Deployment of non-canonical splicing in tunicate genomes is mediated by divergent U2AF function and re-patterning of snRNA m6A modification

Soo T.C.C.*, Leon A., Waymel E., Barais A., Porta B., Chourrout D. and Henriet S.*[#]

Michael Sars Centre, University of Bergen, Bergen, NO-5020, Norway

** These authors share first authorship*

[#] Corresponding author

Abstract

In eukaryotes, the critical function of the spliceosome is to remove introns from newly transcribed pre-messenger RNA to produce functional mRNA. Spliced segments normally begin with a GT and end with an AG, and precise intron removal relies on the recognition of terminal dinucleotides by the spliceosomal RNA (snRNA) and by the U2AF heterodimer. Here, we reveal how the modification of these molecules was instrumental for shifting spliceosome specificity, enforcing splicing accuracy in genomes where 95% of introns escape the GT/AG rule.

We show that the emergence of non-canonical introns in the *Fritillaria borealis* lineage is associated with the duplication of each U2AF subunit. Subunit paralogues can be combined to form multiple U2AF complexes that involve novel protein-protein interaction rules. Binding assays on RNA and transcriptome analysis of human cells expressing U2AF subunits show that divergent paralogues can contribute to the recognition of non-canonical splice signals at the 3' end of introns. In transfected cells, levels of N6-methyladenosine (m6A) modification on snRNA U2 and U6 are impacted by the presence of *F. borealis* U2AF2 and by AHCYL1, a specific partner of the divergent subunit U2AF2β. In striking contrast with other species, *F. borealis* has lost m6A on the U6 snRNA while U1 snRNA has gained a stable, 5'-terminal m6A. We used structure prediction to show how these snRNA modifications could play decisive roles during the recognition of non-GT/AG introns. In conclusion, we argue that spliceosome function can be profoundly impacted by gene neofunctionalization and post-transcriptional modifications, without implying major changes to conserved genetic components.

Introduction

After transcription, the majority of eukaryotic pre-mRNA will undergo splicing, a processing step that establishes open reading frames by removing introns and by stitching exons together. This reaction is carried out by the spliceosome, a large complex composed of snRNPs *e.g.*, ribonucleoprotein particles (RNPs) assembled around a spliceosomal RNA (snRNA). Inefficient and/or inaccurate splicing may produce aberrant transcripts that

35 cause diseases and developmental abnormalities, and one of the main challenges to
36 spliceosome functions is to identify the exact position of introns across thousands of pre-
37 mRNA bases. For most eukaryotic introns, the 5' splice site (5'ss) is a GT dinucleotide and
38 the 3' splice site (3'ss) is an AG. These dinucleotides and their surrounding sequences are
39 recognized by splicing factors and snRNPs (**Fig1A**). The recognition of the intron by the U1
40 snRNP is often considered as the starting step of splicing. The 5' end of U1 snRNA shows
41 extensive complementary to the eukaryotic 5'ss consensus, which includes the 5'GT and
42 adjacent nucleotides located in the exon and in the intron. Accordingly, mutations in the 5'
43 end of U1 lead to an aberrant choice of 5'ss, producing defective transcripts that can be
44 linked to cancers[1]. The 3' end of introns are recognized by the U2AF complex, a
45 heterodimer composed of two highly conserved RNA-binding proteins. The small subunit
46 U2AF1 contacts the 3'ss while the large subunit U2AF2 binds to a polypyrimidine tract
47 (PPT) usually found immediately upstream[2,3]. The binding of U2AF to the 3'end is required
48 for positioning the U2 snRNP on the branch point (BP) sequence. Once U1 and U2 snRNPs
49 are bound to the intron, the U4/U6.U5 tri-snRNP is recruited and the pre-catalytic
50 spliceosome begins to assemble.

51 The GT/AG rule prevails in yeast and mammalian cell lines, organisms that provided key
52 information about the mechanisms of intron recognition in the laboratory. So far,
53 sequencing projects conducted in various species confirmed that GT/AG introns represent
54 the major class. Because of such ubiquity, it remains difficult to assess why these
55 dinucleotides have been selected as dominant splice site signals during eukaryotic
56 evolution. The catalytically activated spliceosome and Group II self-catalytic introns share
57 a conserved structure, yet these machineries process different splice sites. The so-called
58 minor introns represent another class that corresponds to less than 1% of human introns.
59 Minor introns are *stricto sensu* defined as being recognized by minor spliceosome (U12-
60 type) components U11 snRNA and ZRSR2. Their respective functions are equivalent to
61 major spliceosome (U2-type) components U1 snRNA and U2AF[4], but they bind to different
62 RNA sequences. While AT/AC termini represent hallmark signals for minor introns, this
63 class of introns also includes splice sites that differ only to a little extent from the
64 consensus of major introns. Intron removal at sites that differ substantially from the GT/AG
65 consensus was observed in various eukaryotes, even in organisms that lack the minor
66 spliceosome[5,6]. During evolution, splicing machineries with different selectivity have kept
67 an active site organized around a similar architecture[7-9]. The conserved stereochemistry
68 could define the limits of sequence evolution at intron ends, but it also implies that the
69 spliceosome may accommodate a larger variety of splice sites than what genomes usually
70 show.

71 The results gained during our last study challenged the consensus about intron recognition
72 mechanisms, by showing that non-GT/AG introns are abundant and diverse in the
73 larvacean tunicate *Fritillaria borealis*[10]. Non-canonical splice sites correspond to terminal
74 inverted repeats of non-autonomous DNA transposons ("introners") and as shown in other
75 organisms, splicing offers a mechanism to convert gene-disrupting transpositions into
76 introns[11]. Although other deviations have been documented in some unicellular
77 eukaryotes[12,13], the case of *F. borealis* is exceptional by its amplitude, since less than 10%
78 of introns adhere to the GT/AG rule. Like other larvaceans, the U12-type spliceosome is
79 absent in *F. borealis*[10,14], and we could show that the U2-type spliceosome is necessary for
80 removing both canonical and non-canonical introns. How non-canonical introns were
81 recognized remained elusive at that time and until recently[15], technical progress was
82 hampered by the lack of a laboratory culture for this organism.

83 Apart from intron-poor organisms that tend to lose spliceosome genes, the splicing
84 machinery has remained conserved during eukaryotic evolution[16,17]. One could assume
85 that genome-wide modification of splice site identity would involve important changes in
86 protein and RNA sequences that normally recognize introns, but we showed instead that
87 the U2-type spliceosome is extensively conserved in *F. borealis* - including a strict
88 conservation of the 5' end of U1 snRNA. It suggests that the shift in spliceosome selectivity
89 does not rely on major transformation of genetic components but rather on discrete,
90 specific changes that have appeared in the *F. borealis* lineage. The experiments described
91 here allowed us to identify when and where such modifications have been implemented
92 during evolution, and to determine their impact on intron recognition.

93 **Results**

94 *Duplication of U2AF small and large subunit genes coincides with the emergence of a new*
95 *class of introns in larvaceans*

96 Genetic resources for *F. borealis* were initially limited by specimen availability, only
97 permitting to deploy short-read sequencing that produced genome assemblies with an
98 important level of fragmentation[18]. We took advantage of recent developments in long-read
99 sequencing techniques and laboratory culture to improve the genome assembly, and to
100 establish a collection of full-length transcripts. The latest assembly features a smaller
101 number of contigs and higher N50 (**Table S1**), which is useful for aligning full-length cDNA
102 on the *F. borealis* genome. Based on the annotation of >30.000 introns discovered after
103 transcript alignment, we could count only a small fraction of canonical introns (5%), while
104 AG/AC and AG/AT were the most frequent termini seen among non-canonical introns
105 (respectively 18 and 16%) (**Fig1B**). To gain better insight into intron evolution in the
106 larvacean lineage, we also sequenced DNA and RNA of two uncultured Fritillariidae

species, *Fritillaria haplostoma* and *Appendicularia sicula*[15]. Even though these new genomes are still of lower quality, the long-read transcriptomes obtained with all three species allowed us to find gene orthologues and to map introns in another species, *Fritillaria pellucida*, for which no RNA-seq data is available.

