF1 motif variants

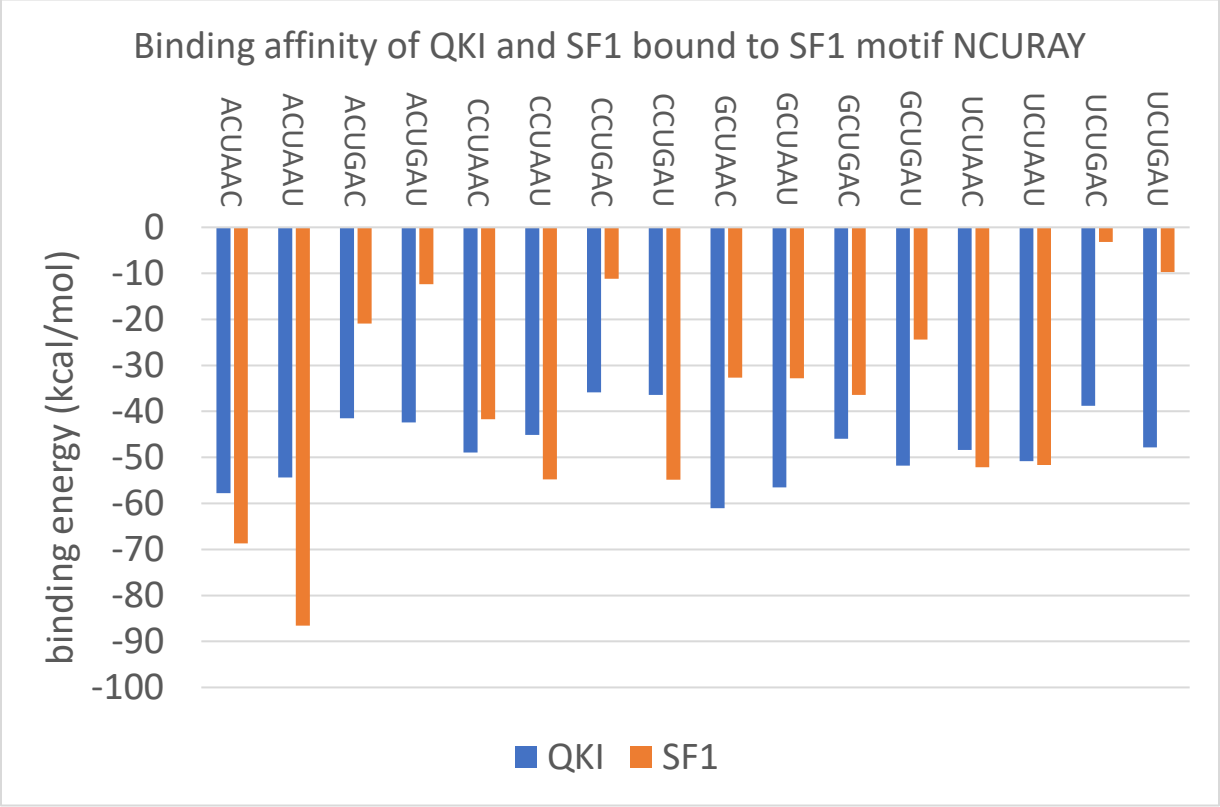

Figure 13: Binding energies of QKI-5 monomer and SF1 bound to SF1 intronic motif NCURAY variants (16 sequences) derived from CAT assay (Peled-Zehavi et al., 2001). QKI-5 outcompetes SF1 in 10 out of 16 cases and SF1 outcompetes QKI-5 in 6 out of 16 cases. The first two motifs (ACUAAC and ACUAAU) are recognized by both QKI-5 (ACUAAY) and SF1 (NCURAY) and the rest by SF1 (NCURAY) only.

To extend the 3D modeling beyond the NUMB gene, I calculated SF1 and QKI binding affinities at the SF1 motif YNCURAY determined by CAT assay (Peled-Zehavi et al., 2001). Visualization of SF1-RNA and QKI-RNA complexes indicate that only six nucleotides interact directly with the protein, so YNCURAY was truncated to the core motif NCURAY. The nucleotides 2-7 of the QKI motif determined by SELEX are ACUAAY (de Miguel et al., 2015), which comprises two variants differing at their sixth nucleotide (ACUAAC and ACUAAU) that represent a subset of the more general SF1 motif NCURAY (sixteen possible variants). In Figure 13, the motif with the highest SF1 binding affinity was ACUAAU and the second highest was ACUAAC, which agrees with the CAT assay (Peled-Zehavi et al., 2001). Significantly, the QKI-5 monomer outcompetes SF1 at most of the variants of the branch point sequence. One limitation of Figure 13 is that only the QKI-5 monomer was used for computing binding energies, whereas QKI-5 can dimerize (Figure 6). We have assessed the effects of dimerization by using additive QKI affinities for WT, Mut1 and Mut2 sequences, which agree with experimental data (Figure 10).

QKI PAR-CLIP binding sites near junctions of interest

Various transcriptomic regions (BED files from online databases) were combined with QKI PAR-CLIP data (Hafner et al., 2010) to determine how far QKI binds from landmarks of interest, specifically the start and end of exons, cassette exons, 5' UTRs, 3' UTRs, and circular RNAs. The cassette exon dataset contains individual skipped exons (a short median nucleotide length of 111) and the circular RNA dataset contains individual exons (e.g. hsa_circ_0014359), and larger sequences of multiple exons or exons and introns (a median nucleotide length of 8266, Table 1).

QKI binds near the splice donor site and splice acceptor site

Knowing where PAR-CLIP-derived QKI sites are distributed near the splice acceptor site could help identify many more candidates for QKI-SF1 competition sites for AS (Figure 14). Genome-wide

distributions of QKI binding around both splice donor and splice acceptor sites show a peak of QKI binding near the splice junction in both the intron (A and A') and all exons, including both constitutive and cassette exons, (B and B') regions, with a higher proportion of sites on the intronic side. The distribution of QKI sites in the exons decays rapidly from the junction, suggesting that QKI sites cluster near the putative SF1 competition sites. In conclusion, genome-wide evidence shows that QKI binds near the splice acceptor site, consistent with its role as an alternative splicing regulator.

