

Real-time Visualization of Spliceosome Assembly Reveals Basic Principles of 3' Splice Site Selection

Dan Larson

dan.larson@nih.gov

NIH <https://orcid.org/0000-0001-9253-3055>

Benjamin Donovan

NIH

Bixuan Wang

University of Maryland

Gloria Garcia

NIH <https://orcid.org/0000-0003-1894-0079>

Stephen Mount

University of Maryland

Article

Keywords:

Posted Date: July 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8176735/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

1
2 **TITLE: REAL-TIME VISUALIZATION OF SPLICEOSOME ASSEMBLY REVEALS BASIC**
3 **PRINCIPLES OF 3' SPLICE SITE SELECTION**
4

5 **Authors:** Benjamin T Donovan¹, Bixuan Wang^{1,2}, Gloria R Garcia^{1,3}, Stephen M Mount², Daniel
6 R Larson^{1,3}
7

8 **Affiliations:**

9 ¹Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA

10 ²Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD
11 20742, USA

12 ³NIH Myeloid Malignancies Program, Bethesda, MD 20892, USA
13
14

15 **ABSTRACT:** The spliceosome is a megadalton protein-RNA complex which removes introns
16 from pre-mRNA, yet the dynamic early assembly steps have not been structurally resolved.
17 Specifically, how the spliceosome selects the correct 3' splice site (3'SS) amongst highly similar
18 non-functional sites is not known. Here, we develop a kinetic model of splice site selection
19 based on single-molecule U2AF heterodimer imaging *in vitro* and *in vivo*. The model
20 successfully predicts alternative splicing patterns and indicates that 3'SS commitment occurs
21 while U2AF is in complex with the spliceosome, not during initial binding. This finding indicates
22 the spliceosome operates in a 'partial' kinetic proofreading regime, catalyzed in part by the
23 helicase DDX42, which increases selectivity to the underlying U2AF binding site while still
24 allowing for efficient forward progression.
25
26
27

28 **ONE-SENTENCE SUMMARY:**
29

30 We apply a kinetic proofreading model to elucidate how transient U2AF binding leads to high
31 fidelity splice site selection.
32
33
34
35

MAIN TEXT

RNA splicing is catalyzed by the spliceosome, a megadalton enzyme¹ consisting of RNA and hundreds of protein components, each serving a role in regulating splicing accuracy and speed. This complexity enables very fine regulation as nearly all human pre-mRNAs undergo some form of alternative splicing². Additionally, despite being a single turn-over enzyme, the spliceosome also exhibits exceptionally high enzymatic activity; recent work indicates that a single intron may be removed in multiple splicing reactions, through a process called recursive splicing^{3,4}. While both alternative and recursive splicing are regulated at the step of initial splice site recognition, there is no unifying mechanism to describe the basis of splice site choice in these different contexts. Moreover, despite the tremendous advances in our structural understanding of the spliceosome⁵, many aspects of the early steps of spliceosome assembly remain unresolved, likely due to the transient nature of these complexes. Here, we use a combination of equilibrium binding assays and time-resolved kinetic assays both *in vitro* and *in vivo* to elucidate splice site selection and splicing fidelity.

3' SPLICE SITES (3'SSs) CONTAIN HIGHLY DEGENERATE U2AF BINDING SITES

The early steps of splicing require recognition of the *cis*-acting sequences in pre-mRNA such as the 5' splice site (5'SS), the branchpoint (bp), and the 3' splice site (3'SS) by RNA binding proteins and snRNPs. Here, we focus on the kinetics of 3'SS selection by the U2AF complex. While the 5'SS is recognized directly by the U1 snRNP, the U2 snRNP is recruited to the 3'SS by the U2AF heterodimer and additional RNA binding proteins. U2AF, comprised of U2AF2 and U2AF1, targets the pyrimidine tract and the -AG dinucleotide marking the 3'SS, respectively^{6,7}. Together, these proteins play a fundamental role in spliceosome assembly from initial binding in the E complex to splice site selection in the A complex⁸⁻¹¹. However, despite the paramount importance of accurate splicing, U2AF binds promiscuously throughout a pre-mRNA^{4,12}. Therefore, we set out to determine the types of binding events that promote splicing progression at specific sites but not others.

We addressed this question first with RNA Bind-N-Seq (RBNS), a high-throughput *in vitro* binding assay to quantitatively measure U2AF binding preferences to every RNA 12mer¹³ (**Extended Data Fig. 1A**). In contrast to previous SELEX assays for U2AF^{6,7}, we reasoned that RBNS would provide new insight into the U2AF binding landscape to all potential sites, including weak and moderate affinity sites which may be lost after multiple rounds of enrichment required for SELEX. In this assay, full-length U2AF heterodimer purified from insect cells (**Fig. 1A**) is mixed with a random pool of RNAs. After equilibration, RNAs in complex with U2AF are isolated by U2AF1-FLAG immunoprecipitation (IP) and deep sequenced (**see methods**). We find that the best predictor of enrichment for a specific RNA is the number of uridines in the binding site (**Fig. 1B**), indicating that U2AF2 is the main determinant of overall binding affinity. Furthermore, while the majority of significantly enriched RNAs do not contain the -AG dinucleotide (**Fig. 1C**), we observe a direct interaction with the -AG sequence on a strong and weaker pY-tract via RNaseI footprinting (**Extended Data Fig. 1A-C**). Additionally, when we compare AG-containing sequences to otherwise identical GA-containing sequences in our RBNS data, we detect an AG preference when the AG is at the 3' end of the binding site (**Fig. 1D**).

Next, we used the computational model Probound¹⁴ to predict U2AF binding affinity to each site identified from a previous PAR-CLIP experiment⁴ using the *in vitro* RBNS data (**Fig. 1E**).

G, Extended Data Fig. 1D-J). There are strong qualitative similarities between the Probound binding model and enrichment. For example, while predicted affinity depends primarily on pY-tract strength, the model predicts subtle stabilization from the AG dinucleotide when it is positioned at the 3' end of the RNA (**Extended Data Fig. 1K**). We compared the Probound predictions (**Fig. 1E**) with *in vitro* gel shift experiments and observed a linear dependence over the range of affinities we tested (**Fig. 1F, Extended Data Fig. 2**) indicating agreement between the Probound binding model and relative U2AF binding affinity measured by gel shift. Surprisingly, we find that U2AF occupies highly degenerate binding sites *in vivo* and that annotated 3'SSs contain only marginally stronger U2AF binding sites than sites at other locations, as shown for example in *RAB7A* (**Fig. 1G**). Therefore, we sought to address how fractional differences in equilibrium binding affinity could result in observed splicing outcomes by examining U2AF binding kinetics to RNA.

We implemented a fluorescence assay to determine the kinetics underlying the U2AF binding affinity landscape. This approach, called **PIFE**¹⁵ (Protein Induced Fluorescence Enhancement), describes a phenomenon where protein binding to a site immediately neighboring a Cy3 fluorophore results in an increase in fluorescence emission. When titrating U2AF against the well characterized Adenovirus Major Late Promoter ('AdML') 3'SS labeled with a Cy3 fluorophore on its 3' end (**see methods**), we measure a 1.5-fold increase in fluorescence emission (**Fig. 1H**). Furthermore, the data can be fit to a non-cooperative Hill function and the resulting K_D closely matches the EMSA measurement $K_{D \text{ PIFE}} = 27 \pm 7 \text{ nM}$ (**Fig. 1I, $K_{D \text{ EMSA}} = 24 \text{ nM}$, Extended Data Fig. 3A-B**), indicating that this approach indeed detects U2AF binding. We extended this assay to the single-molecule level to measure binding kinetics where Cy3-labeled surface tethered RNAs are visualized by Total Internal Reflection Fluorescence microscopy (**TIRF**) (**Fig. 1J**). Time traces were fit to a two-state HMM model where high fluorescence corresponds to RNA in complex with U2AF and low fluorescence corresponds unoccupied RNA (**Fig. 1K**). We measured U2AF binding to two 3'SSs, AdML and a similar 3'SS but with three uridine mutations in the pY-tract (**Fig. 1L, Extended Data Fig. 3A-B, see methods for full sequence**). In the single-molecule assay, we measure essentially the same overall association rate (**Fig. 1M**) and a 1.5 ± 0.2 -fold slower dissociation rate from the AdML sequence (**Fig. 1N, Extended Data Fig. 3C-G**). Therefore, the reduction in binding affinity to the mutant sequence is due exclusively to faster dissociation from RNA. Taken together, we find that U2AF interacts for about 15 seconds to the 'ideal' AdML U2AF 3'SS *in vitro* (**Fig. 1N**). This result was confirmed with competition assays to show the specificity of the PIFE approach (**Extended Data Fig. 3F-G**) as well as by monitoring colocalization between fluorescent U2AF and RNA (**Extended Data Fig. 3I-L**). Considering a median strength U2AF site in human pre-mRNA is ~100-fold weaker (**Fig. 1O**), our analysis predicts U2AF binds on the sub-second time scale to bona fide 3'SSs *in vivo*. It is not known how these transient interactions lead to correct 3' SS selection against the background of highly degenerate sites in pre-mRNA.

SIMULTANEOUS OBSERVATION OF U2AF BINDING AND PRE-mRNA SPLICING IN LIVING CELLS

We speculated specificity is accomplished through cooperative interactions as the spliceosome assembles^{16,17}. To test this idea, we developed an approach to visualize formation of the A complex on endogenous pre-mRNAs. We used CRISPR to integrate a HALO-tag on the C-terminus of U2AF1 (**Fig. 2A**) or U2AF2 (**Extended Data Fig. 4A-I**) and tracked diffusion of single proteins at an acquisition rate of 100 Hz to resolve both bound and unbound events (**Fig. 2B-C, Extended Data Fig. 4**). While we do indeed observe a population of transient binding events on the sub-second time scale (**Fig. 2D**), we also observe a small population of stable binding events that persist for long periods and, when tracking U2AF-HALO over much longer time scales (20 minutes, 0.33 Hz), we observe a broad dwell time distribution that appears linear on a log-log plot

with the longest binding events occurring for over 100 seconds (**Fig. 2E, Extended Data Fig. 4D for kymographs**). These long dwell times, which extend past the timescales of U2AF binding high-affinity sequences *in vitro* (**Fig. 2E**), likely reflect U2AF binding RNA as a member of a larger, stabilizing complex. Indeed, by blocking A complex formation with pladienolide B (PB)¹⁸, we observe a shift in the dwell time distribution towards shorter dwell times (**Fig. 2F**).

We modeled this relationship as shown in **Fig. 2G**. This general model, motivated by previous biochemical and structural data, is intended to encompass a wide range of possible behavior including reversible, irreversible and proofreading steps. We obtain the following equation for the survival curve (**Extended Data Fig. 4K**):

$$\text{U2AF survival}(k_{\text{fwd } 2}, k_{\text{off Int.}}, k_{\text{off A}}, t) = e^{-t(k_{\text{fwd } 2} + k_{\text{off Int.}})} - \left(\frac{k_{\text{fwd } 2}}{k_{\text{fwd } 2} + k_{\text{off Int.}} - k_{\text{off A}}} \right) (e^{-t(k_{\text{fwd } 2} + k_{\text{off Int.}})} - e^{-t(k_{\text{off A}})}) \quad (1)$$

From this fit (see **Extended Data Fig. 4K** for constraints), we find that PB treatment leads to a 2-fold reduction in $k_{\text{fwd } 2}$ ($k_{\text{fwd } 2 - \text{PB}} = 0.017 \pm 0.004 \text{ s}^{-1}$, $k_{\text{fwd } 2 + \text{PB}} = 0.008 \pm 0.002 \text{ s}^{-1}$). Furthermore, the fit values for $k_{\text{off Int.}}$ and $k_{\text{off A}}$ correspond to lifetimes of 9.1 ± 0.4 seconds and 82 ± 21 seconds for the Intermediate and A states, respectively (**Figs. 2F black line, 2H**). In this model, the intermediate state is especially important for refining splice site selection because it constitutes a kinetically branched step, where U2AF may progress to the A complex or dissociate via the proofreading pathway. In contrast, there is only a single pathway out of the final state, which we assume corresponds to U2AF dissociation after successful A complex formation. These experiments establish that U2AF dissociation out of the intermediate state is ~7-fold faster than spliceosome forward progression, indicating that the majority of U2AF molecules in the intermediate state dissociate to terminate spliceosome assembly.

The prediction from this analysis is that U2AF binds for an extended period during splice site selection. Thus, we aimed to simultaneously image splicing and U2AF binding. We overcame the technical challenge of imaging both protein binding (which occurs for seconds) and pre-mRNA splicing (which occurs on the ~15 minute time scale) with orbital tracking spectroscopy¹⁹. First, we created a cell line containing, in addition to U2AF1-HALO, 24xMS2 stem loops integrated about 4 kb upstream of the first 3'SS in *MYH9* (**Fig. 2I**). When transcribed, GFP-MS2 coat proteins bind the stem loops enabling visualization of the RNA as a bright foci which disappears after splicing^{4,20}. Indeed, *MYH9* shows active transcription and splicing (**Fig. 2J**). We analyzed the fluorescent time traces by correlation analysis^{21,22} and measured an average 'time to splice' (τ_{splice}) of 19 ± 2 min, which corresponds to synthesis of the stem loops, elongation to the 3'SS, and subsequent intron removal (**Fig 2K, Extended Data Fig. 4L-M**).

Next, we tracked the *MYH9* TS in three dimensions using orbital tracking spectroscopy (**Fig. 2L, Extended Data Fig. 4N**) while also monitoring the corresponding signal in the U2AF channel. While most measurements begin in the 'high' state when the MS2 signal is visible and end after splicing (**Fig. 2M, top**), some tracks also show the full cycle of transcription and splicing (**Fig. 2M, bottom**). Autocorrelation of the U2AF signal reveals the timescales of U2AF binding in a productive splicing reaction and when splicing is stalled by PB (Control: 102 ± 6 s, PB: 78 ± 6 s) (**Fig. 2N**). Furthermore, the U2AF-RNA cross-correlation curves can be approximated with gaussian distributions where the width represents τ_{splice} (Control = 16 ± 1 min, PB = 27 ± 1 min) (**Fig. 2O**). The broadening of the cross-correlation curve after PB treatment represents slower overall splicing in this condition. Taken together, this data reveals the time scales of U2AF binding during *MYH9* splicing (**Fig. 2P**) and further supports a model where U2AF binds stably to actively spliced pre-mRNAs after progressing through transient initial states.