Next, we compared our intron dataset to a survey of major and minor introns recently established for a variety of genomes[4]. Based on the terminal dinucleotides, we could observe a larger diversity of introns in Tunicates compared to other chordates. In agreement with earlier results[19], intron diversity was highest in the four Fritillariidae genomes (**Fig1B**). However, for the majority of Tunicates genome, the most frequent 3'ss corresponds to the dinucleotide AG, present in GT/AG introns and nearly canonical GX/AG introns (with X being either C, G or A) – with frequencies ranging from 43% in *F. haplostoma* up to 90% in *O. dioica*. We found a typical, uridine-rich PPT upstream the 3'AG, which is also present in other tunicate introns ending with 3'AC or 3'AT. Introns of the *F. pellucida* and *F. borealis* genomes were exceptions, since the 3'AG is less frequent than other termini (respectively 38% and 15%). Instead, the most frequent 3'ss used in these genomes is a 3'AT, found in 44% of *F. pellucida* introns and 34% of *F. borealis* introns, which is preceded by a regular PPT. In *F. borealis*, the 3'AC is the second highest 3'ss in frequency (28%), while the frequency in other species does not exceed 15% of introns. Unlike other species, the 3'AC intron subclass in *F. borealis* is characterized by a cytidine-rich PPT often interrupted by purines.

The atypical PPT of 3'AC introns in *F. borealis* resembles the PPT of minor-like introns[4], which is at odds with the lack of U12-type spliceosome in this species. However, examining the gene complement revealed that *F. borealis* is the only species with two versions of U2AF1 and U2AF2 genes. Versions U2AF1α and U2AF2α share high sequence identity with the single copy that is found in other deuterostomes, while versions U2AF1β and U2AF2β show higher divergence (**Fig2A**). Most differences were found in terminal arginine- and serine-rich (RS) domains, while RNA-binding domains are better conserved (**Fig2B**). The *F. borealis* paralogues were found expressed at comparable levels in the transcriptome and *in situ*, ruling out that divergent versions correspond to pseudogenes (**FigS1**). While *F. pellucida* U2AF1 is homologous to the conserved U2AF1α, we could evidence that its U2AF2 version corresponds to a chimera between U2AF2α and U2AF2β, indicating that the common ancestor with *F. borealis* possessed two copies, that merged into one in *F. pellucida* (**Fig2C**, **FigS2**).

Divergent U2AF subunits form heterodimers with novel RNA binding properties

In principle, four different heterodimers could be formed if all the paralogues of U2AF subunits were expressed in the same cell. Sequence divergence between paralogues

143 suggests these complexes could be associated with distinct biochemical properties,
144 providing new functions during splicing and a division of labor. We supposed that if various
145 U2AF possessed different affinities for RNA, they could provide a flexible way to recognize
146 sequence diversity at the 3' end of *F. borealis* introns, which includes the canonical
147 combination of 3'AG preceded by a uridine-rich PPT and other combinations established
148 with other 3'ss and the cytidine-rich PPT.

149 To explore this hypothesis, we first tested if different combinations of U2AF1 and U2AF2
150 paralogues could form protein complexes. The assembly of U2AF is mediated by
151 interactions between the U2AF homology motif (UHM) of U2AF1 and the U2AF ligand motif
152 (ULM) of U2AF2[20], which we could detect in the *F. borealis* paralogues (**Fig2B**). Using
153 tagged U2AF1α as bait, we could recover U2AF2α and U2AF2β co-expressed in
154 mammalian cell culture (**Fig2D**, lanes 1-6). In comparison, the binding to U2AF2
155 paralogues was strongly reduced with the bait U2AF1β (**Fig2D**, lanes 7-9). One likely
156 reason is the replacement of the conserved F139 of the UHM by a V140 in U2AF1β, which
157 cannot establish π–interaction with the Tryptophane present in the U2AF2 ULM[20] (**FigS3**,
158 **FigS4**). However, we could observe that U2AF1β interacts with U2AF2β slightly better than
159 with U2AF2α, and that complexes formed with U2AF1β are stabilized by monovalent salts
160 – unlike those formed with U2AF1α (**Fig2D**, compare lanes 4-6 with lanes 10-12). Using a
161 bacterial expression system, we could purify the two complexes formed with U2AF2α
162 (U2AF1α/2α, U2AF1α/2β) but only a single complex formed between U2AF1β and U2AF2β
163 (U2AF1β/2β) (**Fig2E**). We could not isolate U2AF1β/2α, suggesting that the weak
164 interaction detected in mammalian cell lysate is stabilized by other molecules absent in
165 bacteria. Removing the N-terminal part of U2AF2β significantly reduced its interaction with
166 U2AF1β (**Fig2F**), suggesting that the loss of affinity associated with the mutation of UHM in
167 U2AF1β can be rescued by unknown interactions established with the RS domain of
168 U2AF2β.

169 We first tested how different U2AFs bind to RNA by using EMSA and oligoribonucleotide
170 probes designed to mimic an intronic 3' AG, preceded either by a canonical PPT (ORN.PyU)
171 or a C-rich PPT (ORN.PyC). We tested a third probe derived from an *F. borealis* intron,
172 where the 3'ss is replaced by an AU, preceded by a sequence rich in purines (ORN.AU)
173 (**Fig3A**). In agreement with the preference for uridine reported for human U2AF2[2], the
174 U2AF1α/2α complex interacted best with ORN.PyU, followed by ORN.AU and ORN.PyC
175 (**Fig3B**, **Fig3C**). The other protein combinations, which include the complexes U2AF1α/2β
176 and U2AF1β/2β and the single subunits U2AF1α and U2AF1β, interacted significantly
177 better with ORN.AU than with other probes - for which levels of gelshift were almost
178 undetectable (**Fig3C**). It is still unclear to us whether the reduced affinity of complexes

179 formed with U2AF2β is linked to specific substitutions present in the RNA binding domains
180 (**FigS5**). However, when we assembled a mutant version missing the N-terminal part
181 (U2AF2$\beta^{\Delta N}$) together with U2AF1α, the resulting complex U2AF1α/2$\beta^{\Delta N}$ could bind ORN.AU
182 strongly, with an affinity close to the one measured for U2AF1α/2α on ORN.PyU (**Fig3C**,
183 **Fig3E**).

184 We predicted that sequence divergence between the conserved U2AF1 and U2AF1β could
185 have an impact on the interaction with RNA. In the first zinc finger of U2AF1α, S38
186 corresponds to a highly conserved residue (S34 in human U2AF1) that is replaced by P39
187 U2AF1β (**FigS3**). Based on the structure of yeast U2AF1 bound to UAGGU[21], the
188 substitution in U2AF1β could interfere with the recognition of the position preceding the
189 3'AG at the end of the canonical introns (**FigS6**). We tested the binding of U2AF1
190 paralogues to RNA with an approach based on RNAcompete[22]. The U2AF1 paralogues were
191 immobilized on beads, and we tested their interaction with a collection of short RNA
192 probes derived from the 3' ends of *F. borealis* introns (**Fig4A**). We examined base
193 composition among the high affinity probes, by looking first at the frequency of
194 trinucleotides combinations with a penultimate A. Compared to U2AF1α, trinucleotides
195 ending with AG are largely depleted in RNA recovered with U2AF1β, whereas those ending
196 with AC are most frequent (**Fig4B**). In U2AF1α-bound RNA, UAG triplets are more frequent
197 than other triplets ending with AG, showing that the protein has conserved the ability to
198 discriminate U_{-3}. Dinucleotides frequencies revealed a preference for C in U2AF1β-bound
199 RNA, and for G and U in U2AF1α-bound RNA. Similar preferences were seen for U2AF1β
200 when comparing it to the *O. dioica* U2AF1 orthologue (**FigS7**). Based on these results and
201 the sequence logos produced with U2AF1-bound RNA (**Fig4C**), we propose that during
202 splicing, the respective targets of U2AF1α and U2AF1β could be subclasses of uridine- and
203 cytidine-rich introns.