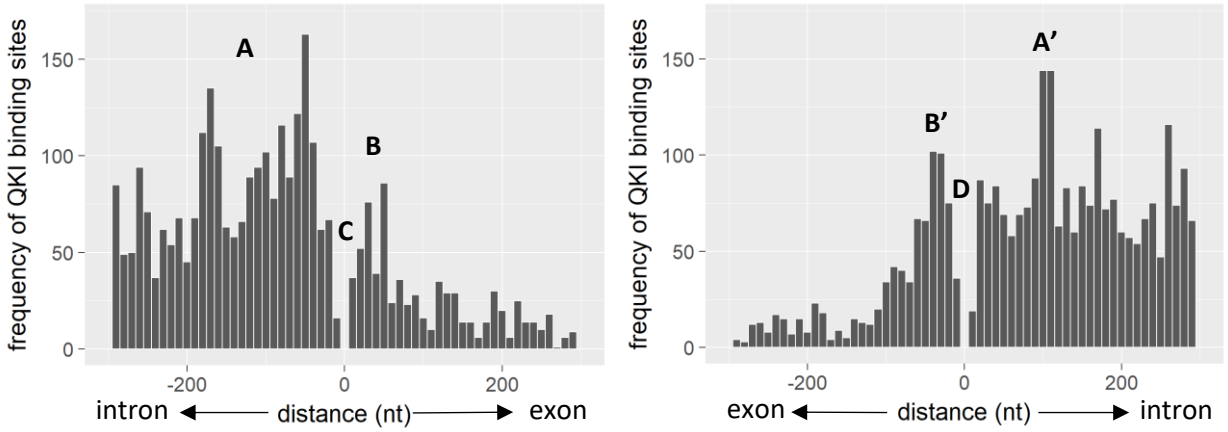

Figure 14: Distributions of QKI sites at splice acceptor (left) and splice donor (right) junctions. A and A' are intronic peaks, B and B' are exonic peaks, C is a valley around the acceptor site, and D is a valley around the donor site. All histograms have the same bin size (10nt).

Ngs.plot.r was used as another way to view upstream, inside, and downstream of the exon (Figure 15). I did not apply the filtering used for my histograms (Figure 7 and 8) to *ngs.plot.r* because *ngs.plot.r* output does not make such incorrect measurements, and *ngs.plot.r* is a one-line command from raw data to output graph so I cannot modify its methods in that way. Both Figure 14 (counts) and Figure 15 (densities) show peaks of QKI binding (A, A', B, and B') and valleys (C and D) in the same positions relative to the splice junction boundaries. It is unknown why QKI binds inside of the exon (B and B').

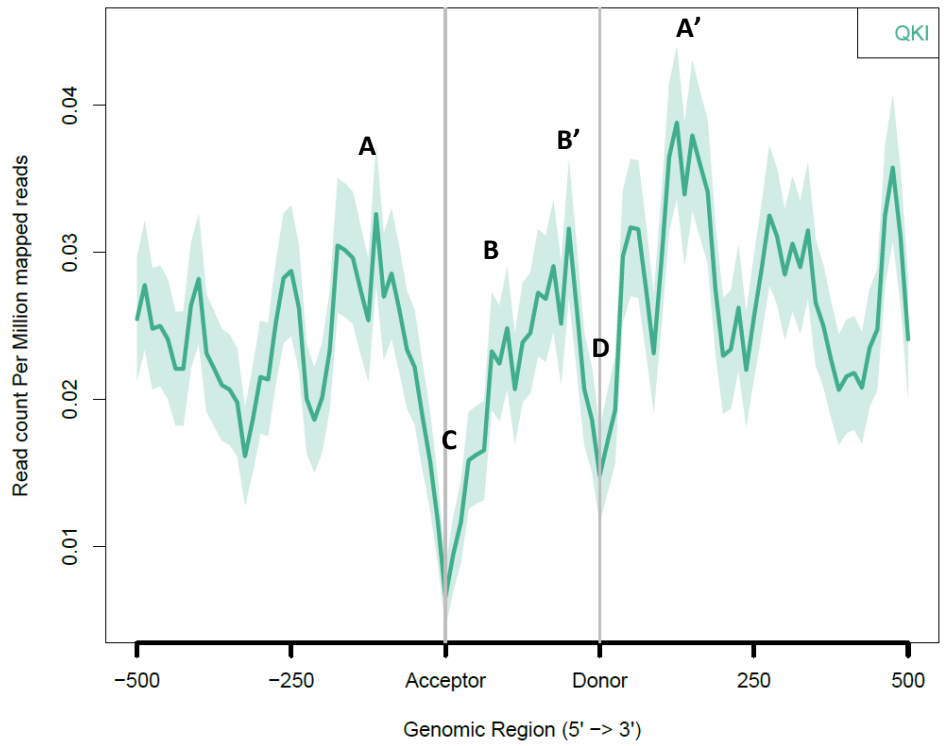

Figure 15: Density plot of QKI binding sites across an idealized exon. The splice acceptor and donor sites are at the 5’ and 3' ends of the exon, respectively. Intronic regions are measured in nucleotides; the exonic region is scaled proportionally to the full length of all exons analyzed. A is an intronic peak, B is an exonic peak, C is a valley around the acceptor site, and D is a valley around the donor site.

QKI binds in introns near exons that can be skipped

Since cassette exons are by definition where exon skipping occurs, it is reasonable to expect that QKI binds near the splice acceptor site of cassette exons. QKI sites near cassette exons (Figure 16) occur mostly in the flanking introns (E and E’), but are not enriched there relative to all exons (there are 15 times as many exons as cassette exons (Table 1, see methods section) and 15 times as many QKI binding sites in all exons (Figure 14 (A and A’) as in cassette exons (Figure 16 (E and E’)). This indicates that QKI can sometimes bind in introns near an exon without causing exon skipping because otherwise Figure 14 (exons) and Figure 16 (cassette exons) should have the same number of QKI binding sites. Since QKI outcompetes SF1 at the SF1 binding motif in 10 out of 16 known SF1 recognition motifs (Figure 13), not all QKI sites near the junctions will have more favorable affinity than SF1. Future research could be to