A KINETIC MODEL FOR SPLICE SITE SELECTION

The core element of our model is ATP-dependent irreversible forward progression ($k_{\text{fwd } 1}$ and $k_{\text{fwd } 2}$) in competition with U2AF dissociation (k_{off} and $k_{\text{off int}}$) (**Fig. 2G**) where we assume that both dissociation steps depend on underlying sequence. Repeating this scheme twice (around the E complex and Intermediate state) is a hallmark of kinetic proofreading and has the potential to improve the fidelity of splice site selection^{23–25} (**Figure S5C-D**). While past work shows that the spliceosome is reversible^{26,27}, here we assume that the backward rate is slow enough to be neglected. We next tested the ability of this model to predict alternative splicing measured by RNA-seq in HBECs. We were particularly interested in alternative 3'SS selection where two closely spaced (>10 nt) 3'SSs compete for U2AF. Applying the model to alternative 3'SS selection (**Fig. 3A, 3B**), we get the following equation for Percent Spliced In (**PSI**) (see **Extended Data Fig. 5A for master equation**):

$$PSI(K_f, K_s, \Delta\Delta G) = \frac{(K_f + 1)(K_s + 1)}{(2 K_f K_s) + (K_f + K_s) \left(1 + e^{\frac{-\Delta\Delta G}{RT}}\right) + \left(1 + e^{\frac{-2\Delta\Delta G}{RT}}\right)} \quad (2)$$

In this equation, $\Delta\Delta G$ (calculated from Probound binding model) reflects the difference in U2AF binding affinity at a pair of potential 3'SSs, where $-\Delta\Delta G = RT \ln \left(\frac{k_{\text{off } S2}}{k_{\text{off } S3}}\right)$ (**Fig. 3A**). Furthermore, the normalized parameters K_f ($k_{\text{fwd } 1} / k_{\text{off}}$) and K_s ($k_{\text{fwd } 2} / k_{\text{off int}}$) (measured in SMT assays) reflect the importance of U2AF binding site strength in determining splicing outcome. For example, if U2AF dissociates more quickly than forward progression (K_f and $K_s \ll 1$), spliceosome advancement to the next assembly intermediate will be a rare event, with most binding events leading to unproductive dissociation. In this case, frequent binding site sampling means that the probability of achieving a specific mRNA isoform depends on the $-\Delta\Delta G$ of U2AF binding (**Fig. 3C, maroon curve**). Importantly, splicing decisions in this 'full proofreading' model achieve higher sensitivity to subtle differences in underlying binding affinity than a simple kinetic model where PSI depends *only* on relative U2AF binding affinity at a single state (**Fig. 3C, gray, 'thermodynamic expectation'**). In contrast, if forward progression is much faster than U2AF dissociation (K_f and $K_s \gg 1$), $PSI \sim 1/2$, and there will be no specificity based on binding site strength, because *any* binding event leads to forward progression (**Fig. 3C, salmon curve**). We call this regime 'no proofreading.'

We tested this model by comparing the median PSI of groups of alternative 3'SSs binned by $-\Delta\Delta G$ calculated from RBNS both in the absence and presence of HeLa NE. Indeed, in the latter case, we now detect the branch point ([Y]UNAY) upstream of the pyrimidine-tract (**Fig. 3D**), demonstrating how additional factors in the extract, such as SF1 and SF3B1 interact with U2AF to promote recognition of genuine 3'SSs over extended pyrimidine-tract motifs. Notably, while it is possible to generate a U2AF binding model that contains an AG by removing sequencing reads with high-affinity binding sites (**Extended Data Fig. 1L-M**), there is no correlation between this model's predicted binding affinity and PSI (**Extended Data Fig. 1N**), suggesting a minor role for U2AF1 binding in initial splice site recognition compared to U2AF2 and branch point binders. Nevertheless, for primary models (**Fig 3D**) we observe a clear sigmoidal relationship between PSI and $-\Delta\Delta G$ for both models (**Fig. 3E**) despite high variation at the level of individual transcripts (**Extended Data Fig. 5B**). The overall shape indicates both equilibrium and non-equilibrium elements and suggests that alternative 3'SS selection does not operate in the 'full proofreading' regime where both K_f and $K_s \ll 1$ and specificity is optimized. Rather, using the measured $K_s = 0.15$ from the slow SMT assay (**Fig. 2H**), we find that the PSI dependence is best fit with $K_f = 1.0$

± 0.5 using the NE model and $K_f = 20 \pm 10$ for the control model (**Fig. 3E**). Strikingly, these data indicate first that NE refines specificity by *limiting* forward progression compared to control, and second that 3'SS selection primarily occurs in the intermediate state ($K_s \ll 1$). Taken together, the spliceosome operates in a 'partial proofreading' regime where U2AF binding leads to rapid initial progression toward the A complex with 3'SS selection occurring *after* initial binding during the intermediate step.

We therefore focused subsequent analysis on understanding the additional spliceosome components that influence U2AF sequence preference during this intermediate step with several approaches. First, we sought to identify additional spliceosome components that influence U2AF sequence preference during the intermediate state by re-analyzing extensive knockdown data for nearly all components of the spliceosome²⁸. We repeated the above analysis (plotting PSI vs. - $\Delta\Delta G$ and fitting using our alternative splicing model (**Fig. 3F**)) to infer kinetic changes from steady state RNA-seq for all ~300 KD experiments (**Fig. 3G**). Notably, while most spliceosome proteins have no global effect on K_s , U2 snRNP components, especially proteins situated close to the pyrimidine tract (**Fig. 3H**), show decreased K_s , suggesting either slower spliceosome progression ($k_{fwd, 2}$) or faster U2AF dissociation ($k_{off, int}$) after protein depletion. Additionally, we observe that U2AF1 knockdown decreases K_s similarly as U2AF2 and U2 snRNP components, which is notable given the minimal role of the AG preference in our Probound analysis (**Fig. 3D**). This observation of aberrant splicing after U2AF1 knockdown emphasizes the importance of U2AF1 beyond its binding to the 3'SS, perhaps by stabilizing U2AF2^{29–32} or by interacting with other splicing regulators^{33,34}.

The main prediction of this analysis is that knockdown of U2 snRNP components slow spliceosome progression. We verified this prediction by SF3B1 knockdown in our live-cell U2AF imaging system and indeed observed faster U2AF dissociation and slower forward progression (**Fig. 3I**). To determine the overall impact on splicing kinetics, we measured splicing dynamics globally using a polyclonal cell line where MS2 stem loops are inserted into 728 unique introns⁴. Overall, we observe slower splicing rate after SF3B1 knockdown (**Fig. 3J, Extended Data Fig. 5I**). In summary, the predictive power of our kinetic model in an orthogonal ensemble steady state dataset suggests future opportunities for leveraging genome-scale perturbations with single-molecule imaging of splicing dynamics.

Finally, we devise a measure of splicing accuracy in our model by considering the scenario where U2AF must identify the correct annotated 3'SS with respect to nearby downstream sites, for example in coding sequences. This selection cannot be identified empirically by RNA-seq due to quality control of mis-spliced transcripts. We calculated a metric called 'predicted splicing accuracy' (PSA) (**Fig. 3K**) – analogous to PSI – and compared this quantity to the thermodynamic expectation where PSA is only due to equilibrium binding affinity. PSA was computed by using the Probound $\Delta\Delta G$ for a U2AF PAR-CLIP peak at the 3'SS vs. the nearest downstream exon site. Importantly, the thermodynamic expectation describes a scenario where spliceosome forward progression is infinitely slow so that equilibrium can be reached. In contrast, our partial proofreading model achieves a higher predicted splicing accuracy despite allowing for efficient spliceosome forward progression, highlighting the tradeoff between speed and accuracy.

DDX42 PROMOTES SPLICING SPECIFICITY BY ACCELERATING U2AF DISSOCIATION

While the knockdown screen (**Fig. 3F-G**) identified many basal A-complex components, it is often hypothesized that proofreading requires a helicase^{35–38} or auxiliary splicing factors^{39–41}. Therefore, we sought to identify additional proteins that interact with U2AF during spliceosome assembly and to apply our single-molecule approach to reveal how these proteins influence spliceosome

progression. More specifically, we used CRISPR to insert a 3XFLAG-tag on the C-terminus of U2AF1 and performed Immunoprecipitation-Mass Spectrometry (IP-MS) on HBEC nuclear extracts⁴². Surprisingly, the most enriched splicing factor was DDX42, a relatively uncharacterized helicase that interacts with SF3B1 during the spliceosome assembly process^{43–45} (**Fig. 4A-C**). Importantly, this protein was not present in the original knockdown screen.

We performed SMT assays after DDX42 knockdown to determine if this protein influences U2AF binding kinetics and subsequent spliceosome assembly (**Extended Data Fig. 6A**). Fast SMT assays indicate DDX42 knockdown has no impact on the time scales of transient sub-second U2AF-RNA interactions (**Extended Data Fig. 6B-C**). However, the slow SMT assay, which only detects stable U2AF binding, including U2AF in complex with the spliceosome, shows slower U2AF dissociation across time scales ranging from 10-300 seconds after DDX42 knockdown (**Fig. 4D, Extended Data Fig. 6D**) but not after depletion of another helicase believed to be involved in formation of the early spliceosome, DDX39B⁴⁶ (**Fig. 4E, Extended Data Fig. 6E**). These results suggest that DDX42 interacts with U2AF during the process of spliceosome assembly, not during initial binding site sampling. We fit the slow SMT data to the same splice model as in **Fig 2G (Eqn. 1, Fig. 4D, Extended Data Fig. 6F-G)** which shows a subtle, $18 \pm 7\%$ decrease in the U2AF dissociation rate from the Intermediate state ($k_{\text{off int}}$) and a larger $44 \pm 13\%$ decrease in U2AF dissociation from the A complex ($k_{\text{off A}}$). Therefore, in contrast to pladienolide B which reduces $k_{\text{fwd}2}$ (**Fig. 2F-H**), DDX42 depletion is best explained with a reduction in $k_{\text{off int}}$ and $k_{\text{off A}}$. One potential interpretation of this data is that DDX42 facilitates U2AF dissociation and competes with forward progression in a similar manner to other spliceosome helicases^{35,47–49}. When forward progression is sufficiently slow, DDX42 facilitates U2AF dissociation from the intermediate state and disassembles the spliceosome via the proofreading pathway. However, when spliceosome forward progression is efficient, DDX42 will facilitate U2AF dissociation from the A complex, enabling spliceosome progression. Thus, DDX42 may act as a ‘molecular timer’ where slowly progressing spliceosomes are subject to proofreading.

In our simple model of splice site selection, specificity is refined at two steps (1) during initial binding and (2) during the intermediate state where U2AF may dissociate via the proofreading pathway or progress to the A complex. Importantly, since there is only one pathway out of the A complex, there will be no specificity refinement at this stage. Thus, we sought to better understand the consequences of slowed dissociation from the intermediate state on alternative splicing outcomes. Our kinetic model of alternative splice site selection (**Eqn. 2**) predicts that the ~20% increase in K_s measured by slow SMT upon DDX42 knockdown will result in a global reduction in PSI for alternative 3'SS selection, especially in the negative $-\Delta\Delta G$ regime where U2AF binds the upstream 3'SS site more tightly (**Fig. 4F**, dashed red curve). To test this prediction, we performed RNA-seq after DDX42 knockdown and applied our alternative splicing analysis. Indeed, when binning ΔPSI values for alternative 3'SSs after DDX42 knockdown into groups of similar relative U2AF binding affinity (**Fig. 4F**, open circles), we observe strong agreement with our SMT data, both in terms of the magnitude of splicing changes and in terms of the trend of the relationship. Importantly, this result establishes that a subtle change in U2AF dissociation from the intermediate state is sufficient to explain global changes in alternative 3'SS selection. Overall, we conclude that DDX42 evicts U2AF from pre-mRNA from the intermediate state. Intriguingly, a recent structural study positions the DDX42 “N-plug” close to the pyrimidine tract⁴⁴. Therefore, we speculate that this region may be responsible for accelerated U2AF dissociation (**Fig. 4G**).

Finally, we interrogated the role of DDX42 in spliceosome progression using additional live-cell assays. In the orbital tracking assay, we observed a substantially slower U2AF dissociation after DDX42 knockdown ($\tau_{\text{U2AF control}} = 1.7 \pm 0.1$ min, $\tau_{\text{U2AF siDDX42}} = 5.8 \pm 0.2$ min) (**Fig.**

4I). We also observed a broader cross-correlation distribution ($\tau_{\text{U2AF control}} = 16 \pm 1$ min, $\tau_{\text{U2AF siDDX42}} = 30 \pm 1$ min) indicative of slower overall splicing (Fig. 4J, Fig. 4K). Lastly, to determine the pervasiveness of this effect, we measured global changes in splicing kinetics using the polyclonal cell line with many MS2 integrations in introns. Overall, we find that DDX42 knockdown results in a global slowing of pre-mRNA splicing (Fig. 4L-M). In summary, we conclude that knockdown of DDX42 results in slower U2AF dissociation, slower overall spliceosome progression, and consequently changes in alternative splicing.

DISCUSSION

Here we illuminate the dynamic operating principles of the early spliceosome. At the heart of this description is a basic kinetic module consisting of equilibrium association of an RNA binding protein, followed by an irreversible, energy-consuming transition to a complex that can be proofread and disassembled as needed (e.g. Fig. 2G). We find that 3'SS generally contain weak U2AF binding sites: U2AF initially binds a potential 3'SS for less than a second (Fig. 1O). Despite these weak sites, the spliceosome is remarkably sensitive to subtle differences in U2AF binding affinity. U2AF recognition has thus adopted a strategy not on discrimination at the initial bound state but on rapid progression to an intermediate which is subject to proofreading. This strategy may be successful because U2AF is highly expressed in the nucleus which could enable very fast association to pre-mRNAs⁵⁰. Furthermore, while operating in 'full' proofreading mode ($K_r \ll 1$, $K_s \ll 1$) would achieve the best binding site discrimination (Fig. 4M), it comes at the cost of inefficient progression (Fig. 4N): our data suggests that about six out of seven spliceosomes are disassembled at the 'intermediate' state. Therefore, the partial proofreading model seems to both (i) discriminate based on binding affinity better than possible for the simple thermodynamic model while (ii) enabling efficient forward spliceosome progression.

The model we propose here is a plausible model based on kinetic proofreading. Remarkably, it unites *in vitro* binding affinity and *in vivo* protein dynamics to quantitatively predict splicing outcomes and mathematically describe how splicing accuracy is achieved in long human introns using the PSA metric. We do not eliminate the possibility that other model topologies would achieve similar results. However, this kinetic proofreading model which is consistent with our data, is also able to incorporate new biochemical findings, including those from an extensive knockdown screen²⁸. For example, we identified the protein components of the intermediate state where proofreading occurs as consisting of U2AF in complex the U2 snRNP (Fig. 3G). Moreover, we discovered a new biochemical interaction between DDX42 and U2AF1 and were able to propose and validate a function based on how knockdown of this factor changes U2AF1 dynamics and 3'SS selection. However, while DDX42 participates in this process by accelerating U2AF dissociation, there are still open questions related to how it impedes spliceosome assembly at off-target sites. The simplest model would be that DDX42 binds U2AF and consumes ATP which dissociates U2AF off RNA. In this view, the dissociation of the intermediate state depends both on the intrinsic U2AF binding affinity and DDX42 activity.

While proofreading mechanisms have long been implicated in splicing^{48,49,51,52}, historically the focus has been on surveillance at later steps to ensure fidelity of the chemical reactions in splicing, i.e. branching and exon-exon ligation. Here we show how inefficiency is essential at the earliest steps of splice site selection to distinguish highly degenerate motifs abundantly present in human introns. In our study, we establish the timing and pervasiveness of an essential proofreading step in early spliceosome assembly which has yet to be structurally resolved. We identify a crucial ~10 second window where U2AF binds in complex with the U2 snRNP and discriminates between highly similar splice sites by continuous re-sampling. As such, the driving operating principle of the human early spliceosome is inefficiency: by increasing the number of

times spliceosomes must assemble at a particular set of competing splice sites, extremely subtle differences in U2AF affinity can be distinguished. More broadly, this work sets the stage to fundamentally understand the dynamic regulation of splicing, how splicing decisions are made in single cells, and how this regulation is altered in disease.