204 *U2AF2α and U2AF2β induce distinct splicing events in mammalian cells*

205 Our results show at least three stable combinations of U2AF may exist in *F. borealis*.
206 Depending on whether conserved or divergent versions of subunits are included, the
207 different U2AFs are susceptible to bind to distinct types of RNA sequences. Testing U2AF
208 function in *F. borealis* is still a challenging task, which implies developing specific
209 approaches for genetic manipulation in this organism. Based on the good conservation of
210 functional domains in *F. borealis* U2AF1 and U2AF2, we assumed these proteins could
211 influence spliceosome function in a heterologous system. Expression vectors containing
212 U2AF1 or U2AF2 paralogues were co-transfected in mammalian cell cultures, and we
213 studied their impacts on the splicing of endogenous transcripts.

214 By inducing the selection of unusual splice sites, the expression of foreign splicing factors
215 could induce the expression of transcripts carrying a premature termination codon (PTC),
216 that are degraded by the nonsense-mediated decay pathway (NMD). To ease the detection
217 of rare and aberrant splicing events, we first used a targeted approach to check the
218 splicing of highly expressed genes, after a cycloheximide treatment to inhibit NMD (**Fig5A**).
219 The treatment allowed to increase the number of splicing isoforms and raised the
220 detection rate of splicing events at GT/AG borders by 1.7-fold, and at other borders by 2.7-
221 fold (**Fig5B**). We used long-read sequencing to compare isoform expression between cells
222 transfected with different combinations of U2AF subunits. For all samples, the expression
223 of non-canonical isoforms (produced by the splicing of at least one GT/AG intron)
224 remained low in comparison to canonical isoforms, but they were substantially more
225 abundant in cells transfected with the U2AF1β/2β combination (**Fig5C**). Each U2AF
226 combination tested could be associated with the expression of a specific set of splicing
227 isoforms, including those whose expression is largely reduced in other samples (**Fig5D**).
228 Some unique isoforms were completely absent from controls cells transfected with a
229 LacZ-expressing construct and taken together, these results show that different U2AFs
230 induce the processing of a specific set of splice sites. For each sample, we next examined
231 the sequence composition in unique introns, whose splicing is not detected in other
232 samples. While no major differences appeared at the level of 3'ss, we could associate
233 U2AF2β expression with the splicing of introns with weaker PPTs (**Fig5E**, **Fig5F**). Our results
234 did not reveal a major impact of U2AF1β on splicing, perhaps due to a competition with the
235 endogenous U2AF1, that could in principle bind to U2AF2α and U2AF2β.

236 We employed a regular RNA-seq approach to examine splicing globally in cells expressing
237 U2AF2 paralogues, against a control sample that expresses a transmembrane envelope
238 (Env)[19]. U2AF2α has a larger impact on gene expression compared to U2AF2β (**Fig6A**).
239 However, GO terms associated with differentially expressed genes (DEGs) primarily
240 identified functional categories linked to protein homeostasis, suggesting effects on gene
241 expression were provoked by the over-expression of transfected genes rather than a
242 specific impact on transcription programs. More than a third of the top DEGs between
243 U2AF2α and U2AF2β correspond to non-coding genes (**Fig6B**). We interpreted this result
244 as a possible consequence of novel splicing events induced by *F. borealis* U2AF2, leading
245 to the degradation of numerous protein-coding transcripts by NMD and to an enrichment in
246 non-coding RNA. Intriguingly, we saw antagonistic effects on the expression of antisense
247 lncRNA transcribed from SMN1 and SMN2 genes, with SMN-AS1 overexpressed in
248 U2AF2α-transfected cells and SMN2-AS overexpressed in U2AF2β-transfected cells. It is
249 still unclear if these differences reflect the effect of U2AF2 paralogues on the processing of
250 lncRNA or on the expression of SMN1 and SMN2 genes[23].

251 After mapping reads, rMATS revealed 42931 alternative splicing (AS) events associated
252 with U2AF2α expression, and 26373 AS events associated with U2AF2β expression. Exon
253 skipping is largely responsible for the excess of events detected with U2AF2α, and it is the
254 most frequent class of AS detected in this sample (15969) (**Fig6C**). Given that U2AF2α and
255 human U2AF2 may share similar preferences for RNA, we suppose that the presence of
256 U2AF2α can contribute to the exon skipping activity already shown for human U2AF. The
257 binding of U2AF to intronic sites was indeed shown to interfere with the selection of 3'ss
258 located immediately downstream[24]. While lower concentrations of U2AF may leave these
259 intronic sites vacant, hybrid U2AF formed with U2AF2α in transfected cells could bind
260 these sites, leading to the strong increase in exon skipping events. Hybrid U2AF that
261 includes *F. borealis* U2AF2α or U2AF2β and endogenous U2AF1 are likely to form in
262 transfected cells, based on our protein interaction assays. However, the lower number of
263 registered events suggest U2AF2β is largely unable to induce exon skipping with the same
264 mechanism – possibly due to different affinity for RNA. We used the set of AS events
265 identified with rMATS to check for sequence differences around the 3'ss. When examining
266 splice sites that are specifically processed in one U2AF2 paralogue transfection but not
267 with the other paralogue or the Env control, the only AS class showing substantial variation
268 was intron retention. Here, splicing induced by U2AF2α is associated with uridine-rich PPT
269 before the 3'ss (**Fig6D**), while cytidines are more frequent in the PPT of introns spliced
270 under U2AF2β expression. Consistent with results gained with the targeted approach,
271 these observations show that U2AF2α or U2AF2β induce the splicing of distinct sets of
272 introns in mammalian cells, which can be distinguished based on their PPT.

273 *U2AF paralogues and AHCYL1 regulate N^6-methyladenosine levels in snRNA*

274 Our experiments revealed an unexpected function of the RS domain of U2AF2β by showing
275 it interferes with the binding of U2AF on RNA. This region of the protein also mediates
276 dimerization with U2AF1β while in humans, the RS domain of U2AF2 is needed for binding
277 to Prp19 and for transcription-coupled splicing[25]. We supposed that the discovery of
278 proteins bound by U2AF2α and U2AF2β could provide novel indications about their
279 function beyond 3'ss recognition, and we used a co-immunoprecipitation assay to recover
280 the proteins that interact with U2AF2α or U2AF2β in mammalian cells (**Fig7A**).

281 Mass spectrometry (MS) revealed Adenosine Homocysteinase-like proteins (AHCYL, also
282 known as IRBIT) as a specific partner of U2AF2β. The proteins AHCYL1 and AHCYL2 (also
283 known as SAHH2 and SAHH3, respectively) share strong sequence identity, and they are
284 homologues of Adenosine Homocysteinase (AHCY, also known as SAHH). Diagnostic
285 peptides identified with MS suggest AHCYL1 as the primary partner of U2AF2β. AHCYL1
286 and AHCYL2 have been reported to interact with AHCY (**Fig7B**), and some of the peptides

287 found in U2AF2β co-IP are also conserved between AHCYL proteins and AHCY. Using
288 specific antibodies, we could confirm that AHCYL1/2 is a protein partner of U2AF2β that is
289 efficiently recovered with co-IP (**Fig7C**).