search the PAR-CLIP data for the 16 known SF1 motifs to learn the frequency of how many times QKI binds to each of the 16 possible binding sites. The calculation of the binding affinity of QKI and SF1 at the 16 SF1 motifs was the last thing done for this project and my advisor did not request proceeding to finding the frequency of QKI binding to the 16 motifs. A next step might be to attempt to compute the affinity of QKI and the frequency of QKI binding to the PAR-CLIP QKI binding sites that is an unknown motif that is not among the 16 known motifs. The PAR-CLIP data contained genomic positions not RNA binding motifs, and finding motifs other than the 16 known ones would be non-trivial because the PAR-CLIP genomic position data had a median of 32nt and the QKI binding site would presumably be in the center of the sequence, but the PAR-CLIP paper (Hafner et al., 2010) admits that the QKI binding site is only approximately in the center, it varies. Therefore, it is not possible to search for the unknown motif based on position because the position varies, and it is not possible to search for the unknown motif based on a character string because the unknown motif is not part of the 16 known motifs so the unknown motif's sequence is not known to science. It is surprising that a QKI binding peak is not observed upstream of cassette exons at a position corresponding to the branch point sequence (although there is some variation in the distance from the branch point sequence to the 3' end of the intron (Figure 1) that would broaden that peak) because that is where QKI is thought to outcompete SF1 to interfere with spliceosome recruitment, causing the exon to be skipped (Figure 2). QKI does have a role in alternative splicing that does not involve cassette exons specifically (e.g. circular RNA formation) and those QKI would bind in places other than the branch point sequence, which might explain the broad distribution of QKI in the introns. Also, there is evidence that SF1 is not essential for constitutive splicing and is also involved in alternative splicing. The majority of SF1 binding sites did not map to the expected position near the splice acceptor site. Instead, target sites were distributed throughout introns, and a smaller but significant fraction occurred in exons within coding and untranslated regions. Since QKI binds to SF1 motifs, this might explain the broad distribution of QKI in the introns,

exons, and untranslated regions. All of these data suggest that there are still many things we do not understand about splicing regulation. It is possible that a QKI binding peak was not observed because more cassette exons remain to be found in the human genome, however there has been a huge amount of tissue-specific transcriptomics done already.

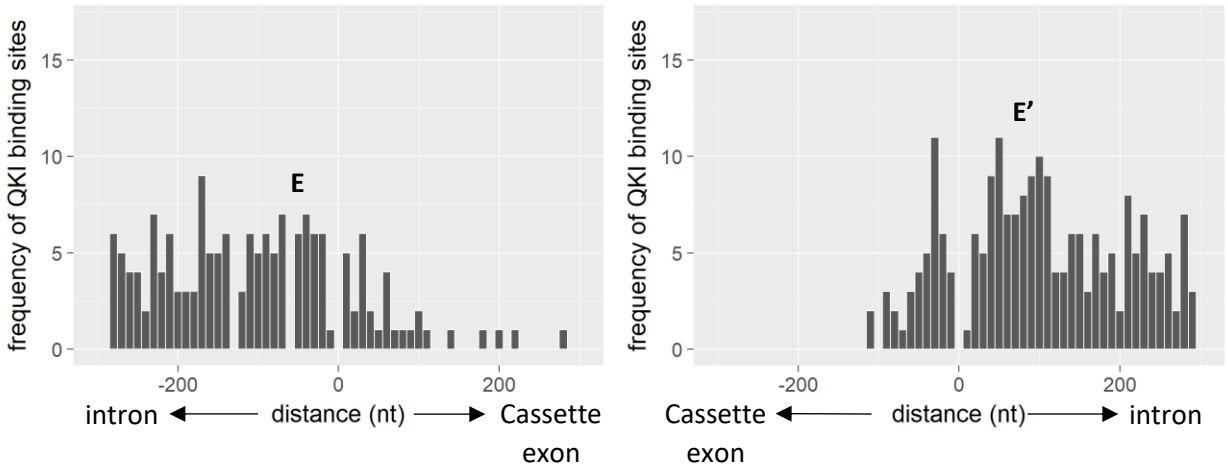

Figure 16: Distributions of QKI sites at splice junctions for cassette exons that can be skipped. Regions surrounding splice acceptor (left) and splice donor (right) sites are shown. E and E' are intronic peaks.

QKI binds near the 5' UTR-CDS and 3' CDS-UTR junctions

QKI has other regulatory roles, including mRNA stabilization by binding to the 3'UTR, for example in MBP mRNA (Li, Z. et al., 2000). Thus, it is interesting to examine QKI binding sites at the 5' UTR-CDS and 3' CDS-UTR junctions. A peak of QKI binding is observed at the start of the 3' UTR (Figure 17) (F'). Interestingly, there is also a peak of QKI binding at the end of the 5' UTR (Figure 17) (F). One hypothesis to explain why peaks of QKI binding occur at those locations is that QKI may play a role in the circularization of mRNA during translation. In addition, there was a large peak of QKI binding at the start of the 5' UTR (G) and the end of the 3' UTR (H) (Figure 18). These results suggest that binding of QKI to those sites may block transcript degradation or promote mRNA stabilization (Li, Z. et al., 2000). *Ngs.plot.r* was used as another way to view QKI binding sites upstream, inside, and downstream of the

gene (Figure 19). Figure 18 (QKI counts) and Figure 19 (QKI densities in the genebody) show similar peaks of QKI binding (G and H). The start of the 5' UTR has a small peak of QKI binding near the TSS (G) and the end of the 3' UTR has a large peak of QKI binding at the TES (H).