REFERENCES AND NOTES:

1. Hoskins, A. A. & Moore, M. J. The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem Sci* **37**, 179–188 (2012).
2. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
3. Sibley, C. R. *et al.* Recursive splicing in long vertebrate genes. *Nature* **521**, 371–375 (2015).
4. Wan, Y. *et al.* Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell* **184**, 2878–2895.e20 (2021).
5. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annual Review of Biochemistry* **89**, 359–388 (2020).
6. Singh, R., Valcárcel, J. & Green, M. R. Distinct Binding Specificities and Functions of Higher Eukaryotic Polypyrimidine Tract-Binding Proteins. *Science* **268**, 1173–1176 (1995).
7. Wu, S., Romfo, C. M., Nilsen, T. W. & Green, M. R. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* **402**, 832–835 (1999).
8. Agafonov, D. E. *et al.* Semiquantitative Proteomic Analysis of the Human Spliceosome via a Novel Two-Dimensional Gel Electrophoresis Method. *Molecular and Cellular Biology* **31**, 2667–2682 (2011).
9. Kotlajich, M. V., Crabb, T. L. & Hertel, K. J. Spliceosome Assembly Pathways for Different Types of Alternative Splicing Converge during Commitment to Splice Site Pairing in the A Complex. *Mol Cell Biol* **29**, 1072–1082 (2009).
10. Gozani, O., Potashkin, J. & Reed, R. A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site. *Mol Cell Biol* **18**, 4752–4760 (1998).
11. Galardi, J. W. *et al.* A UHM–ULM interface with unusual structural features contributes to U2AF2 and SF3B1 association for pre-mRNA splicing. *Journal of Biological Chemistry* **298**, (2022).
12. Shao, C. *et al.* Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol* **21**, 997–1005 (2014).
13. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell* **54**, 887–900 (2014).
14. Rube, H. T. *et al.* Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning. *Nat Biotechnol* **40**, 1520–1527 (2022).
15. Hwang, H., Kim, H. & Myong, S. Protein induced fluorescence enhancement as a single molecule assay with short distance sensitivity. *Proc Natl Acad Sci U S A* **108**, 7414–7418 (2011).
16. Chen, L. *et al.* Stoichiometries of U2AF35, U2AF65 and U2 snRNP reveal new early spliceosome assembly pathways. *Nucleic Acids Res* gkw860 (2016) doi:10.1093/nar/gkw860.
17. Hansen, S. R. *et al.* Multi-step recognition of potential 5' splice sites by the *Saccharomyces cerevisiae* U1 snRNP. *eLife* **11**, e70534 (2022).

18. Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nat Chem Biol* **3**, 570–575 (2007).
19. Kis-Petikova, K. & Gratton, E. Distance measurement by circular scanning of the excitation beam in the two-photon microscope. *Microsc. Res. Tech.* **63**, 34–49 (2004).
20. Bertrand, E. *et al.* Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell* **2**, 437–445 (1998).
21. Coulon, A. & Larson, D. R. Fluctuation Analysis. in *Methods in Enzymology* vol. 572 159–191 (Elsevier, 2016).
22. Coulon, A. *et al.* Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife* **3**, e03939 (2014).
23. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proceedings of the National Academy of Sciences* **71**, 4135–4139 (1974).
24. Burgess, S. M. & Guthrie, C. Beat the clock: paradigms for NTPases in the maintenance of biological fidelity. *Trends in Biochemical Sciences* **18**, 381–384 (1993).
25. Xu, Y.-Z. & Query, C. C. Competition between the ATPase Prp5 and Branch Region-U2 snRNA Pairing Modulates the Fidelity of Spliceosome Assembly. *Molecular Cell* **28**, 838–849 (2007).
26. Tseng, C.-K. & Cheng, S.-C. Both Catalytic Steps of Nuclear Pre-mRNA Splicing Are Reversible. *Science* **320**, 1782–1784 (2008).
27. Hoskins, A. A. *et al.* Ordered and Dynamic Assembly of Single Spliceosomes. *Science* **331**, 1289–1295 (2011).
28. Rogalska, M. E. *et al.* Transcriptome-wide splicing network reveals specialized regulatory functions of the core spliceosome. *Science* **386**, 551–560 (2024).
29. Mackereth, C. D. *et al.* Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature* **475**, 408–411 (2011).
30. Voith von Voithenberg, L. *et al.* Recognition of the 3' splice site RNA by the U2AF heterodimer involves a dynamic population shift. *Proc Natl Acad Sci U S A* **113**, E7169–E7175 (2016).
31. Warnasooriya, C., Feeney, C. F., Laird, K. M., Ermolenko, D. N. & Kielkopf, C. L. A splice site-sensing conformational switch in U2AF2 is modulated by U2AF1 and its recurrent myelodysplasia-associated mutation. *Nucleic Acids Research* **48**, 5695–5709 (2020).
32. Guth, S., Tange, T. O., Kellenberger, E. & Valcárcel, J. Dual Function for U2AF35 in AG-Dependent Pre-mRNA Splicing. *Mol Cell Biol* **21**, 7673–7681 (2001).
33. Graveley, B. R., Hertel, K. J. & Maniatis, T. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* **7**, 806–818 (2001).
34. Wang, Z., Hoffmann, H. M. & Grabowski, P. J. Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity. *RNA* **1**, 21–35 (1995).
35. Feng, Q., Krick, K., Chu, J. & Burge, C. B. Splicing quality control mediated by DHX15 and its G-patch activator SUGP1. *Cell Reports* **42**, 113223 (2023).
36. Zhang, J. *et al.* Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Molecular Cell* **76**, 82–95.e7 (2019).
37. Beusch, I. *et al.* Targeted high-throughput mutagenesis of the human spliceosome reveals its in vivo operating principles. *Mol Cell* **83**, 2578–2594.e9 (2023).

475 38. Maul-Newby, H. M. *et al.* A model for DHX15 mediated disassembly of A-complex
476 spliceosomes.

477 39. Soares, L. M. M., Zanier, K., Mackereth, C., Sattler, M. & Valcárcel, J. Intron removal requires
478 proofreading of U2AF/3' splice site recognition by DEK. *Science* **312**, 1961–1965 (2006).

479 40. Tavanez, J. P., Madl, T., Kooshapur, H., Sattler, M. & Valcárcel, J. hnRNP A1 proofreads 3'
480 splice site recognition by U2AF. *Mol Cell* **45**, 314–329 (2012).

481 41. Zarnack, K. *et al.* Direct Competition between hnRNP C and U2AF65 Protects the
482 Transcriptome from the Exonization of Alu Elements. *Cell* **152**, 453–466 (2013).

483 42. Garcia, G. R. *et al.* U2AF regulates the translation and localization of nuclear-encoded
484 mitochondrial mRNAs. *Molecular Cell* **86**, 1293-1310.e14 (2026).

485 43. Will, C. L. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human
486 Prp5p homologue and an SF3b DEAD-box protein. *The EMBO Journal* **21**, 4978–4988 (2002).

487 44. Yang, F. *et al.* Mechanisms of the RNA helicases DDX42 and DDX46 in human U2 snRNP
488 assembly. *Nat Commun* **14**, 897 (2023).

489 45. Lazear, M. R. *et al.* Proteomic discovery of chemical probes that perturb protein complexes
490 in human cells. *Molecular Cell* **83**, 1725-1742.e12 (2023).

491 46. Fleckner, J., Zhang, M., Valcárcel, J. & Green, M. R. U2AF65 recruits a novel human DEAD
492 box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev* **11**, 1864–1872
493 (1997).

494 47. Tholen, J., Razew, M., Weis, F. & Galej, W. P. Structural basis of branch site recognition by the
495 human spliceosome. *Science* **375**, 50–57 (2022).

496 48. Koodathingal, P., Novak, T., Piccirilli, J. A. & Staley, J. P. The DEAH box ATPases Prp16 and
497 Prp43 cooperate to proofread 5' splice site cleavage during pre-mRNA splicing. *Mol Cell* **39**,
498 385–395 (2010).

499 49. Semlow, D. R., Blanco, M. R., Walter, N. G. & Staley, J. P. Spliceosomal DEAH-Box ATPases
500 Remodel Pre-mRNA to Activate Alternative Splice Sites. *Cell* **164**, 985–998 (2016).

501 50. Cho, N. H. *et al.* OpenCell: Endogenous tagging for the cartography of human cellular
502 organization. *Science* **375**, eabi6983 (2022).

503 51. Mayas, R. M., Maita, H., Semlow, D. R. & Staley, J. P. Spliceosome discards intermediates via
504 the DEAH box ATPase Prp43p. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10020–10025 (2010).

505 52. Mayas, R. M., Maita, H. & Staley, J. P. Exon ligation is proofread by the DExD/H-box ATPase
506 Prp22p. *Nat Struct Mol Biol* **13**, 482–490 (2006).

507 53. Cvitkovic, I. & Jurica, M. S. Spliceosome Database: a tool for tracking components of the
508 spliceosome. *Nucleic Acids Research* **41**, D132–D141 (2013).

509 54. Zhang, X. *et al.* Structure of the human activated spliceosome in three conformational
510 states. *Cell Res* **28**, 307–322 (2018).

511 55. Nilsen, T. W. RNase Footprinting to Map Sites of RNA–Protein Interactions. *Cold Spring Harb*
512 *Protoc* **2014**, pdb.prot080788 (2014).

513 56. Abmayr, S. M., Carrozza, M. J. & Workman, J. L. Preparation of Nuclear and Cytoplasmic
514 Extracts from Mammalian Cells. *CP Pharmacology* **12**, (2001).

515 57. Lambert, N. J., Robertson, A. D. & Burge, C. B. RNA Bind-n-Seq: Measuring the Binding
516 Affinity Landscape of RNA-Binding Proteins. *Methods Enzymol* **558**, 465–493 (2015).

58. Luo, G., Wang, M., Konigsberg, W. H. & Xie, X. S. Single-molecule and ensemble fluorescence assays for a functionally important conformational change in T7 DNA polymerase. *Proceedings of the National Academy of Sciences* **104**, 12610–12615 (2007).
59. Fu, X. *et al.* Identification of transient intermediates during spliceosome activation by single molecule fluorescence microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2206815119 (2022).
60. van de Meent, J.-W., Bronson, J. E., Wiggins, C. H. & Gonzalez, R. L. Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophys J* **106**, 1327–1337 (2014).
61. Elson, E. L. Fluorescence Correlation Spectroscopy: Past, Present, Future. *Biophysical Journal* **101**, 2855–2870 (2011).
62. Fei, D. L. *et al.* Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival. *PLoS Genet* **12**, e1006384 (2016).
63. Garcia, D. A. *et al.* Power-law behavior of transcription factor dynamics at the single-molecule level implies a continuum affinity model. *Nucleic Acids Research* **49**, 6605–6620 (2021).
64. Sergé, A., Bertaux, N., Rigneault, H. & Marguet, D. Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes. *Nat Methods* **5**, 687–694 (2008).
65. Persson, F., Lindén, M., Unoson, C. & Elf, J. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat Methods* **10**, 265–269 (2013).
66. Palangat, M. & Larson, D. R. Single-gene dual-color reporter cell line to analyze RNA synthesis in vivo. *Methods* **103**, 77–85 (2016).
67. Keikhosravi, A. *et al.* High-throughput image processing software for the study of nuclear architecture and gene expression. *Sci Rep* **14**, (2024).
68. Burge, C. B., Tuschl, T. & Sharp, P. A. Splicing of Precursors to mRNAs by the Spliceosomes. *The RNA World, Second Edition* (1999).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
70. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
71. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology* **19**, 40 (2018).
72. Gagnon, K. T., Li, L., Janowski, B. A. & Corey, D. R. Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading. *Nat Protoc* **9**, 2045–2060 (2014).

AKNOWLEDGEMENTS

We thank members of the Larson laboratory for helpful discussions. We thank the many support staff at the NIH core facilities who maintained the instruments and provided helpful advice when conducting experiments, including Dave Ball (NCI Optical Microscopy Core), Jiji Chen (Advanced Imaging and Microscopy Resource), Di Wu (NHLBI Biophysics Core), Gianluca Pegoraro and Adib Keikhosravi (NCI High-Throughput Imaging Facility), Madeline Wong (CCR genomics core), the Frederick Protein Expression Laboratory, the CCR Protein Characterization Laboratory, the CCR sequencing facility, and the NIH High-Performance Computing Group (Biowulf). Finally, we thank Tucker Carrocci and Margaret Rodgers for help preparing flow cells for *in vitro* single-molecule assays.

FUNDING

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the United States Government. However, the findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

AUTHOR CONTRIBUTIONS

Conceptualization: BTB, DRL, SMM
Methodology: BTB, DRL, BW, GRG
Investigation: BTB, BW, GRG
Visualization: BTB, BW
Funding acquisition: DRL
Project administration: DRL, SMM
Supervision: DRL, SMM
Writing – original draft: BTB, DRL
Writing – review & editing: BTB, DRL, BW, GRG, SMM

COMPETING INTERESTS:

Authors declare no competing interests.

DATA AND MATERIALS AVAILABILITY

RNA sequencing data is uploaded to the NCBI Gene Expression Omnibus GEO (GSE275098). Imaging and RNA bind-N-seq will be available on Zenodo upon publication ([temporary link for review](#)). All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