290 The enzymatic activity of AHCY contributes to stabilizing cellular pools of SAM, the
291 substrate of methyltransferases. AHCYL1 is considered catalytically inactive and while it
292 could interfere with AHCY function, the impact on histone methylation is still the subject of
293 debate[26]. Interestingly, AHCYL1 and AHCYL2 bind to ZCCHC4, the methyltransferase
294 responsible for N^6-methyladenosine (m6A) modification of 28S rRNA, but their impact on
295 base modification is not documented[27]. Recent evidence studies have shown that m6A can
296 regulate splicing[28,29] and based on our co-IP results, we decided to check the effect of
297 AHCYL1/2 and *F. borealis* U2AF2 expression on the deposition of m6A on snRNA. Internal
298 m6A is frequent in mammalian U2 and U6 snRNA, while for other snRNA m6A maps to the
299 transcription start nucleotide and is kept at a low abundance by the FTO demethylase[30].
300 Knocking out the expression of mammalian AHCYL1 with CRISPR/cas9 provoked a 1.5-fold
301 increase of m6A signals on U2 and U6 snRNA (**Fig8A, FigS8A, SI**), showing that in normal
302 conditions AHCYL1 represses the modification of these target adenosines. For both *F.*
303 *borealis* U2AF2 paralogues, expression in mammalian cells was associated with reduced
304 m6A levels in U6, while U2AF2α specifically reduced the m6A levels in U2 (**Fig8B, FigS8B,**
305 **SI**). Although we could reveal only modest effects, our results show that AHCYL1 and
306 U2AF2 can regulate the m6A modification of internal adenosines, opening for a new
307 mechanism that is distinct from the FTO-mediated control of N^6 methylation on the 5'-
308 terminal adenine.

309 *Larvacean snRNA shows specific N^6-methyladenosine modification patterns*

310 As *F. borealis* U2AF2 paralogue could reduce m6A levels on mammalian snRNA, we
311 considered the possibility that the snRNA of *F. borealis* could carry base modifications that
312 differ from other species. We examined m6A levels in the snRNA from a variety of
313 organisms, including the larvacean *Oikopleura dioica* and the ascidian *Ciona intestinalis*.
314 The full set of major snRNA could be detected in all samples (**Fig9A**), but signals
315 corresponding to m6A modification were distributed differently in larvacean snRNA
316 (**Fig9B**). First, m6A levels on U6 are considerably decreased in *F. borealis* compared to
317 other species. In yeast, nematodes and humans, the methyltransferase METTL16 is
318 responsible for modifying A_{43} in U6 snRNA, and the loss of m6A in *F. borealis* U6 snRNA
319 does not seem to be caused by the loss of methyltransferase activity, since the genome
320 conserved all m6A writers with demonstrated roles in RNA modification[31] (**Fig9C**).
321 Therefore, we cannot exclude that m6A loss on U6 snRNA is caused by the gain of lineage-
322 specific molecules, which could include U2AF2β. Second, both larvaceans showed

323 prominent m6A signals corresponding to U1 snRNA, while using the same technique, m6A
324 could be detected on mammalian U1 snRNA only in an $FTO^{-/-}$ background[30]. Using a
325 procedure recently made available for direct m6A mapping[32], we confirmed that the signal
326 detected with m6A immunoblot, corresponds to a m6A modification that is present at the
327 first adenine of U1 snRNA in *F. borealis* and *O. dioica* (**Fig9D**). Interestingly, we could not
328 find an FTO homologue outside mammals, showing that in organisms like *C. intestinalis*
329 where m6A modification on U1 snRNA is undetectable, other erasers could handle the
330 control of m6A levels. Intriguingly, the expression of U1 snRNA appears lower in *F. borealis*
331 compared to other species (**Fig9A**). We cannot exclude that the lower signal could be
332 related to the probe's characteristics (labelling efficiency for instance), but it could also
333 show that U1 snRNA plays a more limited role during pre-mRNA processing.

334 *Impact of m6A re-patterning for the recognition of 5'ss*

335 The transcription start adenine A_1 of U1 snRNA also bears a 2'-O-methyl in mammals[30],
336 and pseudo uridylations at present positions Ψ_4 and Ψ_5 in U1 snRNA can regulate splicing
337 as they are directly engaged in the duplex with the 5'ss[33]. In the absence of a strong
338 complementarity with the non-canonical 5'ss of *F. borealis*, it remains difficult to predict
339 whether the 5'end of U1 snRNA will anneal to the same position on the pre-mRNA or if it
340 would instead use a shifted register, as shown by others[34]. Because it has a stabilizing
341 effect on base stacking[35], unpaired $m6A_1$ could help to form duplexes between U1 snRNA
342 and non-canonical 5'ss (**Fig10A**). The m6A present at this position could also regulate the
343 transfer of the 5'ss from U1 snRNP to the U4/U6.U5 tri-snRNP when the pre-catalytic
344 spliceosome is assembled. During this step, Prp28 unwinds the 5' end of U1 snRNA,
345 making the 5'ss accessible to interact with the ACAGAGA box of U6 snRNA[36].

346 Compared to other species, adenines are less frequent beyond the position +2 of *F.*
347 *borealis* introns and we see instead an over-representation of guanines (**Fig10B**). This
348 region is normally targeted by the m6A-modified ACAGAGA box, raising the possibility that
349 the loss of conserved m6A in U6 snRNA is related to 5'ss divergence in this organism.
350 When interacting with the eukaryotic 5'ss consensus, the modified $m6A_{43}$ in U6 snRNA
351 establishes a non-Watson-Crick base pair with the adenine at position +4 of the intron[37]
352 under the *trans* Hoogsteen / Sugar geometry (tHS)[38,39] (**Fig10A**). Reducing m6A levels in the
353 ACAGAGA box favors the splicing of 5'ss where a uridine is present at position +4,
354 suggesting that another geometry can be accommodated in the duplex[37]. By aligning the
355 U6 snRNA on the 5'ss consensus of the two main classes of *F. borealis* introns, we predict
356 that tHS interactions can be established between adenines of the ACAGAGA box and the
357 conserved guanines at positions +2, +4 and +6 of the 5'ss. The loss of m6A modification in
358 the ACAGAGA box could play a critical role in stabilizing tHS pairing with guanine, by

359 enabling an additional hydrogen bond between the N6 amine of A_{43} of U6 and the sugar of
360 the opposite guanine on the 5'ss (**Fig10C**).

361 **Discussion**

362 By placing a variety of organisms under molecular scrutiny, genome projects conducted
363 during the last decade have led to fundamental discoveries about the origins of introns and
364 the evolution of the spliceosome. Important changes to the repertoire of spliceosomal
365 genes were revealed in unicellular organisms, and these results have proven valuable for
366 appreciating how splicing function can be preserved and how its mechanisms have
367 adapted to intron variety[16,40]. Such knowledge is instrumental for deciphering splicing
368 dynamics in the complex genetic environment of the human genome. By revealing how the
369 spliceosome can change its selectivity, we prove that organisms genetically distant from
370 mammals are still relevant for illuminating the function of conserved cellular machinery.
371 On a broader perspective, the results presented in this study open new perspectives about
372 the evolvability of splicing across eukaryotes.

373 *The divergence of U2AF and its impact on 3'ss recognition*

374 In this study, we showed that the divergence between the paralogues of *F. borealis* U2AF
375 subunits affected their interactions with RNA and their ability to interact with protein
376 partners. These changes have an impact on splicing function, as we could document that
377 the presence of diverse *F. borealis* U2AF induce specific splicing events in mammalian
378 cells. The function and RNA-binding properties of U2AF1 and U2AF2 have been extensively
379 documented in mammals, and our results bring novel knowledge by showing how
380 evolution can change the biochemistry of these crucial splicing factors. Documenting with
381 biophysical approaches how the different U2AF complexes present in *F. borealis* interact
382 with RNA, could reveal if the recognition mechanisms canonically employed by U2AF[41]
383 have been adapted to the intron landscape of *F. borealis*, and which mutations in U2AF1β
384 and U2AF2β determine the divergent affinity for RNA. In both U2AF2 paralogues, the N-
385 terminal region upstream the ULM corresponds to a RS domain rich in polar residues. The
386 inhibitory effect on the RNA binding function in U2AF2β might be controlled by
387 phosphorylation, as was shown for the C-terminal RS domain of SRSF1[42]. Although full-
388 length U2AF2α showed a strong affinity for RNA, we cannot exclude that the RS domain
389 also plays a regulatory function in this paralogue. Compared to U2AF2β, one remarkable
390 difference in U2AF2α is a proline stretch between residues 133 and 137, which could
391 reduce the flexibility of the disordered N-terminus and constrain its ability to interact with
392 RRMs (**FigS5**). Further experiments employing mutant U2AF2 should figure out the impact
393 of N-terminal domain phosphorylation and proline conformation on the affinity for RNA.