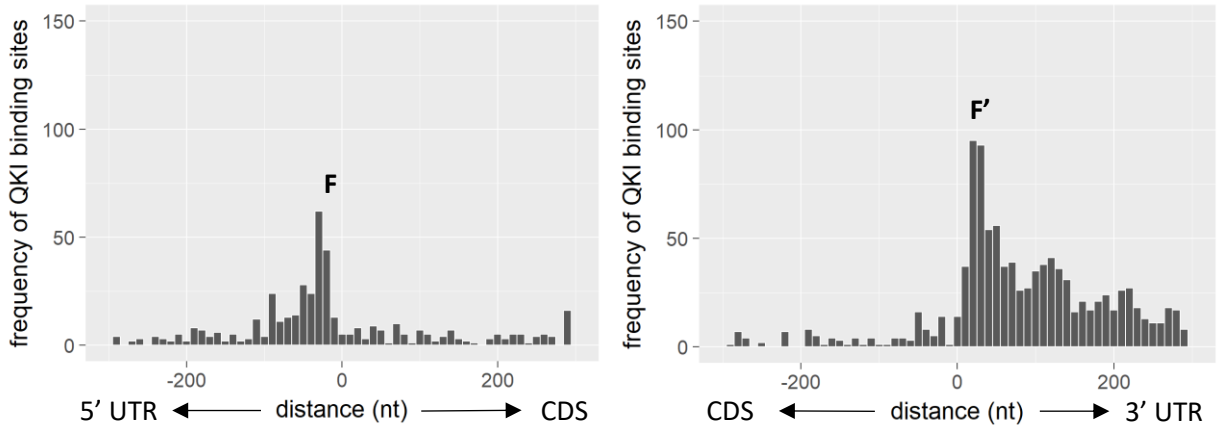

Figure 17: Distributions of QKI sites near the 5' UTR-CDS junction of the first exon (left) and 3' CDS-UTR junction of the last exon (right). CDS stands for Coding Sequence and UTR stands for Untranslated Region. F is a peak in the 5' UTR and F' is a peak in the 3' UTR.

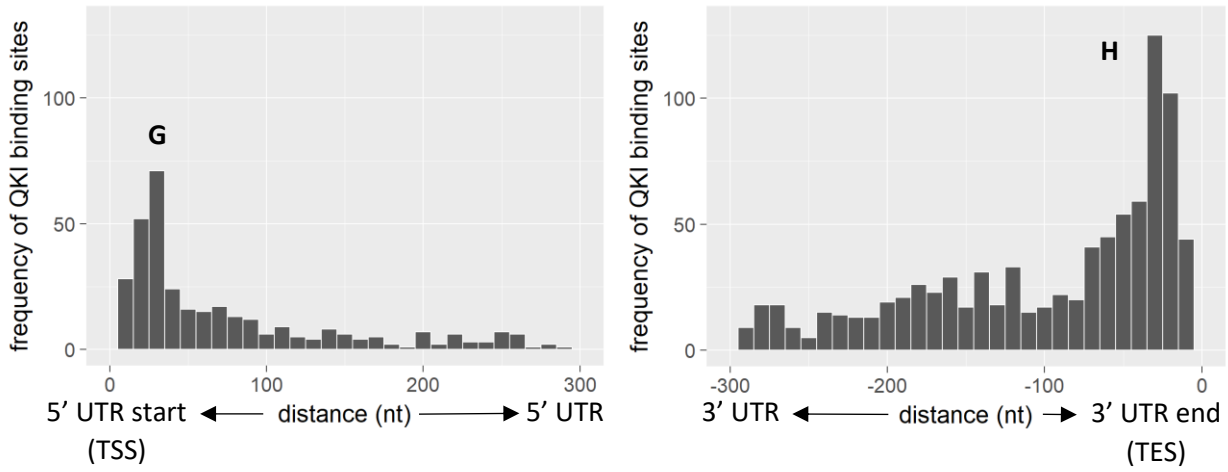

Figure 18: Distribution of QKI binding sites downstream of Transcription Start Sites (TSS, left) and upstream of Transcription End Sites (TES, right). G and H label peaks at the start and end of the UTRs, respectively.

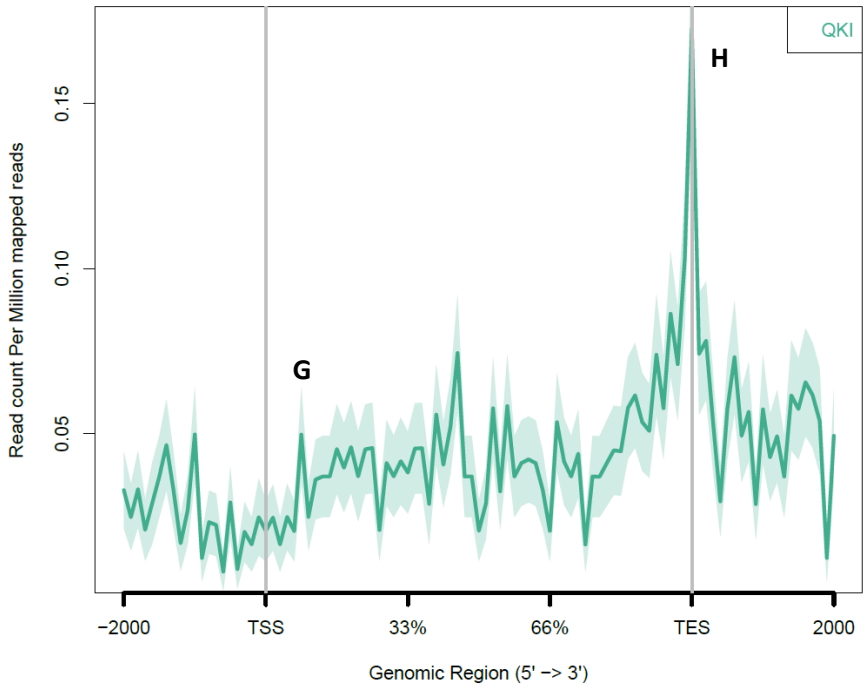

Figure 19: Density plot of QKI binding sites across an idealized gene. TSS is Transcription Start Site and TES is Transcription End Site. The x-axis from -2000 to the TSS and TES to +2000 is measured in nucleotides. The TSS to TES region is measured in percent. There are peaks of QKI binding outside of the transcribed region because the *ngs.plot.r* algorithm looked for binding sites in the genomic DNA, specifically human genome version 19 (hg19) and called the output a genebody graph.