SUPPLEMENTARY MATERIALS

Materials and Methods
Figs. S1 to S6

FIGURES

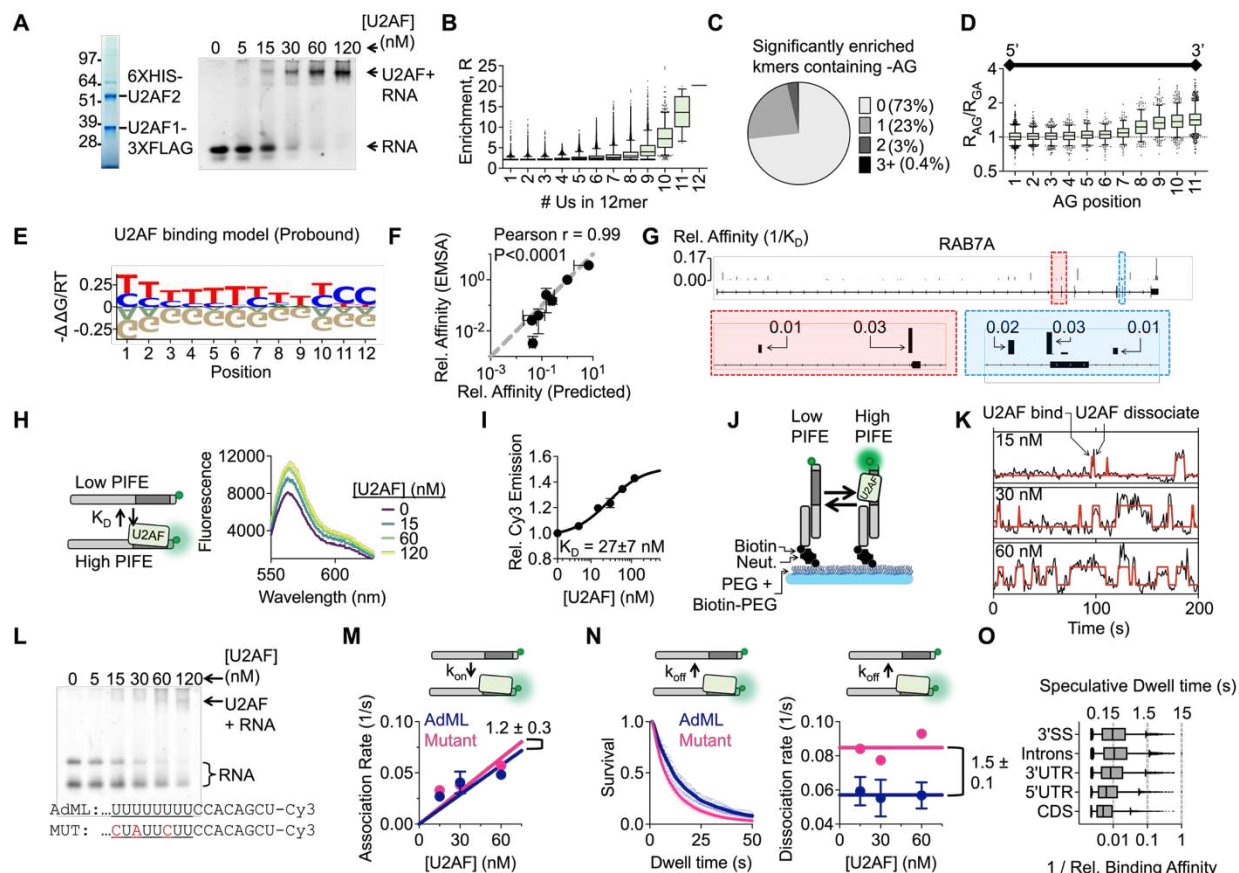


Figure 1: In vitro characterization of U2AF binding transcriptome wide

- A. Full length U2AF was purified from insect cells. We tested functionality by electrophoresis mobility shift assay (EMSA).
- B. Measuring U2AF binding to all possible 12mers in parallel with RNA Bind-N-Seq. U2AF heterodimer was mixed with a random pool of RNAs. Bound RNAs were isolated by U2AF1-FLAG IP and sequenced. The plot displays the distribution enrichment ($R = \text{Freq}_{\text{IP}} / \text{Freq}_{\text{input}}$) segregated by the number of uridines in a 12mer sequence (this data represents enrichment for 20 nM U2AF, $n=2$ replicates).
- C. The distribution of significantly enriched RNAs containing -AG dinucleotide.
- D. The enrichment ratio of an -AG containing sequence relative to an otherwise identical -GA containing sequence segregated by position of the -AG/-GA in the 12mer.
- E. Prediction of free energy contribution of each nucleotide using Probound.
- F. Validating Probound with competitive binding assays resolved by EMSA (see **Extended Data Fig. 2**). Error bars represent one standard deviation.
- G. Assigning U2AF binding affinity to *in vivo* binding sites detected with PAR-CLIP. In the browser shot of RAB7A, the height of each bar represents binding affinity relative to the sequence in (E).
- H. Adopting the PIFE assay to detect U2AF binding *in vitro* by bulk fluorescence. U2AF binding results in a ~1.5 fold increase in fluorescence ($n=2$ replicates). As a control, we also performed the assay in the absence of RNA and find that intrinsic U2AF autofluorescence contributes very little to measured emission.

- I. Plotting relative change in fluorescence when titrating U2AF. Fitting with non-cooperative Hill Function ($K_D = 27 \pm 7$ nM) agrees with EMSA measurements. Error bars represent one standard deviation.
- J. Applying the PIFE assay to the single-molecule level. RNAs are tethered to glass microscope slide through biotin-neutravidin linkage. Single surface-tethered RNAs are illuminated by Total Internal Fluorescence Reflection (TIRF) microscopy.
- K. U2AF binding and dissociation on individual RNAs is inferred from two-state Hidden Markov Model.
- L. EMSA of U2AF binding to an RNA with a mutated pyrimidine tract.
- M. U2AF binding rate to the AdML and mutant 3' SS ($k_{on\ AdML} = 0.0010 \pm 0.0001$ nM⁻¹ s⁻¹, $k_{on\ Mutant} = 0.0011 \pm 0.0002$ nM⁻¹ s⁻¹) (n = 3 replicates for AdML, n=2 replicates for Mutant 3'SS). Error bars represent one standard deviation.
- N. Left: plotting U2AF dwell times for all replicates to AdML (n=9 measurements) and Mutant (n=6 measurements) 3'SS. Dotted lines represent individual replicates and solid lines represent average of all replicates. Right: Comparing U2AF dissociation rate from the AdML and mutant 3'SS at three concentrations of U2AF. U2AF dissociates about 1.5-fold faster from the mutant 3'SS ($k_{off\ AdML} = 0.057 \pm 0.003$ s⁻¹, $k_{off\ Mutant} = 0.085 \pm 0.003$ s⁻¹). Error bars represent one standard deviation.
- O. Comparing the distribution of U2AF binding affinities (relative to consensus sequence in Fig. 1E) for different binding site types. This plot only contains binding affinities within the range that were validated in Fig. 1F. A speculative range of dwell times for these interactions is included on the top x-axis.

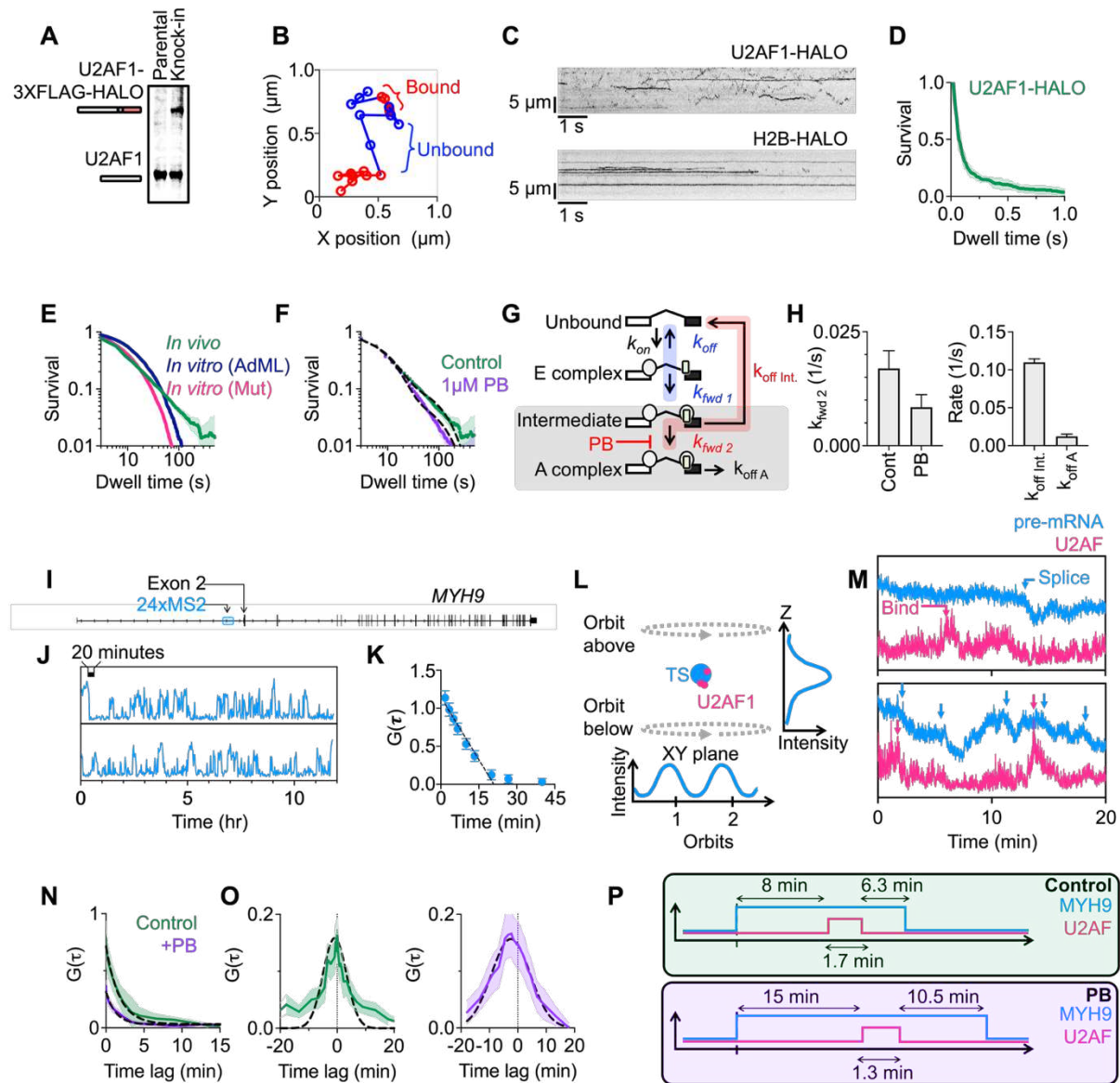


Figure 2: Single-molecule visualization of spliceosome assembly reveals stable U2AF binding through multiple assembly intermediates.

- Western blot of HBEC line containing one copy of U2AF1-3XFLAG-HALO, subsequently referred to as U2AF1-HALO.
- Fast tracking of U2AF1-HALO diffusion throughout the nucleus imaged with Highly Inclined and Laminated Optical Sheet (HILO) illumination at 100 Hz over 10 seconds. Bound and unbound states were inferred using a 2 state HMM model. We analyzed tracks that show complete binding events (i.e. bound event flanked by two unbound events) ($n=2$ replicates).
- Kymographs comparing U2AF1-HALO and H2B-HALO diffusion in fast tracking assays.
- Dwell times of complete binding events measured in fast tracking assays.
- Slow tracking U2AF1-HALO diffusion throughout the nucleus at 0.33 Hz over 20 minutes ($n=3$ replicates, shaded region represents 95% CI). For comparison, *in vitro* U2AF dwell time distributions from Fig. 1N are also included.

- F) Treating cells with pladienolide B (PB) results in a reduction in U2AF1-HALO dwell times (n=3 replicates). Shaded region represents 95% CI.
- G) Interpreting slow SMT data which detects U2AF dissociation from spliceosome complexes with a splice site selection model (intermediate and A complex states highlighted in gray box). Data is fit so that $k_{\text{off Int.}}$ and $k_{\text{off A}}$ are the same for both control and PB datasets. Blue highlighted rates represent K_f , the ratio of spliceosome forward progression with respect to U2AF dissociation at the E complex ($K_f = k_{\text{fwd 1}} / k_{\text{off}}$). Red highlighted rates represent K_s , the ratio of spliceosome forward progression with respect to U2AF dissociation at the intermediate state ($K_s = k_{\text{fwd 2}} / k_{\text{off Int.}}$).
- H) Model fitting reveals pladienolide B reduces $k_{\text{fwd 2}}$, the transition rate from intermediate to A complex by 2-fold. Additionally, our fits indicate U2AF remains bound in the intermediate state for 9.1 ± 0.4 seconds and the A complex for 82 ± 21 seconds. Fit parameters come from a global fit where $k_{\text{off int}}$ and $k_{\text{off A}}$ are shared.
- I) We used CRISPR to integrate a HALO tag on the C-terminus of U2AF1 in a cell line containing 24 MS2 stem loops integrated about 4kb upstream of the first 3'SS in MYH9. These cells also contain GFP-MS2 coat protein (MCP) so that upon transcription of stem loops, MCP binds and enables detection of RNA.
- J) Overnight confocal imaging of MYH9 TS reveals both transcription (transition to high fluorescent state) and splicing (transition back to low fluorescent state).
- K) Fluctuations were analyzed by correlation analysis and indicates a characteristic time of 19 ± 2 min for transcription and splicing. Error bars represent one standard deviation.
- L) 3D orbital tracking: a laser orbits above and below the TS. By monitoring the intensity in the XY plane as well as in Z, the particle is tracked in three dimensions. We tracked MYH9-GFP while monitoring U2AF diffusion in and out of the confocal volume in another channel.
- M) Example traces of MYH9 splicing and U2AF binding. Imaging was performed at 4 Hz over 20 minutes. Splicing manifests as a sudden decrease in GFP signal.
- N) Autocorrelation of time traces in U2AF channel corresponds to dwell times of $\tau_{\text{control}} = 1.7 \pm 0.1$ min, $\tau_{\text{PB}} = 1.3 \pm 0.1$ min. Error bars represent one standard deviation.
- O) U2AF→MYH9 cross-correlations are centered around 0 indicating overlap between MYH9 and U2AF events. Additionally, the width of the cross-correlation curves represent the time to splice ($\tau_{\text{splice control}} = 16 \pm 1$ min, $\tau_{\text{splice PB}} = 27 \pm 1$ min). Error bars represent one standard deviation.
- P) Summary of findings from SMT and OT experiments. U2AF binds for about 2 minutes to a pre-mRNA undergoing the splicing reaction. While U2AF dwell times are shorter after treatment with PB, overall time to splice increases.

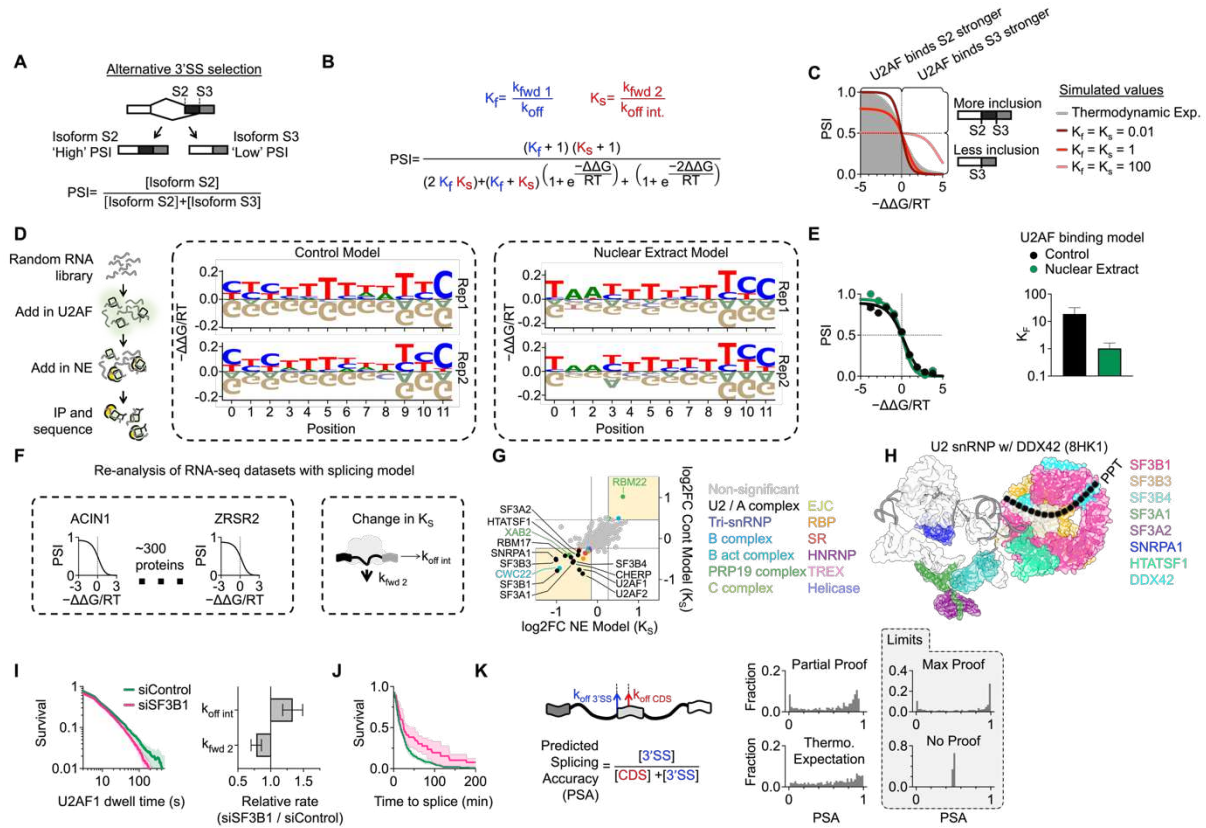


Figure 3: A kinetic model of splice site selection

- A) Alternative 3'SS selection. U2AF may bind to initiate spliceosome assembly at two competing 3'SSs, S2 and S3.
- B) Applying the kinetic model developed in figure 2 to the process of alternative 3'SS selection. This equation relates relative U2AF dissociation rate ($-\Delta\Delta G = RT \ln(k_{off\ S2}/k_{off\ S3})$) to percent spliced in (PSI). While the Probound binding model in Fig. 1 initially reports relative binding affinities instead of dissociation rates, we assume only dissociation rates (and not association rates) depend on sequence and, therefore, use k_{off} instead of K_D in our $-\Delta\Delta G$ calculation.
- C) The predicted relationship between $-\Delta\Delta G$ and PSI described by equation 2. The gray plot shows the expected relationship between PSI and $\Delta\Delta G$ in the scenario where splice site selection depends entirely on the relative equilibrium binding affinity between sites S2 and S3. The maroon curve represents the 'full proofreading' scenario where U2AF reads the 3'SS twice and both K_f and $K_s \ll 1$. Increasing K_f and K_s reduces the role of underlying U2AF binding affinity in splice site choice. For example, when K_f and $K_s = 100$ (salmon curve), PSI asymptotes at 0.5 at negative $-\Delta\Delta G$.
- D) A modified RBNS protocol to measure the influence of nuclear extract on U2AF sequence preference. U2AF binding models for both replicates are shown. We observe a branch point motif ([Y]UNAY) in the nuclear extract models.
- E) We performed RNA sequencing and quantified alternative 3'SS usage in HBECs. PSI values were binned by $-\Delta\Delta G$, the free energy of U2AF binding the S3 site relative to the S2 site for both binding models (control and NE). Circles represent median PSI of each bin. Error bars represent the average standard deviation of PSI for each alternative 3'SS ($n = 2$ replicates). The data was fit to equation 2. We fixed $K_s = 0.15$ based on slow SMT measurements, ($K_{f\ control\ model} = 20 \pm 10$, $K_{f\ NE\ model} = 1.0 \pm 0.5$).