394 *Approaching the new functions of U2AF*

395 Expression of U2AF genes *in situ* shows distinctive patterns in *F. borealis* (**FigS1**). The
396 divergent paralogues are expressed in the germline, while the conserved paralogues are
397 preferentially expressed in tissues like the tail muscle and gut cells. These patterns may
398 reflect a sub functionalization, corresponding for instance to the control of tissue-specific
399 splicing events. Deploying spatial transcriptomics in *F. borealis* will be instrumental for
400 determining if the processing of different classes of introns is linked to cell type, and if it
401 correlates with the expression of different versions of U2AF. Dissecting spliceosome
402 functions in *F. borealis* is still a challenging task and in the short term, we believe that
403 heterologous systems such as mammalian cell cultures will remain instrumental for
404 gaining answers about non-canonical intron recognition. With genetic engineering to
405 replace the endogenous U2AF genes with *F. borealis* U2AF1β and U2AF2β, we could gain a
406 background-free system that is adequate to reveal the links between RNA binding and
407 splice site selection.

408 By showing that U2AF2β binds AHCYL1 in mammalian cells, we revealed an unexpected
409 link between splicing factors and regulators of the SAM metabolism. A variety of roles have
410 been attributed to AHCYL1, but how it interferes with AHCY function remains largely
411 unknown[26]. The protein sequence of AHCYL1 is remarkably well-conserved in tunicate
412 genomes, which are otherwise characterized by frequent gene loss and remarkably high
413 evolution rates[43]. It suggests a strong selective pressure for preserving AHCYL1 functions,
414 and future studies should test its impact on animal development and gene expression.

415 *Evolutionary and chemical constraints to the selectivity of the spliceosome*

416 Most components of the human spliceosome are conserved in intron-rich yeasts, but
417 cases of gene loss were reported in several lineages of unicellular eukaryotes[16,17]. For
418 these organisms, the loss of snRNP genes appears to be linked to reduction of intron
419 diversity, often in a context of rapid evolution. These studies have been instrumental for
420 revealing the function of genes that have been lost during evolution, and they recently
421 showed that the massive reduction of splice site diversity observed in Saccharomycotina
422 genomes coincided with the loss of non-essential proteins that enforce splicing fidelity by
423 repressing the usage of cryptic splice sites[16]. Nematodes and larvacean tunicates are the
424 only metazoan phyla that have lost all snRNA components of the U12-type spliceosome,
425 and for these genomes where a substantial number of non-GTAG introns remain, there is
426 no apparent link between the loss of spliceosome genes and the diversity of splice
427 sites[4,10,43]. We suppose that losing the U12-type equipment can impact the ability to
428 process a larger variety of introns, which in the long term could compromise the genome's
429 resistance to splice site erosion or to invasions of introner transposons carrying weak

430 splice sites[10,11]. A viable solution during evolution would be to compensate the loss of
431 flexibility in the splicing machinery by introducing new components, or by modifying the
432 function of existing ones.

433 Mutation of essential genes such as U2AF1 and U2AF2 can compromise cellular functions,
434 and the gain of extra copies also bears the risk of giving rise to negative dominants. By
435 analogy to essential splicing factors that have a functional homologue in the minor
436 spliceosome, the risks associated with maintaining U2AF duplicates could be outweighed
437 by a critical need for removing non-canonical introns. Retrospectively, the case of *F.*
438 *borealis* U2AF appears as a common example of neofunctionalization mediated by gene
439 duplication. In this organism, the situation of U1 snRNA genes appears more subtle. Here,
440 the answer to the massive change of 5'ss identity was not evolving U1 snRNA sequences
441 adapted to new introns - *F. borealis* has two U1 snRNA genes whose 5' ends are strictly
442 identical[10]; but instead, changing the basal level of $m6A_1$ in U1 snRNA. This could be an
443 advantage, as the possibility to duplicate and evolve divergent copies of U1 snRNA could
444 be constrained to a much higher level by the critical role of U1 during splicing. The
445 presence of m6A in mammalian U1 snRNA has been examined in a handful of studies only,
446 and the results gained in our study open the exciting possibility that the regulation of this
447 modification could have an impact on U1 function and splice site selection. The effect of
448 $m6A_1$ on U1 snRNA function during splicing should be examined with adequate
449 experimental design because on its own, transcriptome analysis of $FTO^{-/-}$ cells cannot
450 permit to isolate the effects caused by the presence of $m6A_1$ in U1 snRNA[30], from those
451 caused by the loss of m6A on pre-mRNA sequences[44].

452 We propose that the large diversity of intron ends present in the *F. borealis* genome
453 ultimately reflects the inherent flexibility of the spliceosome's catalytic site, which can
454 potentiate the effect of snRNA base modifications during splice site selection. Splicing
455 depends on the conservation of non-Watson-Crick base pairing between the terminal
456 nucleotides of introns[7]. The isostericity of G-G and A-C terminal pairs (respectively formed
457 with GT/AG and AT/AC introns) has been raised as the molecular basis to explain why the
458 major spliceosome can process AT/AC introns in laboratory experiments[6]. Based on
459 mutant analysis[45], G-G and A-C intron termini were predicted to adopt a *trans* Watson-
460 Crick / Watson-Crick (tWW) conformation in the human spliceosome. In *F. borealis*,
461 despite the extreme variety of splice sites, the termini of at least 55% of introns could form
462 tWW interactions isosteric or nearly isosteric to G-G *i.e.*, with isodiscrepancy indexes
463 under 3.5[46] (**TableS2**). If we consider an alternative *cis* Watson-Crick / Hoogsteen (cWH)
464 conformation that is supported by CryoEM data of the *S. cerevisiae* spliceosome[8], this
465 proportion will correspond to 45% of *F. borealis* introns. Examining the presence of other
466 base modifications on the snRNA and on the pre-mRNA and ultimately, high-resolution

467 experimental structures of the *F. borealis* splicing machinery will give definitive answers to
468 this hypothesis.

469 **Acknowledgments**

470 Simon Henriet acknowledges financial support from the Norwegian Cancer Society
471 (Pioneer grant #227930 – "Non-canonical splicing as a source of aberrant genetic
472 information and genomic"). We thank Ave Tooming-Klunderud at the Norwegian
473 Sequencing Centre in Oslo for help during Pacbio sequencing experiments. We thank
474 Teshome Tilahun Bizuayehu and Jonas Brandenburg at the Michael Sars Centre for help
475 during ONT sequencing experiments. We thank the Norwegian Research and Education
476 Cloud for generous access to computing infrastructure. Larvacean samples were provided
477 by Anne Aasjord, animal facility manager at the Michael Sars Centre. Ascidian samples
478 were provided by the Chatzigeorgiou lab at the Michael Sars Centre. Mouse ES cells were
479 provided by Max Jordi Makem Pekouankouang at the Michael Sars Centre.

480 **Material and Methods**

481 *Animal samples*

482 Specimens of *Fritillaria borealis* were sampled from laboratory populations established
483 with specimen collected near Bergen, on the west coast of Norway[15]. Specimens of
484 *Fritillaria haplostoma* and *Appendicularia sicula* were collected directly near Bergen, on
485 the west coast of Norway[10,15]. Specimens of *Fritillaria pellucida* were collected near La
486 Jolla, (CA, USA)[10].

487 *Genome sequencing*

488 We extracted high-molecular weight genomic DNA from mature *F. borealis* gonads
489 dissected under microscope, with the Chomzynsky procedure. The DNA corresponding to
490 190 individuals was pooled, and 1.5 µg was used for long-read sequencing with Oxford
491 Nanopore Technologies (ONT) platforms. We prepared ONT libraries following the
492 manufacturer's recommendations (Ligation sequencing DNA V14, ONT) but increasing
493 incubation times for adapter ligation to one hour and bead elution time to 14 hours to
494 improve DNA recovery. Between 250 and 300 ng of ONT library were sequenced on a
495 MinION flow cell (R10.4.1). Available pores rapidly decreased during the run, presumably
496 due to pore blocking by polysaccharide contaminants present in DNA preparations. To get
497 enough coverage for genome assembly, we collected over 29.10^4 reads with N50=4.9 kb
498 produced by sequencing three independent libraries.