QKI binds in circular RNAs and their flanking introns

The current model (Figure 3) is that QKI dimerizes to juxtapose two sites along a pre-mRNA to promote circular RNA formation. Unexpectedly, I found only slight binding signals at the start and end of annotated circular RNAs (Figure 20). One factor that could contribute to this result is that the circular RNA database circBase contains only 33 cell types (Glažar et al., 2014), whereas the human body has about 200 total cell types (AAAS, 2012), so only a fraction of all circular RNAs may be present in circBase. The circBase paper does not address sampling depth, but does say that circBase contains data from all studies of large-scale circular RNA identification published before 2014 (Glažar et al., 2014) and the circBase website says that the most recent update to the circBase *H. Sapiens* genome was in July 2017 (Maass et al., 2017). My research tests whether there are QKI binding sites near each circular RNA in circBase, so if circBase had contained more circular RNAs there would be more chances to find nearby

QKI binding sites. Also, the QKI binding site PAR-CLIP data contains only 1 cell type, HEK239 cells, and circBase finds only 239 circular RNAs in HEK239 cells. It was expected that there would be a peak of QKI binding to the exon-adjacent portion of flanking introns, because when a previous experiment tested whether QKI binds the SMARCA5 pre-mRNA, by performing RNA-immunoprecipitation (RIP) assays and using qRT-PCR to quantify QKI occupancy within the introns adjacent to the circRNA-forming exons, they found that QKI binds to the exon-adjacent sites (Conn et al., 2015) at a level comparable to its binding to a site in the previously explored QKI target, NUMB (Zong et al., 2014). The small peak at (I) in Figure 20 is not in the flanking intron, however it could be expected because QKI can bind to both the intron and the start of an exon (Figure 9). One might expect the amount of QKI binding sites in the flanking introns in Figure 20 to be higher than the QKI binding sites inside the circular RNA because that was what was found in SMARCA5, but in Figure 20 the QKI binding sites in the flanking introns were lower than inside the circRNA.

Upon closer examination, I found that most QKI binding sites associated with circRNAs were near longer circular RNAs of at least 3000nt (data not shown). Previous work has indicated that sequence complementarity in the flanking introns may be the major driver of formation of circular RNAs (Conn et al., 2015; Rybak-Wolf et al., 2015). Considering possible structural conformations, it may be expected that bringing the ends of short circular RNAs into close proximity for cleavage and ligation requires less assistance from auxiliary factors, so that RBP dimerization may be more important for the efficient formation of longer circular RNAs (Figure 3). Future experiments to test this idea could provide more insight into the role of QKI in circular RNA formation.

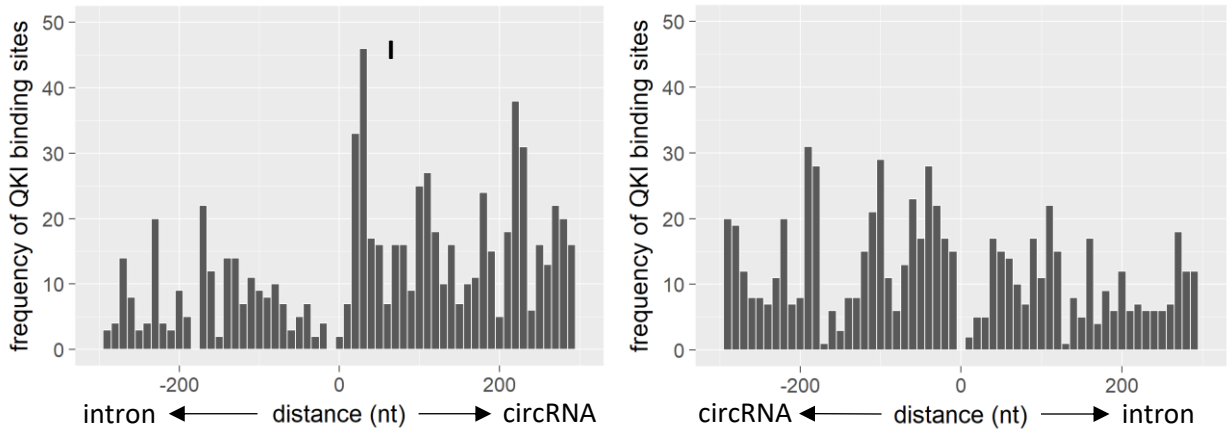

Figure 20: Distribution of QKI sites near the 5' (left) and 3' (right) ends of circular RNAs. There is slight enrichment at (I).

Discussion

Previous studies have suggested that QKI competes with SF1 at the splice acceptor site to control exon skipping and that QKI has a role in the biogenesis of circular RNAs. I have used molecular modeling and bioinformatics tools to examine these findings in greater detail. Molecular modeling of QKI/SF1 competition at a set of binding sites in the NUMB gene and all SF1 consensus motif variants confirmed that QKI outcompetes SF1 for most putative SF1 target sites. A genome-wide search (Figure 14 to Figure 20) focused on histograms of junctions and idealized plots of all exons and the entire gene body:

- intron-exon junction and exon-intron junction (all exons)
- *ngs.plot.r* of idealized (the distance between the start and end of the exon is measured in percent not nt distance) exon (all exons)
- intron-cassette exon junction and cassette exon-intron junction
- 5' UTR-CDS junction and CDS-3' UTR junction
- start of the 5' UTR (TSS) and end of the 3' UTR (TES)
- *ngs.plot.r* of the entire gene body, as an idealized gene (the distance between the start and end of the gene is measured in percent not nt distance)

- start and end of circular RNA

The model for the first step of cassette exon skipping (Figure 6) is that one QKI monomer binds to the intronic splice acceptor site and then sometimes recruits another monomer 1-20nt downstream, in the start of the exon, to compete with SF1. The model for the next step that happens after QKI outcompetes SF1 (Figure 2B) is that SR attaches to the next downstream U2AF, causing exon skipping. The model for circular RNA formation (Figure 3) is that in a pre-mRNA QKI monomers bind to different introns and juxtapose those two sites to promote circular RNA formation. Below, I discuss several issues related to the model for the roles of QKI-SF1 competition, the possible role of QKI in stabilizing mRNAs by binding to their 3' UTR and blocking translation by binding to the 5' UTR, and limitations of available data and our computational procedures.