- 722 F) Analyzing knockdown data from the Valcárcel lab using the alternative splicing model. The
723 relationship between PSI and binned $\Delta\Delta G$ for each protein knockdown was fit using
724 equation 2. K_f is fixed to 1 and we measure the change in K_s . We classified hits as
725 knockdowns that produced a change in K_s that exceeds any of the 7 siControl
726 measurements.
- 727 G) Plotting the change in K_s using either the control or NE binding model. Hits are colored by
728 complex.
- 729 H) A Cryo-EM structure of the U2 snRNP from the Shi Lab (PDB 8HK1) (35) highlighting the
730 proteins identified in knockdown screen. We estimated the location of the py-tract ("PPT").
731 While not present in screen, we also highlight the helicase DDX42.
- 732 I) U2AF dwell time distribution after SF3B1 knockdown measured by SMT, where we
733 measure a 22 ± 8 % reduction in k_{fwd2} as well as a $33 \pm 15\%$ increase in $k_{off\ int}$. Data is
734 comprised of 5 separate knockdowns and error bars represent SEM.
- 735 J) Measuring global changes in splicing kinetics using the polyclonal gene trap cell line where
736 728 introns are tagged with MS2 stem loops: $\tau_{sicontrol} = 19 \pm 3$ min, $\tau_{siSF3B1} = 43 \pm 19$ min.
737 Data is comprised of 3 separate knockdowns and error bars represent SEM.
- 738
- 739 K) Determining predicted splicing accuracy (PSA) at 3'SSs relative to sites 10-50 nt
740 downstream. The histograms compare PSA distribution for an optimal non-proofreading
741 'bind-once' model "thermo expectation" ($K_f = 0$) and the partial proofreading model ($K_f =$
742 1 and $K_s = 0.15$). The partial proofreading model predicts higher splicing accuracy. For
743 reference we include the performance limits of the full proofreading model, "Max Proof"
744 ($K_f = 0$ and $K_s = 0$) and "No Proof" ($K_f = 1e6$ and $K_s = 1e6$). U2AF binding locations come
745 from PAR-CLIP measurements and $\Delta\Delta G$ was determined using the NE Probound model.
746
747

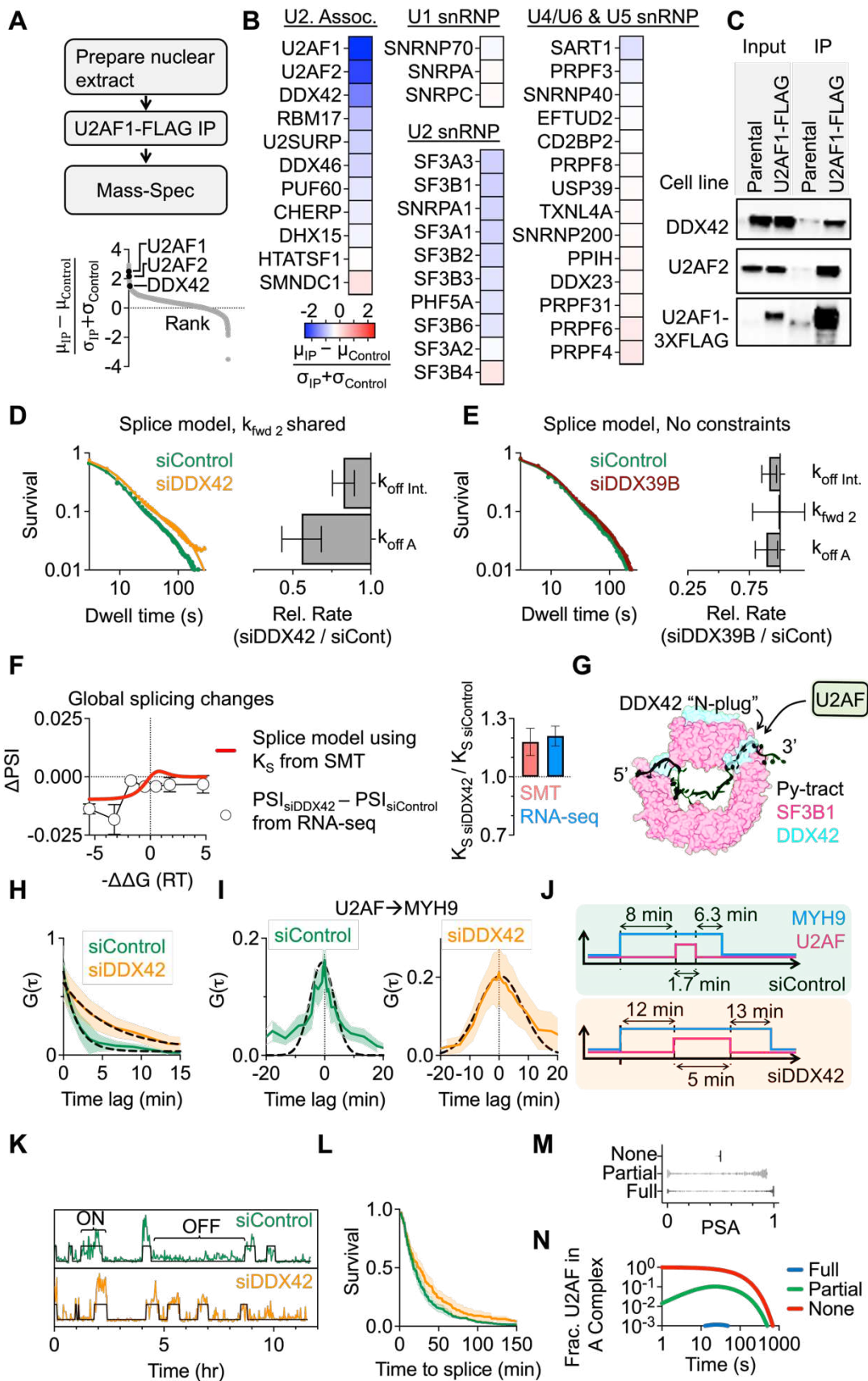


Figure 4. DDX42 accelerates U2AF dissociation from intermediate state and A complex

- A) Top: Workflow for U2AF1-FLAG IP-MS from HBEC nuclear extract. Bottom: Ranking identified proteins by signal-to-noise (where control is parental HBEC cell line). Other than U2AF1 and U2AF2, DDX42 is the most enriched splicing factor (n = 3 replicates).
- B) Plotting enrichment of proteins associated with U1, U2, and the tri-snRNP from our U2AF1 IP-MS assay. The protein components of each snRNP were determined using the spliceosome database⁵³.
- C) Western blot after U2AF1-FLAG IP confirms U2AF1-DDX42 interaction.
- D) Comparing U2AF1 dwell time distributions in the slow SMT assay after electroporation with DDX42 or Control siRNA. Fitting to equation 1 reveals a reduction in $k_{off\ Int.}$ and $k_{off\ A.}$ Error bars in bar chart represent standard error of the mean from 3 replicates.
- E) Measuring U2AF dwell times after knockdown of DDX39B (n = 1 replicate). Error bars represent standard deviation of 1000 bootstrapping iterations.
- F) Predicting the change in PSI ($PSI_{siDDX42} - PSI_{siControl}$) after DDX42 knockdown based on the change in K_s measured in slow SMT. Red dashed lines represent predicted PSI, white circles represent mean ΔPSI for binned $-\Delta\Delta G$ values from RNA sequencing. We include all identified alternative 3'SSs in this analysis. Bar plots compare change in K_s from SMT and RNA-seq measurements.
- G) Cryo-EM structure of SF3B1 in complex with DDX42 (PDB: 8HK1)⁴⁴. The pyrimidine tract RNA from mature Bact spliceosome (PDB: 5Z56)⁵⁴.
- H) Autocorrelation of the U2AF time traces from the orbital tracking measurements after DDX42 knockdown show that, on a pre-mRNA undergoing the splicing reaction, U2AF dwell times are longer ($\tau_{bound\ U2AF\ siDDX42} = 5.8 \pm 0.2\ min$, $\tau_{bound\ U2AF\ siCont} = 1.7 \pm 0.1\ min$). Error bars represent one standard deviation.
- I) U2AF→MYH9 cross-correlation broadens after DDX42 knockdown, indicating slower overall splicing ($\tau_{splice\ U2AF\ siDDX42} = 30 \pm 1\ min$, $\tau_{splice\ U2AF\ siCont} = 16 \pm 1\ min$). Error bars represent one standard deviation.
- J) Interpretation of SMT and orbital tracking data: DDX42 knockdown prolongs both U2AF binding and MYH9 splicing.
- K) Example time traces of gene trap cell line showing transcription and splicing events.
- L) Measuring changes splicing kinetics globally with the gene trap cell line $\tau_{1/2\ siControl} = 18\ min$, $\tau_{1/2\ siDDX42} = 26\ min$. Shaded regions represent one SEM (n=4 replicates for siControl, n=2 replicates for siDDX42).
- M) Comparing predicted splicing accuracy (as in figure 3L) for 3'SSs in competition with binding sites in nearby downstream regions for three regimes of the proofreading model, full ($K_f = 0.15e-3$, $K_s = 0.15e-3$), partial ($K_f = 1$, $K_s = 0.15$), and none ($K_f = 1.5e3$, $K_s = 1.5e3$).
- N) Comparing the fraction of bound U2AF molecules in the A complex for the three scenarios introduced in N. While figure N establishes that full proofreading is the most accurate, it is also very slow, at best, only ~0.1% of bound U2AF molecules are in the A complex at a specific time.

MATERIALS AND METHODS

Expression and purification of full length U2AF heterodimer

DNA fragments coding for 3X-FLAG-U2AF1 and 6X-HIS-U2AF2 were commercially synthesized (IDT) and cloned downstream of the p10 and polyhedron promoters respectively in the pFASTBAC baculovirus expression vector. To prepare the 3X-FLAG-U2AF1-HALO plasmid, we cloned in the HALO tag using In-Fusion cloning kit (Takara). Tni-FNL cells were transduced with baculovirus packaged with the expression vectors in the NIH protein expression laboratory. Cells were stored as pellets at -80°C until ready to purify. Cell pellets were thawed on ice and resuspended in Buffer A (50 mM tris pH 7.5, 500 mM NaCl, 5 mM imidazole, 10% glycerol, protease inhibitor tablet (Pierce Protease Inhibitor Tablets, EDTA-free)). After thawing, cells were lysed by sonication with 15 cycles (15 s on/10 s off) using the Bioruptor Plus (Diagenode) at 4°C. The first 5 cycles were set to 'high' power and the last 10 cycles were set to 'low' power. After lysis, Cells were clarified by centrifugation at 23000xg for 20 min. Supernatant was slowly loaded onto a 5 mL HisTrap HP IMAC column (Cytiva) and then installed onto a BioRad NGC FPLC. We performed 30 mL wash with buffer A to remove unbound protein followed by a 60 mL wash with Buffer B (50 mM tris pH 7.5, 500 mM NaCl, 60 mM imidazole, 10% glycerol) and then 15 mL buffer C (50 mM Tris pH 7.5, 500 mM NaCl, 340 mM imidazole, 10% glycerol). To ensure each U2AF2 molecule was in complex with U2AF1, we finished the purification by U2AF-IP with anti-FLAG M2 magnetic beads (Sigma Aldrich). Before mixing with U2AF, beads were washed 4x in Buffer D (10 mM tris pH 7.5, 500 mM NaCl, 0.1% tween 20, 0.5 mM DTT, protease inhibitor tablet. U2AF was eluted with 25 mM Tris pH 7.5, 100 mM NaCl, 200 ug/ml 3xflag, 1mM DTT. Protein concentration and purity was assessed by SDS PAGE. Protein was stored at -80C after adding glycerol to 10%.

For fluorescent labeling of U2AF1-HALO in complex with U2AF2, we mixed equal molar ratio of JFX650 HALO Tag ligand (Promega) with U2AF before FLAG tag purification. We let the reaction incubate for 30 minutes at 4°C before proceeding with purification.

Expression and purification of U2AF heterodimer

U2AF1 (1-188) and U2AF2 (85-342) were used for RNaseI footprinting assays. These expression vectors were a gift from the Kielkopf lab and purified as previously described³¹.

RNaseI footprinting assay

RNaseI footprinting assays were performed as described previously⁵⁵.

Briefly, recombinant U2AF1(1-188) and U2AF2(85-342) was incubated with either AdML-WT or AdML-U4A4 RNAs with a Cy5 label on their 3'-end (**Extended Data Fig. 1B**) for 10 minutes at 4°C to allow complex formation. RNase I was then added at either 0.0003 or 0.003 Units with carrier tRNA (400 ng/uL final) for 5 minutes at 4°C to achieve approximately 1 cleavage event per RNA molecule. Reactions were quenched by addition of denaturing urea loading dye, and samples were immediately analyzed by denaturing 7M urea 20% polyacrylamide gel electrophoresis.

An RNA size ladder was generated by controlled alkaline hydrolysis. Reactions contained RNA (1 µM), yeast tRNA (400 ng/µL, as carrier), and 4 mM EDTA (pH 8) in a total volume of 20 µL. RNA cleavage was initiated by addition of NaOH and samples were incubated at 4 °C for 15 or 25 min to generate partial hydrolysis products. Reactions were quenched by addition of 4 µL 1 M Tris-HCl (pH 8) and then kept on dry ice until analysis by urea-polyacrylamide gel electrophoresis.