499 The DNA from whole bodies of four mature *F. haplostoma* individuals was prepared as
500 described above and prepared for Illumina sequencing (Nextera XT DNA Library

501 Preparation Kit). These libraries were barcoded and pooled with *A. sicula* libraries prepared
502 for an earlier study, using whole-genome amplified DNA (WGA-DNA) obtained from
503 individual specimens[10]. Pooled libraries were sequenced with 150bp PE on a 1.5B NovaSeq
504 X flow cell at the Norwegian Sequencing Centre (NSC, Oslo, Norway), producing 56.10^7
505 and $87.\ 10^6$ reads for the *F. haplostoma* and for the *A. sicula* libraries, respectively. We
506 prepared ONT libraries for *A. sicula* using WGA-DNA produced from single individuals, as
507 described above. We sequenced 400 ng of *A. sicula* ONT library on a PromethION 2 solo
508 flow cell (R10.4.1) for 16.5 hours, producing 2.10^6 reads with reads N50=6.4 kb.

509 *Transcriptome sequencing*

510 We extracted total RNA (Nucleospin RNA XS, Macherey-Nagel) from *F. borealis* (24 larvae,
511 3 immature and 8 mature adults), *F. haplostoma* (6 mature adults) and *A. sicula*
512 specimens (one mature adult). Individual transcriptomes were converted to cDNA and
513 amplified with barcodes for a total of 17 cycles, following Pacbio's guidelines for isoform
514 sequencing (Iso-Seq Express Template preparation). SMRTbell libraries were prepared
515 from pools of barcoded cDNA and sequenced with two SMRT cells on a Sequel II
516 instrument at the NSC. Circular Consensus Sequencing (CCS) reads and Full-Length Non-
517 Concatemer reads were produced with the Iso-Seq pipeline (SMRT Link v10.2).

518 *Genome assembly*

519 We used long-read sequencing data to improve an existing assembly of the *F. borealis*
520 genome produced with short-read data from a single individual[18]. We processed ONT reads
521 with Guppy software v6.4.8 configured for GPU-accelerated basecalling with superior
522 accuracy model (dna_r10.4.1_e8.2_260bps_sup). We assembled ONT reads using Flye
523 v2.9.4, and we merged long- and short-read assemblies together using Quickmerge v0.3.
524 The hybrid assembly was polished with Illumina reads using Pilon v1.24, resulting in a last
525 version with 4063 contigs, for a total size of 79.7 Mb and N50=39.9 kb.

526 To produce an *A. sicula* genome assembly, we used sequencing information obtained from
527 WGA-DNA corresponding to a single individual. ONT reads were basecalled with Dorado
528 V0.8.3, configured with superior accuracy model. We assembled ONT reads with Flye
529 v2.9.4 and polished the assembly with Illumina reads produced from the same individual.
530 The final assembly contains 13.10^3 contigs, for a total size of 65 Mb and N50=6.9 kb.

531 We assembled 75.10^6 paired-end Illumina reads from the *F. haplostoma* samples with
532 Megahit assembler[47]. Heterozygous contigs were removed from the assembly with the
533 Redundans pipeline, using the initial set of Illumina reads for scaffolding and gap-closing
534 steps. The final assembly contains 55.10^3 contigs, with a total size of 120 Mb and N50=2.2
535 kb.

536 *Gene and intron annotations*

537 Gene evidence and intron positions were obtained by aligning Pacbio transcripts to the
538 genome assemblies. Only high-quality Pacbio isoforms were employed, corresponding to
539 clusters produced by Full-Length Non Concatemer (FLNC) reads that belong to the same
540 barcode.

541 For *F. borealis*, we first used Cutadapt v1.16 to remove the splice leader sequence present
542 at the beginning of transcripts[10]. We mapped a collection of 249123 trimmed transcripts on
543 the genome assembly with minimap2, without an alignment guiding option that uses
544 conservative positions flanking canonical splice sites (--splice-flank=no). We produced
545 34034 unique isoforms after collapsing transcripts with Python scripts[48]. After aligning
546 transcripts to the genome, gap intervals were used to determine intron positions. We first
547 collected 28845 introns whose borders could be aligned without ambiguity, corresponding
548 to situations where the last position of the upstream exon and the first position of the
549 downstream exon have different bases. In these non-ambigous introns, we could detect
550 88% of the time an adenine at the penultimate position, consistent with the strong
551 conservation observed in various intron classes of other eukaryotes[4,7]. To precisely map
552 6203 additional introns with ambiguous borders, we applied a conservative set of rules :
553 #1, intron ends must have terminal dinucleotides GT/AG; #2, introns that do not comply to
554 #1 must have terminal dinucleotides AG/AC or AG/AT; #3, introns that do not comply to #2
555 must have a penultimate adenine.

556 To annotate introns in the *F. haplostoma* and the *A. sicula* genomes, we used PacBio's Iso-
557 Seq package to align transcripts over genome assemblies and to collapse redundant
558 isoforms. As described above, we first established intron borders based on splice signals
559 discovered with a set of non-ambiguous introns positions (1507 and 1381 introns for *F.*
560 *haplostoma* and *A. sicula*, respectively). For both genomes, the following rules were
561 applied to map additional introns with ambiguous borders: #1, intron ends must have
562 terminal dinucleotides GT/AG; #2, introns that do not comply to #1 must have terminal
563 dinucleotides GC/AG; #3, introns that do not comply to #2 must begin with either
564 GC/AG/AT/GG and end with either AG/AC/AT; #4, introns that do not comply to #3 must
565 have a penultimate adenine.

566 To annotate introns in the *F. pellucida* genome, we first used BLASTX to find matches
567 against a collection of 30987 proteins predicted from the *F. borealis* isoform collection.
568 Intron borders were found based on gap intervals. We kept 450 non-ambigous cases for
569 which no gaps are present and amino-acid identity is strictly conserved in at least two
570 codons out of five, in each flanking exon.

571 *Homology search*

572 Gene orthologues were annotated with reciprocal blast. First, we used human protein
573 sequences as queries against transcriptomes and genome assemblies of the species of
574 interest. We searched for orthologues in the annotated transcriptome and in the genome
575 assembly of *O. dioica*[43] and for other larvaceans, we employed the genomic resources
576 described in this study. Orthologues from other organisms were searched in the ENSEMBL
577 and ANISEED[49] databases. Top hits were checked for protein length and domain
578 conservation, then used as queries for reciprocal BLAST against the UNIPROT database to
579 confirm annotation.

580 *Expression and purification of U2AF heterodimers*

581 The full-length open-reading frames (ORFs) of U2AF subunits were amplified from the
582 cDNA of *F. borealis* with Taq polymerase, and A-tailed PCR products were cloned in pGEM-
583 T vector for sequencing. We employed a seamless cloning strategy (In-Fusion HD, Takara)
584 to subclone U2AF ORFs in bacterial expression vector. We first replaced positions 100-168
585 of MCS-1 in pRSFDUET-1 with the U2AF2 ORF in frame with the 6xHis tag. Then, we
586 replaced the whole MCS-2 with the U2AF1 ORF, in frame with an N-terminal tag that
587 consists of 30 amino acids of the Twin-Strep-tag, followed by amino acids 2-56 of the B1
588 domain of Streptococcal Protein G[50]. Deletion mutants were engineered with seamless
589 cloning.