QKI binding sites are more common in the intronic splice donor site and splice acceptor site than in all exons (Figure 14) and cassette exon (Figure 16) near the splice acceptor site and splice donor site. The biological function of QKI binding to the splice donor site of all exons and cassette exons remains to be discovered. The peak of QKI binding at the splice acceptor site suggests that QKI may preferentially bind to the intron, near the splice acceptor site, as a monomer and sometimes recruits a second monomer 1-20nt downstream in the start of the exon (Figure 6) to form a dimer that spans both the intronic and exonic sites. This scenario is consistent with our QKI/SF1 binding affinity results (Figure 13), which imply that the QKI monomer outcompetes SF1 for most binding motif sequence variants. An intriguing next step would be to compare QKI binding affinity and motif at intronic and exonic sites. However, at present there is insufficient data to compare QKI binding affinity in introns vs exons (Figure 10 investigates only one gene and Figure 13 is a limited NCURAY SF1 motif, which may not represent all of the *in vivo* QKI binding sites). Another feature of QKI binding sites is a broad distribution in the introns (Figure 14 and Figure 16). A plausible explanation is that an intron may contain multiple splice sites, for

example a twintron contains multiple splice sites (Drager and Hallick, 1993). Further work will be required to examine such patterns in more detail. Future research should compute the affinity of QKI to the binding sites found in the PAR-CLIP data, which may have different sequences than the 16 versions of the SF1 motif tested in Figure 13.

This paper also examines a current hypothesis that QKI dimerizes to juxtapose two sites along a pre-mRNA to promote circular RNA formation (Conn et al., 2015). However, our results (Figure 20) suggest QKI sites tend to be evenly distributed with no clear peaks of QKI binding near exons that form circular RNAs. However, most QKI binding sites associated with circRNAs were near longer circular RNAs of at least 3000nt (data not shown), so my histograms may be obscuring this signal. One fix might be to split the circRNA data into distance ranges (e.g. 0-500, 500-1000, etc.) and plot a pair of histograms (one for the intron-circRNA junction and one for the circRNA-intron junction) for each distance range. Future *in vivo* research could validate or refute that QKI is not enriched in flanking introns (Figure 20) by testing whether QKI binds to exon-adjacent sites in flanking introns not just in SMARCA5 (Conn et al., 2015) but in more cases.

An unexpected finding is that QKI sites accumulate near the UTR-CDS junctions (Figure 17). In eukaryotes, mRNAs circularize to improve translational efficiency (Wells et al., 1998). Therefore, a plausible explanation of QKI binding to the UTR-CDS is that QKI dimerization plays a role in bringing the two ends of the mRNA together to improve translational efficiency. This is an entirely different process than circular RNA formation caused by QKI-5.

Another unexpected finding is that QKI sites accumulate near the start of the 5’ UTR and the end of the 3’ UTR (Figure 18). A likely explanation is that QKI may bind to those locations to block mRNA

degradation and repress translation, as has been shown for the *C. elegans* QKI homolog GLD-1. In *C. elegans,* GLD-1 binds to the 5' UTR to repress translation by repressing ribosome assembly on the bound mRNA or by repressing subsequent ribosome elongation (Artzt et al., 2010; Lee and Schedl, 2004). Moreover, binding of GLD-1 protects *gna-2* mRNA from nonsense-mediated mRNA decay (NMD), likely by binding to the 5' UTR and repressing translation of the upstream open reading frames (Lee and Schedl, 2004). Thus, QKI might have similar functions.

A potentially significant limitation of our analysis was that the QKI binding sites determined by PAR-CLIP originated from a single cell type (HEK295), so many other regulatory splice sites may not be found because they are not part of the HEK295 transcriptome. By contrast, genomic annotations for introns, exons, UTRs and circular RNAs are based on dozens of cell types. Therefore, my analysis provides only a partial view of the transcriptome regulated by QKI.

The main conclusions of this research are as follows:

1. Binding affinity calculations based on structural modeling support the experimental finding that QKI monomer outcompetes SF1 at the NUMB exon 12 splice acceptor site.
2. Structural modeling also indicates that the QKI monomer frequently shows stronger affinity than SF1 at most known SF1 motifs.
3. Analysis of genome-wide binding data shows that QKI binds at both the splice donor site and the splice acceptor site of exons and cassette exons, near the 5' UTR-CDS junction in the first exon and 3' CDS-UTR junction in the last exon, and near the start of the 5' UTR and the end of the 3' UTR. In contrast, QKI binding is not detected at high frequency near circular RNAs.

Thus, my computational results corroborate and provide a biophysical explanation for experimental data indicating that QKI regulates alternative splicing by outcompeting SF1 for binding to the same sites (Zong et al., 2014).

One next step might be to search for circular RNAs in cell types beyond the 33 cell types from circBase and expand the QKI PAR-CLIP beyond just HEK295 cells (which have 239 circular RNAs), because the frequency of QKI binding near the start and end of circular RNAs with the current 33 cell type dataset was very low (Figure 20), and both advances might lead to identification of more QKI binding sites. The cell type with the most circular RNAs (38,983) in circBase is human fetal frontal cortex tissue from independent donors (Glažar et al., 2014). Circular RNAs are extraordinarily enriched in the mammalian brain, and many mRNAs are translated in an activity-dependent fashion at the synapse, where they play a role in synaptic plasticity (Rybak-Wolf et al., 2015). Circular RNAs are highly enriched in the synaptoneurosome compared to the whole-brain lysate and cytoplasm, and CRISPR-Cas9 editing of the mouse genome to remove the locus encoding the circular RNA Cdr1 resulted in dysfunction of excitatory synaptic transmission (Piwecka et al., 2017). Circular RNA also regulates differentiation (e.g. Epithelial to Mesenchymal Transition (Conn et al., 2015)), and circular RNAs are upregulated during neuronal differentiation (Rybak-Wolf et al., 2015). The first cell type to use to search for more circular RNAs might be embryonic cell types not yet tested for circular RNA, because embryonic cells differentiate into all the tissues in the adult body. The first type of differentiation to investigate would be neurogenesis, because circRNA is most highly enriched in neural tissue and is involved in neurogenesis and synaptic plasticity. The second cell type to use to search for even more circular RNAs might be cancer types not yet tested for circular RNA, because the list of circular RNAs involved in cancer continues to grow and in the vast majority of those, the function of the change in circular RNA is unknown (L S Kristensen et al., 2018).

Another possible next step is to examine whether there is a correlation between QKI binding frequency and circular RNA length, a suggested by preliminary analyses here.