Electrophoresis Mobility Shift Assays (EMSAs)

5 nM RNA was mixed with 0-120 nM U2AF in binding buffer (25 mM Tris pH 7.0, 150 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.05% Tween 20, 0.02 ug/ul tRNA) and incubated at 4C for 15 minutes. Binding was resolved in a native PAGE gel (1XTBE, 8% polyacrylamide (29:1 National Diagnostics), 0.1% APS, 0.1% TEMED, 10% Glycerol) run at 75 volts for 2 hours at 4C. Gel was pre-run for 1 hour before loading.

We measured binding to the following sequences (changes to mutant sequence in bold). The red region is a DNA sequence used for tethering RNAs to microscope slide in single-molecule PIFE assays.

AdML	5' - TGGCGACGGCAGCGAGGCC UGUCCCUUUUUUUUCCACAGCU-Cy3
Mutant	5' - TGGCGACGGCAGCGAGGCC UGUCCCU CUAUUC UCCACAGCU-Cy3

Preparation of Nuclear Extract

We prepared nuclear extract from 25L of suspension HeLa S3 cells at a concentration of 5×10^8 cells per liter from the Cell Culture Company as previously described⁵⁶.

RNA Bind-N-Seq (RBNS)

Our protocol for RNA Bind-N-Seq closely follows the protocol developed by the Burge Lab⁵⁷. The random RNA library was prepared by annealing the T7 oligo to the RBNS template, which is a single-stranded DNA oligo from IDT (sequences are listed in table below). We then performed *in vitro* transcription for 4 hours using the NEB HiScribe T7 RNA synthesis kit. After RNA synthesis, we degraded the DNA template by DNaseI digestion for 15 minutes. RNA was purified using the NEB Monarch RNA cleanup kit and stored at -80°C. We confirmed that RNA was the correct length by tapestation analysis (Agilent).

U2AF (0 nM, 20 nM, 80 nM, or 320 nM) was incubated with 500 nM of random RNA library in U2AF binding buffer without tRNA (25 mM Tris pH 7.0, 150 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.05% Tween 20) at 4°C for 30 minutes with constant mixing on a tube rotator. We performed each binding reaction in duplicate. Final volume was 250 uL per reaction. RNA in complex with U2AF was isolated by U2AF1-FLAG IP using anti-FLAG M2 magnetic beads (Sigma Aldrich). Beads were washed 4x in RBNS wash buffer (25 mM Tris pH 7.5, 150 mM KCl, 60 ug/mL BSA, 0.01% Tween 20) and then resuspended in 200 uL of binding buffer. We added 20 uL of washed beads to each RBNS reaction and incubated for 30 minutes at 4°C. Bound RNA was pelleted with magnetic beads and incubated in 1000 uL RBNS wash buffer for 15 minutes at 4°C. Next, the RNA was pelleted and resuspended in elute buffer (25 mM HEPES pH 7.3, 100 mM NaCl, 200 ug/ml 3XFLAG (sigma), 1 mM DTT). U2AF was removed from RNA by phenol-chloroform extraction and ethanol precipitation. Adding 4 uL glycogen to each ethanol precipitation significantly improved the ability to visualize such a small pellet. After ethanol precipitation, the quality of RNA was assessed by tape station.

For experiments with nuclear extract (and the associated controls), we made the following changes to the protocol. We prepared a reaction containing U2AF (0 nM, 5 nM, 20 nM, 80 nM), U2AF ATP binding buffer (25 mM Tris pH 7.0, 150 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.05% Tween 20, 0.5 mM ATP), and 10% nuclear extract. In parallel, we performed the same assay without nuclear extract. At this concentration of nuclear extract, we observe about 50% of RNAs are bound (**Figure S5F**). Each binding reaction was incubated for 30 minutes at 4°C. However, to preserve complex formation, we performed a 'light' wash for this assay. To remove unbound RNAs, we added 750 uL wash buffer directly to each reaction. After a two minute incubation, we removed supernatant.

Even in these minimal wash conditions, we still observe a strong correlation between predicted and measured binding affinity (**figure S2D**). Therefore, while we use the U2AF binding model from the stringent wash data for figure 1, we use the nuclear extract model exclusively in figures 3 and 4.

For all RBNS assays (stringent wash and light wash), we prepared sequencing libraries in the same way. Isolated RNA was converted to cDNA using the ImProm-II reverse transcription system (Promega) with the RT primer (see table). We also performed reverse transcription on the random RNA library not subjected to U2AF or IP (input sample). We performed PCR on each sample using NEBNext Multiplex Oligos for Illumina (NEB). We assessed PCR product by agarose gel and chose the number of amplification cycles to avoid side products (0 nM-15 cycles, 20 to 320 nM-11 cycles, input-9 cycles). Finally, DNA bands corresponding to about 160 bp were purified from 8% Native PAGE and incubated in 400 μ L gel extraction buffer (10 mM Tris, pH 7.0, 300 mM NaCl, 2 mM EDTA) for 1 hour at 65°C. DNA was further cleaned using a DNA Clean & Concentrator Kit (Zymo). The correct DNA length (160 bp) was confirmed by tape station.

DNA was pooled into a single tube (67 fMoles per reaction) and sent for deep sequencing at the NCI sequencing facility. Sequencing was performed on a NovaSeq (Illumina) in an S2 flow cell with 2x100 read length. We sequenced to a depth of between 149 million to 450 million reads per binding reaction which allowed us to determine enrichment scores for very long U2AF binding sites. For all samples, at least 90% of reads had a quality score above Q30.

Analysis, performed using the RBNS pipeline (https://github.com/cburgelab/RBNS_pipeline), was too computationally intensive for a single computer. Therefore, the pipeline was performed on Biowulf, a linux cluster maintained by the NIH High Performance Computing center.

T7 Oligo	TAATACGACTCACTATAGGG
RBNS Template	CAGACGTGTGCTCTTCCGATCT-N40- AGATCGGAAGAGCGTCGTGTAGGGCCCTATAGTGAGTCGTATTA
RT primer	CAGACGTGTGCTCTTCCGATC

Predicting relative binding affinities of *in vivo* U2AF binding sites

We randomly selected 5 million sequences from each sample of the RNA Bind-N-Seq assay using seqtk (<https://github.com/lh3/seqtk>) and applied ProBound Apps for binding mode modeling (<https://github.com/RubeGroup/ProBound>) and motif visualization (<https://github.com/BussemakerLab/PyProBound>). We generated U2AF binding models as 12mer energy logos. Binding models were generated by comparing sequence differences between each U2AF concentration and the input fraction.

Predicted binding affinities were assigned to peaks from a previous U2AF1 PAR-CLIP experiment⁴ using ProBoundTools (<https://github.com/BussemakerLab/ProBoundTools>). We determined the U2AF binding affinity of every 12mer in a single PAR-CLIP peak and assigned the highest affinity value as affinity of that binding site. We did not consider the small fraction of PAR-CLIP peaks that spanned less than 12 nt.

When analyzing alternative 3'SS data, we calculated the binding affinity for the 12mer sequence immediately preceding the -AG dinucleotides. To ensure that we're looking at competition between two distinct 3' splice sites and not a single pyrimidine tract followed by two splices, we only analyzed alternative 3' splice sites that were greater than 10 nt apart.

To search for a secondary binding model, we calculated the U2AF binding affinity for each sequence in the RBNS output using the primary 12mer proBound model, which represents a strong polypyrimidine tract, and only picked those with affinity values < 0.1 as input. Fitting models of length 5 returned 3 binding modes: a polypyrimidine tract, a G-rich pattern, and the TTAGG motif. Affinity distribution across concentrations indicates that the G-rich pattern is similarly enriched in control and U2AF samples and is likely to be an artifact.

Competitive Binding Assays

Testing U2AF binding directly to sites identified from PARCLIP would require copious amounts of U2AF. Therefore, we performed competitive binding assays where 50 nM U2AF was prebound to 10 nM Cy5-(U)₁₂ RNA and competed off with increasing concentrations of unlabeled competitor RNA. For each titration we measure the IC₅₀, the concentration of unlabeled RNA required to compete off 50% of the Cy5-RNA. The relative affinity of two binding sites is the ratio of their IC₅₀ values. Binding assays were resolved by EMSA (see EMSA section for details) using the RNA oligos in the table below. While probound predicts a very broad range of affinities, we were only able to validate U2AF binding to sequences with binding affinity greater than 370-fold weaker than the strongest motif. Binding reactions were performed over multiple days. To account for experimental variability, each competitive binding assay was performed in parallel to the '3B' sequence. Reported relative binding affinity is IC₅₀ / IC₅₀ 3B Seq.

Labeled Oligo	Cy5-UUUUUUUUUUUU
Seq 2A	UUUUUUUUUUGGG
Seq 2B	UUUUUUUUUCCCC
Seq 3B	UUUUUUUUUUUUUU
Seq 4B	UUUUUUUUUUUUUU
MYH9_A	UCCUUUACUUCU
MYH9_B	UUAACUCCCUAG
MYH9_D	CAUCAUCUCUCC
MYH9_H	UUUCAGACCGAG

Ensemble PIFE assay

U2AF binding to 5 nM Cy3-RNA was determined by protein induced fluorescence enhancement (PIFE), in which Cy3 fluorescence increases upon protein binding as previously described^{15,58}. Each binding reaction was 60 uL. Cy3 was excited at 510 nm. We measured Cy3 emission measured in 2 nm steps from 550-630 nm. Fluorescence enhancement was greater when the Cy3 was placed after the U2AF1 binding site as opposed to before the U2AF2 binding site. RNA was incubated for at least 5 min with varying concentrations of U2AF in 'U2AF binding buffer without tRNA' (25 mM Tris pH 7.0, 150 mM NaCl, 0.5 mM DTT, 10% glycerol, 0.05% Tween 20). Fluorescence spectra were obtained using a Horiba fluorometer and analyzed in MATLAB to (1) remove background autofluorescence from buffer and protein and (2) to determine the relative change in Cy3 fluorescence compared to 0 nM. Background autofluorescence was determined by performing the same fluorescence scan but without RNA (protein + buffer).

Slide Preparation

A quartz microscope slide (Alfa Aesar) was patterned with 16 small holes to form 8 flow cells (each flow cell has an inlet and an outlet). #1.5 coverslips (Corning) and microscope slides were soaked in toluene, ethanol, and then plasma cleaned. Slides were then incubated in 100 uM mPEG-Si and biotin-PEG-Si (Laysan Bio) overnight in anhydrous toluene. Functionalized quartz slides and coverslips were assembled into microscope flow cells where channels are defined with

parafilm. Before each experiment, the flow cell was treated sequentially with 1 mg/ml BSA, 40 ug/ml neutravidin, and biotin-labeled RNA in 'U2AF binding buffer -tRNA' + 0.5% Tween 20. For colocalization assays, slide preparation was performed as previously described⁵⁹.

smPIFE data acquisition

To attach U2AF binding sites to the biotinylated surface, we annealed either the 'AdML' or 'Mutant' RNA to a DNA handle that was biotinylated on its 3' end. All oligos were purchased through IDT.

AdML	5' - TGGCGACGGCAGCGAGGC CUGUCCCUUUUUUUUCCACAGCU-Cy3
Mutant	5' - TGGCGACGGCAGCGAGGC CUGUCCCU CUAUUC UCCACAGCU-Cy3
smHandle	5' - GCCTCGCTGCCGTCGCCA -intCy5-biotin

Annealed RNAs were added to the flow cell. After a 5 minute incubation, unbound molecules were removed by a wash with 'U2AF binding buffer -tRNA'. We then added imaging buffer containing the desired U2AF concentration into the flow cell. Imaging was performed for 500 frames at 1 Hz. The imaging buffer contained 10 mM Tris-HCl (pH 7), 150 mM NaCl, 10% glycerol, 0.5% v/v Tween-20, 0.1 mg/ml BSA, 2 mM Trolox, 0.0115% v/v COT (Cyclooctatetraen, Sigma 138924), 0.012% v/v NBA (3-Nitrobenzyl alcohol, Sigma 146056), 450 ug/ml glucose oxidase (Sigma G2133) and 22 ug/ml catalase (Sigma C3155).

smPIFE data analysis

In each image stack, the imageJ "find maxima" function was applied to determine the coordinates of each RNA molecule. Time traces were generated by measuring the average intensity of a 3x3 pixel area centered on each molecule throughout the entire stack. We typically identify about 1000 RNA molecules per field of view and attempted to analyze as many RNAs as possible in each video. We ended up analyzing $22 \pm 7\%$ and $28 \pm 6\%$ of all identified molecules for AdML and Mutant RNAs, respectively. While we attempted to include as many RNAs as possible in the analysis, we discarded traces with multiple photobleaching steps (RNA aggregates), traces that photobleached too quickly (<50 seconds), and uninterpretable traces. Below is a table of the number of molecules identified in each measurement and the number of molecules sent for further analysis:

[U2AF] (nM)	RNA	Replicate	# Identified	# Analyzed	% Analyzed
15	AdML	1	565	93	16
		2	822	122	15
		3	1163	251	22
30		1	513	72	14
		2	462	126	27
		3	1055	378	36
60		1	720	168	23
		2	465	95	20
		3	1103	288	26

[U2AF] (nM)	RNA	Replicate	# Identified	# Analyzed	% Analyzed
15	Mutant	1	1661	310	19
		2	420	137	33

30		1	903	329	36
		2	1232	353	29
60		1	1216	333	27
		2	721	176	24

Selected traces were cropped and analyzed using ebFRET⁶⁰ which infers kinetic parameters based on global HMM fitting. We fit the data to a two state HMM model where the high state represents U2AF in complex with RNA and the low state represents unbound RNA. It becomes difficult to observe binding fluctuations at high concentrations of U2AF where the fluctuations begin to occur very quickly. Nevertheless, we observe a concentration dependent on-rate which indicates the observed fluctuations represent U2AF binding events.

To test the specificity of the PIFE signal, we performed a single-molecule competition assay where U2AF binding was measured in the presence of a potent competitor (UUUUUUUUUCCC) and an off-target RNA (CCCCAGCUCAG). The buffer conditions were identical to the other smPIFE assays except this time we also included tRNA (0.02 ug/ul) as in gel shift assays. We performed a correlation analysis on all traces with fano-factor < 30 using fcs_app in IDL (<https://github.com/CBIIT/Larson-Lab-CCR-NCI.git>) for the following conditions:

- Cy3-AdmL RNA + 0 nM U2AF (3 replicates)
- Cy3-AdmL RNA + 20 nM U2AF (2 replicates)
- Cy3-AdmL RNA + 20 nM U2AF + 500 nM 2B competitor (unlabeled) (3 replicates)
- Cy3-AdmL RNA + 20 nM U2AF + 500 nM 'off-target' competitor (unlabeled) (2 replicates)

The measured autocorrelation for Cy3-AdmL RNA + 0 nM U2AF represents the timescales of Cy3-fluctuations that do not correspond to U2AF binding (i.e. photobleaching and transient blinking). To remove the contribution of these fluctuations to the assays containing 20 nM U2AF, we subtracted the 0 nM autocorrelation curve from all other conditions. These are the curves shown in **Extended Data Fig. 3G**. Each corrected autocorrelation curve was fit to a single exponential decay where the time constant represents the sum of the dissociation rate and association rate ($k_c = k_{off} + k_{on, apparent}$)⁶¹. Using $k_{off} = 0.057 \pm 0.003 \text{ s}^{-1}$ from **Fig. 3E**, we determined $k_{on, apparent}$ and fraction bound for each condition.