590 We transformed tandem expression construct in *E. coli* BL21-AI. Clones were grown to
591 OD=0.8 in terrific broth (TB) and protein expression was induced with 50 μM IPTG, in the
592 presence of 0.4% L-arabinose and 0.1 M $ZnCl_2$ during 16 h at 15°C. Cell pellets were lysed
593 on ice in HHS buffer (HHS: Hepes 20 mM, NaCl 0.5 M, 10% glycerol, 1% Tween-20, pH 7.8)
594 supplemented with 2 mg/mL lysozyme, 1 mM TCEP, 1 mM PMSF and 1x Protease Inhibitor
595 Cocktail (cOmplete, EDTA-free; Roche). Lysates were sonicated on ice three times with 10
596 second pulses, then incubated 30 minutes on ice with DNase I, 2U/mL RNase cocktail
597 (Ambion), and 3% BioLock (IBA). Lysates were centrifuged at 18000 g for 30 minutes at 4°C,
598 and protein complexes assembled with U2AF1 were captured on a Strep-Tactin XT 4 flow
599 gravity flow columns (IBA). Column-bound complexes were washed with 20 column
600 volume (CV) of HHS, 5 CV of HLS buffer (HLS: Hepes 20 mM, NaCl 0.1 M, 5% glycerol, 0.5%
601 Tween-20, pH 7.8), and 10 CV of HLS buffer supplemented with 5 mM biotin. Purified
602 complexes were eluted 6 times with 1 CV of HLS supplemented with 0.1 M biotin. We
603 performed buffer exchange on pooled fractions by centrifugation on 10 kDa MWCO
604 columns (Zeba Spin, Thermo) equilibrated with HLS buffer.

605 *Electrophoretic mobility shift assays*

606 We used seamless cloning to insert the EMSA probe sequence in pUC19 downstream a T7
607 promoter and immediately upstream a SapI restriction site. Constructs were linearized
608 with Sap I and used as templates for in vitro transcription (IVT) reactions (MEGAshortscript
609 T7 transcription kit, Ambion). IVT products were treated with Turbo DNase (Ambion),
610 purified with Phenol/Chloroform extraction and treated with Calf Intestinal Alkaline
611 Phosphatase (Thermo). Dephosphorylated probes were purified on RNA Clean &
612 Concentrator-5 (RNA-CC5, Zymo Research). We end-labelled 100 ng of probe with T4 PNK
613 (NEB) and 10 µCi [γ^{32}P]-ATP at 6000 Ci/mmol, and reaction products were purified on 8%
614 PAGE/TBE-Urea 8M. For EMSA, 100 cpm probe were denatured 2 minutes at 80°C, placed
615 on ice, and incubated 1 h on ice in EMSA buffer (HLS supplemented with 50 ng/µL BSA, 1%
616 RNase inhibitor (Protector, Roche), 0.5 mM TCEP, 0.5 mM EDTA) in the presence of purified
617 U2AF. RNPs were run on 6% PAGE/TBE 0.5x for 2 h at 4°C under 5V/cm. Gels were fixed
618 with a 10% EtOH/5% AcOH solution and dried under vacuum. We recorded signals with
619 phosphor screens and a Typhoon FLA scanner (GE), and measured signal intensity with
620 ImageJ. Dissociation constants were determined with the nlsfit function of the R package
621 easynls, using an equation for one-site model: Kd= [P][R]/[PR] where P and R are the free
622 U2AF and probe, respectively, and PR is the U2AF-probe complex. Our conditions provide
623 at least 5-fold molar excess of U2AF compared to the probe.

624 *RNA-compete assays*

625 We designed a pool of 5633 unique sequences derived from *F. borealis* introns, which
626 include the 20 last nucleotides of the introns, followed by 10 nucleotides of the
627 downstream exon. The T7 promoter sequence followed by an extra guanine residue, was
628 added at the 5' end of the pool. A SapI restriction site was added at the 3' end, followed by
629 a target sequence for PCR priming. After chemical synthesis (Twist Biosciences), we
630 produced IVT templates from oligo pools with PCR reactions and primers specific to intron-
631 flanking sequences. To preserve complexity and to avoid the formation of unspecific
632 products, amplification rounds were performed with at least 60% of the initial pool as
633 template during a maximum of 6 PCR cycles. Template linearization, IVT reactions and 5'
634 phosphate removal was performed as described above. We checked the integrity of RNA
635 pool on 15%PAGE/TBE-Urea 8M.

636 We used seamless cloning to insert the full-length ORF of U2AF1 in pGEX-6P-1, in frame
637 with the N-terminal GST. Constructs were transformed into *E. coli* BL21, selected clones
638 were grown to OD=0.8 in TB and protein expression was induced 16 h at 18°C with 0.4 mM
639 IPTG. Cell lysis and nucleic acid removal was performed as described above, in LY buffer
640 (LYB: 25 mM Hepes, 0.35 M NaCl, 10% glycerol, 0.5% Tween-20, pH 7.5). After 10000 g
641 centrifugation for 30 minutes at 4°C, fusion proteins were captured on Glutathione-

642 sepharose 4B (Cytiva) for 2 h at room temperature, washed with 30 bead volume of LYB
643 and eluted 4 times with 1.5 bead volume of LYB supplemented with 50 mM glutathione. We
644 exchanged the elution buffer to LYB as described above.

645 For RNA-binding assays, 2 picomoles GST-U2AF1 were bound to 2 µL Glutathione High-
646 Capacity Magnetic Agarose Beads (Millipore) for 1 h at room temperature in 0.2 mL of RB
647 buffer (RBB: 25 mM Hepes, 0.5 M NaCl, 5% glycerol, 1 mM EDTA, 1 mM DTT, 2 mM $MgCl_2$,
648 0.5% Tween-20). After three washes with RBB to remove unbound proteins, we added 1 µM
649 of RNA pool to the beads in a final volume of 0.5 mL RBB, together with 10 µg BSA, 4 µg
650 Heparin, and 2 µL of RNase inhibitor. Samples were incubated 2 h at 4°C on wheel, and
651 beads were washed three times with 0.5 mL ice-cold RBB. Proteins were digested with
652 Proteinase K in 10 mM Tris, 1% SDS, 2 mM EDTA for 15 minutes at 37°C, and we recovered
653 RNA with RNA-CC5. We prepared Illumina libraries using NEBNext Small RNA library prep
654 kit (NEB), and samples were sequenced on a MiSeq instrument at the NSC. After adapter
655 trimming, an average of 84% reads could be mapped on the original collection of introns
656 with Bowtie2 (--very-sensitive-local).

657 We analyzed an average of 945.10^3 reads per library. We filtered out introns with low read
658 count (no sample with >10 reads). Next, read counts were normalized against the total
659 number of reads per sample and for each intron, we measured enrichment using:
660 E =log[(C1+2)/(C2+2)], where C1 and C2 correspond to read counts measured for different
661 U2AF1 paralogues. We first selected introns with E>0.12 when comparing differential
662 enrichment between either *F. borealis* U2AF1α and U2AF1β, or *O. dioica* U2AF1 and F.
663 borealis U2AF1β. We scored introns with their read count, and we selected a group of top-
664 scoring introns representing 80% of the total read count kept for sequence analysis.
665 Composition statistics and motif detection were performed on RNA pool sequences
666 clipped from their terminal dinucleotides.

667 *Mammalian cell expression and transfection*

668 We used seamless cloning to introduce *F. borealis* U2AF2 paralogues in pCDNA3.1, in N-
669 terminal fusion with a 6xHis tag. U2AF1 paralogues were introduced in pSF-CMV, in fusion
670 either with a N-terminal GST, or with the Twin-Strep tag and solubilization domain as
671 described above. Constructs were delivered to HEK293T cell grown in DMEM using
672 Lipofectamine 3000 (Invitrogen), and cultures were harvested 48 h post-transfection.