To advance our understanding of exon skipping by QKI, one might search for the SF1 motif YNCURAY near the branch-point sequences that were predicted by the machine learning algorithm Branchpointer, because there may be more introns out there than are currently annotated, which could harbor additional QKI binding sites (Signal et al., 2018). This set of data could provide more putative QKI-SF1 competition sites, which can be screened using our 3D modeling. Branchpointer identified branchpoint elements solely from genomic sequences and gene annotations (Signal et al., 2018). When applied to all introns in human gene annotations, Branchpointer found branchpoints in 87% of introns and annotated 353,177 branchpoints in 216,974 introns in human genes. A large fraction (47%) of branchpointer-annotated introns contained more than one branchpoint. Multiple-branchpoint introns had higher exon expression, higher intron conservation, and were typically shorter than single-branchpoint introns. Searching through the Branchpointer data to identify more potential QKI-SF1 competition sites to find more possible cassette exons *in silico* could guide *in vivo* experiments to verify if those are cassette exons. This would in turn expand our understanding of alternative transcription in the transcriptome, which might be a step towards further understanding of embryonic development, tissue differentiation, and cancer.

References

AAAS. (2012). The Cells in Your Body - Science NetLinks.*2018,*

Artzt, K., Wu, J.I., Feng, Y., Bankston, A., Ryder, S.P., Massi, F., and Sette, C. (2010). Post-Transcriptional Regulation by STAR Proteins: Control of RNA Metabolism in Development and Disease Springer US).

Beuck, C., Qu, S., Fagg, W.S., Ares, M., and Williamson, J.R. (2012). Structural analysis of the Quaking homodimerization interface. J. Mol. Biol. *423,* 766-781.

Bhagat, Y.S. (2013). Introns: structure and functions.*2018,*

Black, D. (2016). Pre-mRNA Splicing Lecture 1 Douglas Black. Fig The most complex RNA processing reaction is pre-mRNA splicing. Most genes in metazoan (multicellular) - ppt download.*2018,*

Brosseau, J., Lucier, J., Nwilati, H., Thibault, P., Garneau, D., Gendron, D., Durand, M., Couture, S., Lapointe, E., Prinos, P.*, et al.* (2014). Tumor microenvironment-associated modifications of alternative splicing. RNA (New York, N.Y.) *20,* 189.

Busch, A., and Hertel, K.J. (2013). HEXEvent: a database of Human EXon splicing Events. Nucleic Acids Res *41,* D124.

Cha, I.E., Hoblitzell, K.L., and Rouchka, E.C. (2008). Alternative Splicing Events.*2018,*

Chen, A., Paik, J., Zhang, H., Shukla, S.A., Mortensen, R., Hu, J., Ying, H., Hu, B., Hurt, J., Farny, N.*, et al.* (2012). STAR RNA-binding protein Quaking suppresses cancer via stabilization of specific miRNA. Genes & Development *26,* 1459.

Chen, L. (2016). The biogenesis and emerging roles of circular RNAs. Nature Reviews Molecular Cell Biology *17,* 205.

Clancy, S. (2008). RNA Splicing: Introns, Exons and Spliceosome | Learn Science at Scitable. In Nucleic Acid Structure and Function, Moss, Bob ed., Nature Education 1(1):31)

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

Conn, S., Pillman, K., Toubia, J., Conn, V., Salmanidis, M., Phillips, C., Roslan, S., Schreiber, A., Gregory, P., and Goodall, G. (2015). The RNA Binding Protein Quaking Regulates Formation of circRNAs. Cell *160,* 1125-1134.

Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biology *12,* R79.

Costanzi, C., and Pehrson, J.R. (1998). Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. Nature *393,* 599-601.

Darbelli, L., and Richard, S. (2016). Emerging functions of the Quaking RNA-binding proteins and link to human diseases. WIREs RNA *7,* 399-412.

de Miguel, F.J., Pajares, M.J., Martínez-Terroba, E., Ajona, D., Morales, X., Sharma, R.D., Pardo, F.J., Rouzaut, A., Rubio, A., Montuenga, L.M., and Pio, R. (2015). A large-scale analysis of alternative splicing reveals a key role of QKI in lung cancer. Clinical Microbiology Newsletter *37,* 33.

Drager, R.G., and Hallick, R.B. (1993). A complex twintron is excised as four individual introns. Nucleic Acids Research *21,* 2389-2394.

Feracci, M., Foot, J.N., Grellscheid, S.N., Danilenko, M., Stehle, R., Gonchar, O., Kang, H., Dalgliesh, C., Meyer, N.H., Liu, Y.*, et al.* (2016). Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. Nature Communications *7,* 10355.

Gan, H.H., and Gunsalus, K.C. (2015). Assembly and analysis of eukaryotic Argonaute–RNA complexes in microRNA-target recognition. Nucleic Acids Res *43,* 9613-9625.

Gan, H.H., and Gunsalus, K.C. (2013). Tertiary structure-based analysis of microRNA-target interactions. RNA (New York, N.Y.) *19,* 539-551.

Girard, L.R., Fiedler, T.J., Harris, T.W., Carvalho, F., Antoshechkin, I., Han, M., Sternberg, P.W., Stein, L.D., and Chalfie, M. (2007). WormBook: the online review of Caenorhabditis elegans biology. Nucleic Acids Res. *35,* 472.

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA (New York, N.Y.) *20,* 1666-1670.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A., Munschauer, M.*, et al.* (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell *141,* 129-141.

L S Kristensen, T B Hansen, M T Venø, and J Kjems. (2018). Circular RNAs in cancer: opportunities and challenges in the field. Oncogene *37,* 555-565.

Lee, M., and Schedl, T. (2004). Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in C. elegans. Genes & Development *18,* 1047-1059.

Li, X., Yang, L., and Chen, L. (2018). The Biogenesis, Functions, and Challenges of Circular RNAs. Molecular Cell *71,* 428-442.

Li, Z., Zhang, Y., Li, D., and Feng, Y. (2000). Destabilization and Mislocalization of Myelin Basic Protein mRNAs in quaking Dysmyelination Lacking the QKI RNA-Binding Proteins. J. Neurosci. *20,* 4944-4953.

Liu, Z., Luyten, I., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A., and Sattler, M. (2001). Structural Basis for Recognition of the Intron Branch Site RNA by Splicing Factor 1. Science *294,* 1098-1102.

Lodish, H., Berk, A., Matsudaira, P., and Kaiser, C.A. (2004). Molecular Cell Biology 5th Edition, Modern Genetic Analysis 2nd Edition & Cd-rom Macmillan Higher Education).