Colocalization Analysis

In each image stack, the imageJ “find maxima” function was applied to determine the coordinates of each RNA molecule in the FAM and Cy3 channel. In this analysis, we only considered locations where both a Cy3 and FAM spot were identified. Once the locations of RNA molecules were identified, we performed a few unbiased quality control steps on the JFX650 images. In particular, full-length U2AF is prone to aggregation so we masked out large static blobs that were apparent in maximum-intensity projected images. We also corrected the JFX650 movies for small amounts of drift using the Fast4DReg imageJ plugin. After these processing steps, we curated traces similarly as with the smPIFE assay. In figure 3G, we describe the fraction of RNAs outside blobs where we observe U2AF binding.

We measured U2AF-JFX650 binding to two different constructs. The first was an RNA nearly identical to the AdmL RNA used in the smPIFE assay. We also performed a negative control assay where we monitored U2AF binding to the ‘Handle block’ oligo annealed to smHandle.

smHandle	5' - GCCTCGCTGCCGTCGCCA -intFAM-biotin
AdML_U2AF2/1_V4	5' - TGGCGACGGCAGCGAGGC CUGUCCCUAA <u>CAUUUUUUU</u> UCCACAGCU-Cy3
Handle_block	5' -TGGCGACGGCAGCGAGGC

Preparing Cell Lines

We tagged endogenous U2AF2, U2AF1, and H2B with HALO-tags for single-molecule tracking experiments in Human Bronchial Epithelial Cells (HBECs). We used Keratinocyte SFM (Gibco 17005042) for all cell culture. For the U2AF2 cell line, we inserted a HALO-tag on the C-terminus of U2AF2 followed by P2A and a blasticidin marker for selection. For the U2AF1 cell line, we inserted a HALO-tag on the C-terminus of U2AF1 followed by 3XFLAG, T2A, and a blasticidin marker for selection. For orbital tracking assays, we also HALO-tagged the C-terminus of U2AF1 in the MYH9-MS2 cell line created earlier⁴. Finally, for the H2B cell line, we inserted a HALO-tag on the C-terminus immediately after a TEV site.

Tagging was accomplished with CRISPR-CAS9. Guide RNAs were cloned into the PX458 plasmid to target the C-terminus. Subsequent HALO-tag integration was accomplished by homology directed repair. These are the genomic coordinates of the homology arms:

Name	Assembly	chr	Left homology arm		Right homology arm	
			Start	End	Start	End
U2AF2-HALO	GRCh38.p14	19	55,673,561	55,674,065	55,674,069	55,674,219
U2AF1-HALO	GRCh38.p14	21	43,093,105	43,093,960	43,092,220	43,093,098
H2BC8-HALO	GRCh38.p14	6	26,216,266	26,217,066	26,215,502	26,216,265

We were not able to generate homozygous knock in cell lines, likely because (a) integration of ~30kDa HALO tag is inefficient and (b) our selection approach enriches for both homozygous and heterozygous integration. We ruled out the possibility that a homozygous integration renders the cells inviable by performing HALO-tag knock in an HBEC cell line where one copy of U2AF1 is frameshifted⁶². Our measurements of U2AF dwell times in these cells agree with the original cell line expressing both tagged and untagged U2AF (**Figure S4H-I**).

Single-molecule tracking

Single-molecule tracking of HALO-tagged proteins was performed with Highly Inclined Laminated Optical sheet (HILO) microscopy on a custom-built microscope. The microscope is built on a Rapid Automated Modular Microscope system (RAMM, Applied Scientific Instrumentation) with synchronized lasers and cameras using an Arduino board. Images are acquired using micromanager (Open imaging). The microscope contains an Olympus 100x 1.49 NA oil immersion TIRF objective which corresponds to a pixel size of approximately 150 nm. The automated stage (Applied Scientific Instrumentation) is fitted with an incubator to maintain CO₂ and temperature (Oko lab and Tokai Hit). The emission path contains a quad-band dichroic (ZT405/488/561/647 Chroma technology) which is further separated by two long pass filters (T588lpxr, T660lpxr) and emission filters (525/50, 609/58, 736/128, Semrock) and directed onto 3 separate EMCCD cameras (Photometrix).

Tracking was performed in Human Bronchial Epithelial Cells (HBEC) labeled with between 0.5 pM – 1 pM JF646 dye (Promega). At this concentration of dye, each frame contained about 5-10 labeled U2AF or H2B molecules. JF646 was stored at 500 nM in DMSO at -20 C and serially diluted to 0.5 pM in Keratinocyte SFM (Gibco 17005042) (KSFM). After 15 minute incubation at

37C, we removed free dye by washing 3x in KSFM, followed by a 10 minute incubation at 37C, followed by another 3x KSFM wash before imaging.

These experiments are analyzed using the MatlabTrack pipeline (https://github.com/davidaalejogarcia/PL_HagerLab.git)⁶³. Slow tracking experiments were conducted at a rate of 1 frame every 3 seconds. During each frame, cells were illuminated for 100 ms with a 647 nm laser (Coherent Obis 1196627) set to power 1 mW (with a 10% neutral density filter). Fast tracking measurements were performed at a rate of 1 frame every 10 ms with constant 15 mW laser illumination. Particle detection was accomplished using SLIMfast⁶⁴. From these localizations, we created tracks using the nearest neighbor approach with the following constraints (maximum jump → 4, shortest track → 2, gaps to close → 2).

The diffusive parameters of a 'bound' molecule were determined by tracking H2B-HALO as previously described⁶³. Additionally, because the vast majority of H2B is stably integrated into chromatin, disappearance of an H2B-HALO 'spot' corresponds to photobleaching. Therefore, the U2AF dwell time distribution from slow tracking assays was corrected for photobleaching using the approach outlined in this paper⁶³.

We analyzed fast tracking data in vbSPT⁶⁵ where tracks were fit to two states, 'bound' and 'unbound' using the default priors and the following settings:

- time step = 0.01
- dim = 2
- trjLin = 2
- bootstrapNum = 10
- fullBootstrap = 0
- init_D = [0.001, 16]
- init_tD = [2, 20]*timestep.

We observe three types of tracks: U2AF is unbound for the entire trajectory, U2AF is bound for the entire trajectory, and U2AF transitions between bound and unbound state during the trajectory. In this very fast imaging condition with high laser power, we suspected that many long trajectories ended because of photobleaching and not because diffusion out of illumination volume. Therefore, when determining dwell times at these short time scales, we only analyzed the minority of tracks that show a bound event separated by two unbound events. In this way, we can be sure that we are observing dissociation and not photobleaching.

Single-molecule FISH

Nascent RNA was visualized by single-molecule FISH (smFISH) as previously described^{4,66}. Human Bronchial Epithelial Cells were plated onto #1.5 coverslips, washed 3X with HBSS and fixed by incubating for 10 minutes in 4% EM Grade Paraformaldehyde (Electron Microscopy Sciences) in PBS. After 1X PBS wash, cells were permeabilized by incubating in 70% ethanol at 4C for 1-2 days. Before adding FISH probes, coverslips were incubated 10 minutes in PBS twice and then 5 minutes in wash buffer (10% formamide, 2X SSC). FISH probes were diluted to 50 nM in hybridization buffer (2× SSC (v/v), 10% dextran sulfate (w/v) and 10% deionized formamide (v/v) and added dropwise to parafilm. Coverslips were placed cell-side down in hybridization buffer and incubated at 37C for at least 4 hrs. After incubation, coverslips were incubated at 37C twice for 30 minutes in wash buffer. Coverslips were then washed with 2X SSC followed by PBS. After air drying, coverslips were placed cell side down into mounting media spotted onto microscope slides (Prolong Gold with DAPI, Thermo) and stored at -20°C. We used the same probe sets as in Wan et al 2021.

Single-molecule FISH experiments are performed on a custom microscope setup consisting of a RAMM (ASI) chassis and a Zeiss 40x, 1.4 N.A. objective. The emission path is filtered using both a quad bandpass filter and an automated emission filter wheel for imaging DAPI, GFP, Cy3 and Cy5 (VCGR-SPX-P01-PC, Chroma). We've installed an automatic stage (ASI) for imaging multiple fields of view using an ORCA-Flash4 V2 CMOS camera (Hamamatsu). The microscope is controlled using micromanager (Open imaging). Image analysis is performed in python (https://github.com/Lenstralab/smFISH_analysis.git).

Confocal Imaging of Transcription and Splicing Kinetics

Splicing kinetics were measured on an automated high-throughput dual spinning disk confocal microscope (Yokogawa Cell Voyager 7000S) as previously described⁴. Briefly, imaging was performed on 96 well glass bottom plates (Cellvis-P96-1.5H-N) seeded with 40,000 cells about 16 hours before imaging. For the fluorescence imaging, we used a 488 nm laser to visualize GFP-MS2 coat proteins at a rate of one z-stack (16 images, 0.5 μ m increments) every 100 seconds for 12 hours. To save space, we performed on-the-fly maximum intensity projection which is what we used for analysis. Cells were segmented and registered using HiTIPS⁶⁷. Transcription site tracking was performed in IDL using localize (<https://github.com/CBIIT/Larson-Lab-CCR-NCI>).

3D Orbital Tracking

Orbital tracking was performed on an inverted Zeiss Observer Z.1 microscope. The microscope contains a Zeiss 63X 1.2 NA water immersion objective and an automated stage (ISS) that is temperature and CO₂ controlled (Oko Lab and Tokai Hit). U2AF1-HALO (labeled with JF646 (Promega)) was detected with a long pass filter (T647lpxr) followed by a bandpass filter (ET706/95m) (Chroma). The GFP channel is filtered using a ET525/50m filter (Chroma). Data is collected with avalanche photodiodes (Excelitas). Tracking was performed using SIM FCS software (ISS).

To simultaneously image RNA splicing and U2AF binding, we integrated a HALO tag on the C-terminus of U2AF1 in a cell line containing 24XMS2 stem loops 4 kb upstream of the 3'SS of MYH9 intron 1. While it has long been known that decent matches to a 3'SS occur about once every 500 bp throughout a pre-mRNA⁶⁸, our analysis of U2AF binding site locations indicate that MYH9 intron 1 appears to be an ideal candidate due to lack of nearby competitor sites. We observe no intronic binding sites in the 4000 nt region that spans from the MS2 integration to 3'SS. Furthermore, there is about a 7000 nt window centered on the 3'SS of intron 1 completely devoid of U2AF binding sites. Finally, the MYH9 intron 1 3'SS contains a strong pyrimidine tract (**Extended Data Fig. 5G**). We believe these features of MYH9 makes it an ideal candidate for this measurement.

We selected cells with visible MS2 spots for imaging. These represented cells actively transcribing MYH9 pre-mRNA. Based on single-molecule FISH measurements, the majority of MYH9 TS produce 1-2 pre-mRNAs per burst (**S4L**). Therefore, when tracking begins, we typically observe only 1-2 downward steps in MYH9 fluorescence (**Figure S4O**) which corresponds to splicing events. Furthermore, very few traces show increases in fluorescence which corresponds to transcription. Since we begin tracking only after transcription of the stem-loops, we begin the measurement with no understanding of where we are in the lifetime of the intron. For this reason, we are unable to measure the total time to splice by autocorrelation as previously^{21,22}. However, cross-correlation, which relates the timing of U2AF binding during the lifetime of the pre-mRNA, shows a peak near 0 minutes which means MYH9 splicing and U2AF binding overlap. Furthermore, perturbation (whether DDX42 knockdown or pladienolide B

treatment) broadens the cross-correlation curve indicating slowed splicing relative to the unperturbed control.

Cells were labeled with 5-10 pM JF646 for 15 minutes at 37°C. While this is 10 to 20-fold higher dye concentration than the SMT measurements, many U2AF molecules will still go unlabeled. However, increasing the dye concentration further adds to the background fluorescence signal. Therefore, we determined that this dye concentration was best suited to observe single U2AF binding events while minimizing background signal. Not all traces show clear binding fluctuations in the U2AF channel (**figure S4O**), likely due to under labeling.

While these measurements indicate that U2AF association to MYH9 occurs about 8 minutes after transcription of the stem loops, it is important to note that because this assay requires under labeling of U2AF the actual initial time for U2AF to bind is likely much faster. However, we reason that since we use the same dye concentration for all assays, this result reflects the relative change in spliceosome assembly after perturbation.

RNA sequencing

RNA extraction was performed on ~1 million Human Bronchial Eplithelial Cells (HBEC) using the Qiagen RNeasy mini kit (74106). Genomic DNA and degraded using the Turbo DNase kit (Invitrogen, AM1907). We verified complete depletion of genomic DNA by qPCR. All samples had an RNA integrity value (RIN) above 9.9. We sent 600 ng of each sample for sequencing library preparation by the CCR genomics core which performed rRNA depletion using NEBNext rRNA Depletion Kit v2 (E7400), and sequencing library preparation using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760). Sequencing was performed using Illumina NextSeq 2000 P3 (200 cycle) reagents at a depth of 163 to 193 million reads per replicate. Raw sequencing files were aligned using Bowtie2⁶⁹ which also removed reads corresponding to rRNA and ERCC spike-in sequences (ThermoFisher 4456740). After transcript quantification with Salmon⁷⁰, we evaluated differential splicing with SUPPA2⁷¹.

We obtained raw sequencing files of splicing regulator knockdown by Rogalska et al. from EMBL²⁸. Read alignment was performed using STAR (<https://github.com/alexdobin/STAR>) and the human GRCh 38 reference genome. Transcript expression quantification was performed with salmon (<https://github.com/COMBINE-lab/salmon>). We analyzed alternative splicing using SUPPA (<https://github.com/comprna/SUPPA>) and calculated the relative U2AF binding affinity by applying the binding mode to alternative 3'SS sequences.

The analysis in **Fig. 3G** was performed by plotting $-\Delta\Delta G$ vs PSI for each of the ~300 knockdowns using the control and nuclear extract binding models from **Fig. 3D**. We then fit the relationship using the steady state kinetic model of alternative 3'SS selection (equation 2, main text). The nuclear extract model was difficult to fit because the binned PSI values are extremely close to 1 when $-\Delta\Delta G < 0$. In this regime, any reduction in K_s produces essentially no change in PSI because PSI cannot go above 1 (**Extended Data Fig. 5H**). Therefore, we report the 95% one-sided confidence intervals for the upper limit of K_s for each protein knockdown as the lower limits all equal 0.

Kinetic model of spliceosome assembly

Dwell time distributions from single-molecule tracking assays were fit to the kinetic model illustrated in figure 2. However, based off our *in vitro* measurements which suggest rapid U2AF dissociation, we assume that the first two states were not detectable in the slow SMT assay where data was acquired at a rate of one frame every three seconds. Therefore, the kinetic model was

consisted of 2 states (**Extended Data Fig. 4K**), the intermediate and the A complex, which we modeled with the following differential equations:

$$\frac{\partial[Int]}{\partial t} = -k_{off\ int}[Int] - k_{fwd\ 2}[Int]$$

$$\frac{\partial[A]}{\partial t} = k_{fwd\ 2}[Int] - k_{off\ A}[A]$$

This model contains three rates $k_{off\ int}$, $k_{off\ A}$, $k_{fwd\ 2}$, and we assume that all U2AF molecules start in the intermediate state when detected in the SMT assay. Therefore, we solved this system of differential equations with the initial condition of $Int(0) == 1$ and $A(0) == 0$ and defined survival as the concentration of U2AF molecules in either the Intermediate state or the A complex (Survival = $[Int] + [A]$):

$$U2AF\ survival = e^{-t(k_{fwd\ 2} + k_{off\ int})} - \left(\frac{k_{fwd\ 2}}{k_{fwd\ 2} + k_{off\ int} - k_{off\ A}} \right) (e^{-t(k_{fwd\ 2} + k_{off\ int})} - e^{-t(k_{off\ A})})$$

Pladienolide B blocks A complex formation by preventing SF3B1 recognition of the branchpoint. Therefore, when comparing the U2AF dwell time distribution +/- Pladienolide B, we assumed that only $k_{fwd\ 2}$ is affected and that $k_{off\ int}$ and $k_{off\ A}$ are unchanged. This allowed us to simultaneously fit both dwell time distributions (**Fig. 2F**) using shared values for $k_{off\ int}$ and $k_{off\ A}$ (reported in **Fig. 2H**). When comparing U2AF dwell time distributions in the siControl and siDDX42 conditions (**Fig. 4D**), we tested the following assumptions by measuring the Bayesian information criterion (BIC) for the following models (**Extended Data Fig. 6F**):

1. DDX42 blocks spliceosome forward progression in a similar manner to pladienolide B, ($k_{off\ int}$ and $k_{off\ A}$ unchanged upon knockdown).
2. DDX42 affects U2AF dissociation from the intermediate state ($k_{fwd\ 2}$ and $k_{off\ A}$ unchanged upon knockdown).
3. DDX42 affects U2AF dissociation from the intermediate state and A complex ($k_{fwd\ 2}$ unchanged upon knockdown).

Kinetic model of alternative 3' splice site selection

To model the steady-state splicing decisions, we developed a kinetic model comprising a system of differential equations:

$$\frac{\partial[Unbound]}{\partial t} = [I] - 2 \times k_{on}[Unbound] + k_{off\ S2}[E_{S2}] + k_{off\ S3}[E_{S3}] + k_{off\ int\ S2}[Int_{S2}] + k_{off\ int\ S3}[Int_{S3}] = 0$$

$$\frac{\partial[E_{S2}]}{\partial t} = k_{on}[Unbound] - k_{off\ S2}[E_{S2}] - k_{fwd\ 1}[E_{S2}] = 0$$

$$\frac{\partial[E_{S3}]}{\partial t} = k_{on}[Unbound] - k_{off\ S3}[E_{S3}] - k_{fwd\ 1}[E_{S3}] = 0$$

$$\frac{\partial[Int_{S2}]}{\partial t} = k_{fwd\ 1}[E_{S2}] - k_{off\ int\ S2}[Int_{S2}] - k_{fwd\ 2}[Int_{S2}] = 0$$

$$\frac{\partial[Int_{S3}]}{\partial t} = k_{fwd\ 1}[E_{S3}] - k_{off\ int\ S3}[Int_{S3}] - k_{fwd\ 2}[Int_{S3}] = 0$$

$$\frac{\partial[A_{S2}]}{\partial t} = k_{fwd\ 2}[Int_{S2}] - k_{off\ A}[A_{S2}] = 0$$

$$\frac{\partial[A_{S3}]}{\partial t} = k_{fwd\ 2}[Int_{S3}] - k_{off\ A}[A_{S3}] = 0$$

In these equations, the term I is a constant that represents an unchanging concentration of RNA production per unit time. This assumption ensures a steady influx into the system and is crucial for balancing the rates of production and splicing.

We solved this system of equations using Mathematica (Wolfram Research) and defined percent spliced in (PSI) as follows:

$$PSI = \frac{[A_{S2}]}{[A_{S2}] + [A_{S3}]}$$

To formulate the equation for PSI in terms of $\Delta\Delta G$, we made the following substitutions

$$k_{off2} = k_{off3} e^{-\Delta\Delta G/RT}$$

$$k_{off\ int\ 2} = k_{off\ int\ 3} e^{-\Delta\Delta G/RT}$$

While typically $\Delta\Delta G$ is described in terms of K_D and not k_{off} , this is a safe assumption considering we measure almost the same association rate (k_{on}) for U2AF binding the AdML and Mutant sequence (**Fig. 1M**). After this step, the equation is now in terms of U2AF dissociation rates from the downstream 3'SS (k_{off3} and $k_{off\ int\ 3}$) where the relative difference in U2AF dissociation between the two sites is reflected in $\Delta\Delta G$. After simplification and rewriting in terms of $K_f = k_{fwd1}/k_{off1}$ and $K_s = k_{fwd1}/k_{off\ int}$, we get the following equation for PSI:

$$PSI = \frac{(K_f + 1)(K_s + 1)}{(2 K_f K_s) + (K_f + K_s) \left(1 + e^{\frac{-\Delta\Delta G}{RT}}\right) + \left(1 + e^{\frac{-2\Delta\Delta G}{RT}}\right)}$$

We considered multiple models to describe alternative 3'SS selection. In addition to the kinetic proofreading model, we also wanted to consider models with backward rates opposing k_{fwd1} and k_{fwd2} . We have included a comparison of the models in **Figure S5C-E**. Not surprisingly, the kinetic proofreading scheme exhibits a steeper transition between around $-\Delta\Delta G = 0$, indicating higher sensitivity to the underlying binding site. The overall shape of these curves depend on equilibrium constants K_f and K_s . In the *partial* kinetic proofreading scheme where one constant is much lower than another ($K_s = 1$, $K_f = 0.15$), inclusion of a backward rate provides almost no specificity advantage over a standard Michaelis-Menten kinetic scheme (**Figure S5D, Model 2 vs Model 1**). Importantly, while it is clear that all models exhibit the same sigmoidal shape, the difference in U2AF off-rate from the alternative ($k_{off\ S3}$) relative to the canonical ($k_{off\ S2}$) site would have to be much larger for Models 1 & 2 to heavily favor production of S3 isoform (**refer to tables in S5D**). For these reasons, we favor the kinetic proofreading model of alternative 3'SS selection. In principle, the kinetic proofreading model is not incompatible with backward rates opposing k_{fwd1} and k_{fwd2} as long as they're sufficiently slow so that most U2AF molecules leave intermediate complex via $k_{off\ int}$ or $k_{fwd\ 2}$.

As is clear in **Fig. 3C**, this relationship is asymmetric, especially when K_f and $K_s \gg 1$. This is due to how PSI is defined. While we present our kinetic models in terms of equilibrium constants K_s and K_f as well as $\Delta\Delta G$, it is easiest to address this point where the PSI is expressed exclusively in terms of rates. For model 1 (**Extended Data Fig. 5C**) this is:

$$PSI = \frac{k_{fwd} + k_{off\ 3}}{2 * k_{fwd} + k_{off\ 2} + k_{off\ 3}}$$

In this form of the equation, the asymmetry is clear. When $k_{off\ 2}$ is much larger than k_{fwd} and $k_{off\ 3}$, PSI goes to 0 as expected. However, in the alternative scenario, where $k_{off\ 3} > k_{off\ 2}$, the equation will only go to 1 if k_{fwd} is sufficiently small.

Mass-spectrometry

Human Bronchial Epithelial Cells (HBECs) expressing endogenously tagged U2AF1-3XFLAG underwent subcellular fractionation as previously described⁷². In addition, a parental HBEC line lacking the endogenous U2AF1 tag was prepared to serve as a control. This control was critical for identifying nonspecific interactions, ensuring that observed binding events were attributable to the specific interactions of U2AF1-3XFLAG. The experiment was conducted with three biological replicates per sample group, each consisting of approximately 20 million cells. Briefly, cells washed with PBS at room temperature were detached using trypsin, collected in ice-cold PBS, and transferred to pre-chilled 15 ml conical tubes to inhibit degradation. After centrifugation at 500 x g for 5 minutes at 4°C, the supernatant was removed, and the cell pellet weighed. The pellet was then washed with ice-cold 1X PBS and resuspended in hypotonic lysis buffer (HLB) at a ratio of 1 ml per 75 mg of cells, with a 1X final concentration of protease inhibitor (PI) cocktail and phosphatase inhibitor (PhI) solution. Following a 10-minute incubation on ice with intermittent vortexing, the suspension was centrifuged at 800 x g for 10 minutes to collect the cytoplasmic fraction. The nuclear pellet underwent four washes with HLB before being resuspended in nuclear lysis buffer (NLB) at half the initial HLB volume, with PI and PhI added. Nuclei were then sonicated ten times, alternating between 30 seconds of sonication and 30 seconds of cooling. Subsequent centrifugation at 18,000 x g for 15 minutes at 4°C clarified the nuclear extracts.

For immunoprecipitation (IP), 30 µL of magnetic beads (specific for FLAG-tag affinity) were prepared following the manufacturer's instructions (Sigma M8823), washed with NLB, and incubated with 500 µL of nuclear extract overnight at 4°C with rotation to ensure thorough binding. Afterward, beads were washed three times with NLB to remove nonspecifically bound components.

On-bead digestion and TMTpro labeling: Samples were digested on bead by treatment with 50µL each of lysis buffer, reducing buffer, and alkylation buffer provided with the EasyPep kit (Thermo #A40006) then incubated at 25°C 1.5hrs in the dark. Samples were treated with 20µL of 100ng/µL trypsin/LysC and incubated at 37°C overnight for 20hrs with shaking at 1000rpm. Supernatant was then moved to a clean tube and treated with 50µL of 5µg/µL TMTpro in 100%ACN and incubated at 25°C for 2hrs. Samples were then quenched with 50µL of 5% hydroxylamine, 20%FA and incubated for 20 minutes before combining samples within each plex. Each plex was cleaned using EasyPep mini spin columns (Thermo #A40006) and drying eluted peptides by speed-vac.

LC-MS/MS analysis: Each plex was loaded twice onto a Dionex U3000 RSLC in front of a Orbitrap Eclipse (Thermo) equipped with an EasySpray ion source. Solvent A consisted of 0.1%FA in water and Solvent B consisted of 0.1%FA in 80%ACN. Loading pump consisted of Solvent A and was operated at 7 µL/min for the first 6 minutes of the run then dropped to 2 µL/min when the valve was switched to bring the trap column (Acclaim™ PepMap™ 100 C18 HPLC Column, 3µm, 75µm I.D., 2cm, PN 164535) in-line with the analytical column EasySpray C18 HPLC Column, 2µm, 75µm I.D., 25cm, PN ES902). The gradient pump was operated at a flow rate of 300nL/min and each run used a linear LC gradient of 5-7%B for 1min, 7-30%B for 133min, 30-50%B for 35min, 50-95%B for 4min, holding at 95%B for 7min, then re-equilibration of analytical column at 5%B for 17min. All MS injections employed the TopSpeed method with three FAIMS compensation voltages (CVs) and a 1 second cycle time for each CV (3 second cycle time total) that consisted of the following: Spray voltage was 2200V and ion transfer temperature of 300 °C. MS1 scans were acquired in the Orbitrap with resolution of 120,000, AGC of 4e5 ions, and max injection time of 50ms, mass range of 350-1600 m/z; MS2 scans were acquired in the Orbitrap using TurboTMT method with resolution of 15,000, AGC of 1.25e5, max injection time of 22ms, HCD energy of 38%, isolation width of 0.4Da, intensity threshold of 2.5e4 and charges 2-6 for MS2 selection. Advanced Peak Determination, Monoisotopic Precursor selection (MIPS), and

EASY-IC for internal calibration were enabled and dynamic exclusion was set to a count of 1 for 15sec. The only difference in the methods was the CVs used, one method used CVs of -45, -60, -75 and the second used CVs of -50, -65, -80.

Database search and post-processing analysis: Raw files from each plex were batched together as fractions and searched with Proteome Discoverer 2.4 using the Sequest node. Data was searched against the Uniprot Human database from Feb 2020 using a full tryptic digest, 2 max missed cleavages, minimum peptide length of 6 amino acids and maximum peptide length of 40 amino acids, an MS1 mass tolerance of 10 ppm and MS2 mass tolerance of 0.02 Da. Variable modifications of oxidation on methionine (+15.995 Da) and TMTpro (+304.207) on lysine and peptide N-terminus as well as fixed carbamidomethyl on cysteine (+57.021). Percolator was used for FDR analysis and TMTpro reporter ions were quantified using the Reporter Ion Quantifier node and normalized on total peptide intensity of each channel. Proteins were filtered using a FDR cutoff of 1% and have quantitative values in at least 4 samples to be included in the final list. An absolute log2 fold change of 0.6 and pvalue of 0.05 were used to determine if a protein/phosphopeptide was differentially regulated. TMTpro channel assignment were:

Condition	Nuclear Fraction
WT_Control_1	126
WT_Control_2	127N
WT_Control_3	127C
3X_Control_1	131N
3X_Control_2	131C
3X_Control_3	132N

The list of protein in each snRNP were compiled using the 'class_family' complumn from spliceosome database ⁵³ (U2 Associated: "17S U2 snRNP associated," U2: "17S U2 snRNP", U1: "U1 snRNP," U4/U6 & U5 snRNP: combined "U4/U6 snRNP", "U5 snRNP" & "tri-snRNP").

DDX42 knockdown

DDX42 knockdown was performed by electroporation with the neon kit (Invitrogen). We electroporated 500K cells with 20 pMoles of two separate siRNAs pools (Invitrogen silencer AM16708 and Dharmacon siGENOME smartpool M-012393-00-0005) into 120 uL Buffer R (Invitrogen) for a final concentration of 333 nM siRNA. Cells were electroporated with 4-10 ms pulses at 1230 volts. Optimal siRNA concentration was determined by titration (**figure S6A**). For control assays, we used 333 nM siGENOME non-targeting siRNA pool #1 (D-001206-13-05, Dharmacon). Knockdown efficiency was assessed by western blot using a DDX42 polyclonal antibody from Invitrogen (PA5-54828) 3 days post electroporation. For RNA sequencing experiments, this reaction was scaled up to 750K cells per DDX42 knockdown.

Western Blotting

Cells were trypsinized, washed with PBS, and then pelleted (5 minutes at 300 x g). After resuspending in 2x SDS loading buffer, heated at 95 °C for 10 min, lysates were resolved on 4–12% Bis-Tris gels in MOPS running buffer at 200 V. Proteins were transferred to nitrocellulose membranes using the trans-blot turbo transfer system (biorad). Membranes were blocked in 5% non-fat milk in TBST for 1 h at room temperature, then incubated with primary antibody for 1 hr at room temperature. Following washes in TBST, membranes were incubated with secondary

1462 antibody (1:2500) for 1 h at room temperature. After additional washes, proteins were detected
1463 using chemiluminescence.

Antibody	Company	Ref
U2AF1	Abcam	AB172614
U2AF2	Santa Cruz	SC-53942
H2B	Abcam	AB52484
DDX39B	Cell signaling	47258
SF3B1 (SAP155)	MBL	D221-3
DDX42	Invitrogen	PA5-54828
Beta Actin	Sigma Aldrich	A2228

1464
1465
1466 **Data Analysis**
1467 All reported measurements represent mean $\pm$ standard deviation unless otherwise noted.
1468 Parameter estimates were obtained by fitting experimental data to the models described in the
1469 text. Reported uncertainties reflect variability across independent replicates. No formal hypothesis
1470 testing was performed. Single-molecule measurements were aggregated within each replicate
1471 prior to analysis, so uncertainty was calculated using individual replicates, not individual
1472 molecules.

1473
1474
1475

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DonovanSplicesomev23extendedData.pdf](#)