Lu, W., Feng, F., Xu, J., Lu, X., Wang, S., Wang, L., Lu, H., Wei, M., Yang, G., Wang, L.*, et al.* (2014). QKI impairs self-renewal and tumorigenicity of oral cancer cells via repression of SOX2. Cancer Biology & Therapy *15,* 1174-1184.

lumen. (2013a). Eukaryotic Transcription | Boundless Biology.*2018,*

lumen. (2013b). RNA Processing in Eukaryotes | Boundless Biology.*2018,*

Maass, P., Glažar, P., Memczak, S., Dittmar, G., Hollfinger, I., Schreyer, L., Sauer, A., Toka, O., Aiuti, A., Luft, F., and Rajewsky, N. (2017). A map of human circular RNAs in clinically relevant tissues. J Mol Med *95,* 1179-1189.

McInnes, L.A., and Lauriat, T.L. (2006). RNA metabolism and dysmyelination in schizophrenia. Neuroscience & Biobehavioral Reviews *30,* 551-561.

Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., and Proudfoot, N.J. (2008). Primary microRNA transcripts are processed co-transcriptionally. Nature Structural & Molecular Biology *15,* 902-909.

NCBI Gene. (2018). H2AFY H2A histone family member Y [Homo sapiens (human)] - Gene - NCBI.

Novikov, L., Park, J.W., Chen, H., Klerman, H., Jalloh, A.S., and Gamble, M.J. (2011). QKI-Mediated Alternative Splicing of the Histone Variant MacroH2A1 Regulates Cancer Cell Proliferation. Molecular and Cellular Biology *31,* 4244.

Ohe, K., and Mayeda, A. (2010). HMGA1a Trapping of U1 snRNP at an Authentic 5′ Splice Site Induces Aberrant Exon Skipping in Sporadic Alzheimer's Disease. Molecular and Cellular Biology *30,* 2220-2228.

Peled-Zehavi, H., Berglund, J.A., Rosbash, M., and Frankel, A.D. (2001). Recognition of RNA Branch Point Sequences by the KH Domain of Splicing Factor 1 (Mammalian Branch Point Binding Protein) in a Splicing Factor Complex. Molecular and Cellular Biology *21,* 5232-5241.

Piwecka, M., Glažar, P., Hernandez-Miranda, L.R., Memczak, S., Wolf, S.A., Rybak-Wolf, A., Filipchyk, A., Klironomos, F., Cerda Jara, C.A., Fenske, P.*, et al.* (2017). Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science *357,*

Ponder, J. (2015). Tinker 7.1.2.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26,* 841-842.

Ram, O., and Ast, G. (2007). SR proteins: a foot on the exon before the transition from intron to exon definition. Trends in Genetics *23,* 5-7.

Richard, S., and Galarneau, A. (2005). Target RNA motif and target mRNAs of the Quaking STAR protein. Nature Structural & Molecular Biology *12,* 691-698.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R.*, et al.* (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Mol. Cell *58,* 870-885.

Salzman, J. (2016). Circular RNA Expression: Its Potential Regulation and Function. Trends in Genetics *32,* 309-316.

Sanford, J.R., Longman, D., and Cáceres, J.F. (2003). Multiple roles of the SR protein family in splicing regulation. Progress in Molecular and Subcellular Biology *31,* 33.

Schrödinger. (2018). The PyMOL Molecular Graphics System, Version 1.2r3pre.

Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer*, et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495,* 333-338.

Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003). Structural Basis for the Molecular Recognition between Human Splicing Factors U2AF65 and SF1/mBBP. Molecular Cell *11,* 965-976.

Shao, Y., and Chen, Y. (2016). Roles of Circular RNAs in Neurologic Disease. Front. Mol. Neurosci. *9,*

Shimoyama, M., De Pons, J., Hayman, G.T., Laulederkind, S.J.F., Liu, W., Nigam, R., Petri, V., Smith, J.R., Tutaj, M., Wang, S.*, et al.* (2015). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res *43,* D750.

Sidman, R.L., Dickie, M.M., and Appel, S.H. (1964). Mutant Mice (Quaking and Jimpy) with Deficient Myelination in the Central Nervous System. Science *144,* 309-311.

Signal, B., Gloss, B.S., Dinger, M.E., Mercer, T.R., and Hancock, J. (2018). Machine learning annotation of human branchpoints. Bioinformatics *34,* 920-927.

Skandalis, A., Frampton, M., Seger, J., and Richards, M.H. (2010). The adaptive significance of unproductive alternative splicing in primates. RNA (New York, N.Y.) *16,* 2014-2022.

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

Teplova, M., Hafner, M., Teplov, D., Essig, K., Tuschl, T., and Patel, D.J. (2013). Structure–function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. Genes Dev. *27,* 928-940.

Tyner, C., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C.M., Gibson, D., Gonzalez, J.N., Guruvadoo, L.*, et al.* (2017). The UCSC Genome Browser database: 2017 update. Nucleic Acids Res *45,* D634.

Wang, Y., Vogel, G., Yu, Z., and Richard, S. (2013). The QKI-5 and QKI-6 RNA Binding Proteins Regulate the Expression of MicroRNA 7 in Glial Cells. Molecular and Cellular Biology *33,* 1233-1243.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA (New York, N.Y.) *14,* 802-813.

Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Molecular Cell *2,* 135-140.

Wilson, D.S., and Szostak, J.W. (1999). In Vitro Selection of Functional Nucleic Acids. Annual Review of Biochemistry *68,* 611-647.

Wu, J.I., Reed, R.B., Grabowski, P.J., and Artzt, K. (2002). Function of quaking in Myelination: Regulation of Alternative Splicing. Proceedings of the National Academy of Sciences of the United States of America *99,* 4233-4238.

Xiaofeng Zhang, Chuangye Yan, Xiechao Zhan, Lijia Li, Jianlin Lei, and Yigong Shi. (2018). Structure of the human activated spliceosome in three conformational states. Cell Research *28,* 307-322.

Zong, F., Fu, X., Wei, W., Luo, Y., Heiner, M., Cao, L., Fang, Z., Fang, R., Lu, D., Ji, H., and Hui, J. (2014). The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. PLoS Genetics *10,* e1004289.