673 *Long-read sequencing of cDNA*

674 Cells transfected with U2AF1 and U2AF2 constructs were treated with 0.15 M
675 cycloheximide for 8 h prior harvest. Controls were treated with 0.1% DMSO. Cells pellets
676 were washed in PBS, and RNA was extracted with TRIzol reagent (Invitrogen), following the

677 manufacturer's procedure. Samples were treated with Turbo DNase, and 1 µg RNA was
678 converted to cDNA with Superscript IV reverse transcriptase, using a combination of
679 random 10-mers and oligo-dT primers. For each target gene, we amplified cDNA with
680 specific primers in separate PCR reactions, for a total of 20 cycles. PCR products from the
681 same cDNA were pooled together and purified with AMPure XP SPRI beads (Beckman). We
682 prepared libraries with barcodes to separate cDNA samples, following ONT's procedure
683 for barcoding and sequencing amplicons (SQK-LSK114). We sequenced 55 ng of library
684 pool on a PromethION 2 solo flow cell (R10.4.1) for 2 h, producing 6.10^6 reads with N50=3.4
685 kb. Reads over 500 bp with quality score over Q9 were basecalled with Dorado V0.8.3,
686 configured with high accuracy model. We trimmed adapters and primers with porechop,
687 and processed the reads with TALON v6.0[51], with a database corresponding to regions of
688 the human genome targeted by the gene-specific primers. For each sample, we
689 considered only isoforms whose sequencing coverage across samples is superior to 50,
690 and we found splicing events based on exons positions mapped by TALON. Introns were
691 classed as canonical (ending with GT/AG) or non-canonical (other combinations of
692 terminal dinucleotides). Isoforms were classed as canonical unless their splicing involved
693 at least one non-canonical splice site.

694 For each isoform of each target gene, we scaled expression across samples with:
695 E= $(C - Mc)/Sc$ where E is the scaled isoform expression value in the sample, C is the
696 isoform read count in the sample, Mc is the mean of the isoform read count across
697 samples and Sc is the standard deviation of the isoform read count across samples. For
698 each sample, we collected intron positions corresponding to the splicing of isoforms with
699 E > 0.75, and we established a list of specific positions absent from other samples. We
700 analyzed sequence composition using a non-redundant collection of introns. The PPT
701 score of each intron was calculated with the rules established by Clark & Thanaraj[52].

702 *Transcriptome analysis of U2AF2-expressing cells*

703 We used Trizol to extract total RNA from HEK293T cells transfected with either U2AF2
704 constructs described above, or with the pCDNA3.1 vector containing the env gene of a LTR
705 retrotransposon[19]. We enriched mRNA with the Dynabeads mRNA Purification kit (Ambion)
706 and barcoded sequencing libraries were prepared from poly-A^+ RNA with NEBNext Ultra II
707 RNA Library Prep kit (NEB). Pooled libraries were sequenced 150bp PE on a 1.5B
708 NovaSeqXPlus flow cell at the NSC.

709 Basic trimming of raw reads was performed with FastP v0.24.0 to remove adapter
710 sequences and keep Q20 reads[53]. The data was mapped with STAR v2.5.2b[54] to *H. sapiens*
711 GRCh38.p14 primary genome assembly using the latest gene annotation file retrieved from
712 GENCODE[55]. The mapping results were used to run an additional STAR two-pass mapping

713 for further analysis of AS events. We used Salmon v1.10.3[56] to quantify gene expression
714 and we conducted Differential Gene Expression (DGE) analysis with R packages tximport
715 v1.32.0 and DeSeq2 v1.44.0[57,58]. Differentially expressed genes (DEGs) were filtered with
716 significance threshold of -0.58 ≤ Log2FoldChange value ≥ 0.58 and adjusted p value padj <
717 0.05. To examine enriched functional categories and pathways in DEGs, we performed GO
718 and KEGG Over-representation analysis (ORA) using the clusterProfiler R package
719 v4.12.6[59]. We determined AS events with two-passed STAR genome aligned bam files using
720 the rMATS turbo software v4.3.0[60] with --novelSS function enabled to allow the detection of
721 novel splice sites. The AS events detected were filtered using a significance threshold of
722 FDR ≤ 0.05.

723 *Identification of U2AF2 partners*

724 Cells transfected with either U2AF constructs or a control containing the *env* gene were
725 washed in PBS and lysed in HLS, supplemented with 1 mM TCEP, 1 mM PMSF, and
726 protease inhibitors (cOmplete, Roche). Protein concentration was determined with
727 bicinchoninic acid assay, and for each transfection, a sample equivalent to 6 µg protein
728 was used for immunoprecipitation. Samples diluted in 1 mL HLS were first pre-cleared by
729 incubation with Anti-HA magnetic beads (Sigma, equivalent to 70 µL suspension) for 1 h at
730 4°C on wheel. The supernatant was transferred to Anti-V5 magnetic beads (Sigma,
731 equivalent to 70 µL suspension) and incubated 2 h at 4°C on wheel. Beads were washed
732 five times with 1.5 mL ice-cold HLS and resuspended in Laemmli buffer. Bound proteins
733 were eluted for 5 min at 70°C and separated on Bis-Tris SDS-8% PAGE. After Coomassie
734 staining, bands between 37 and 100 kDa were excised. Peptides were digested in-gel and
735 analyzed on an Orbitrap Eclipse Mass Spectrometer at the Proteomics Unit of the
736 University of Bergen. For validation of protein interactions, samples separated on gel were
737 transferred to a PVDF membrane and tested with either SAHH Antibody (sc-271389, Santa
738 Cruz Biotechnology) or AHCYL1/SAHH-3 Antibody (sc-271581, Santa Cruz Biotechnology).

739 *Generation of AHCYL1 KO cells*

740 We transfected HEK293T cells with a set of CRISPR/Cas9 KO plasmids directed against the
741 human AHCYL1 gene (sc-417109, Santa Cruz Biotechnology). Three days post-
742 transfection, cells were harvested and washed in PBS supplemented with 2% BSA and 1
743 mM EDTA. Fluorescent cells expressing GFP and the Cas9 gene were sorted on a Wolf G2
744 benchtop cell sorter. Gating was set by examining the fluorescence of cell cultures
745 transfected with pEGFP-N1 or with a LacZ-expressing construct. After sorting, GFP-positive
746 cells and GFP-negative were pelleted, resuspended in complete culture media and grown
747 four days at 37°C before assaying AHCYL1 expression and m6A levels.

748 *Detection of N6-methyladenosine in snRNA*

749 Total RNA was extracted with Trizol, following the manufacturers' recommendations. To
750 examine m6A levels, total RNA was separated on 10% PAGE-Urea and transferred on a
751 positively charged nylon membrane (Hybond N^+, Amersham). We detected N6-
752 methyladenosine using a m6A antibody (ab151230, Abcam). To quantify total snRNA
753 amounts, we performed Northern Blot using a pool of radioactive, end-labeled
754 oligoribonucleotides probes complementary to U1, U2, U4, U5 and U6. Membranes were
755 hybridized to probes in ULTRAhyb-Oligo (Ambion) during 14 h at 42°C, then washed twice
756 at 42°C with 2x SSC / 0.5 % SDS. Signals were recorded and quantified as described above.
757 When measuring cellular m6A levels, each RNA sample was tested with m6A antibody and
758 ran in parallel for Northern Blot. The m6A signal was quantified with Image Lab (Bio-Rad)
759 and normalized against the intensity of the corresponding radioactive signal.

760 *Mapping of m6A in U1 snRNA*

761 We designed primer extension assays based on procedures described by Hong et al.[32].
762 Briefly, 1.5 µg total RNA was annealed to 150000 cpm of gel-purified DNA primer
763 complementary to nucleotides 13-33 of U1 snRNA and 5' end-labelled with [γ^{32}P]-ATP.
764 After 2 minutes at 90°C, hybridization mix were placed 2 minutes at 50°C, and 10 U
765 Superscript IV reverse transcriptase (Invitrogen) were added in RT buffer (50 mM Tris-HCl, 4
766 mM $MgCl_2$, 10 mM DTT, 50 mM KCl, 30 µM dATP, 30 µM dGTP, 30 µM dCTP, pH 8.3) in the
767 presence of 30 µM of either dTTP or 4SedTTP (Jena Biosciences). After 20 minutes
768 incubation at 50°C, cDNA synthesis was stopped by addition of Urea gel loading buffer,
769 and denatured elongation products were resolved on 20% PAGE-Urea sequencing gels.
770 Signal was recorded on fixed gels and processed as described above.

771

772

773

774

775

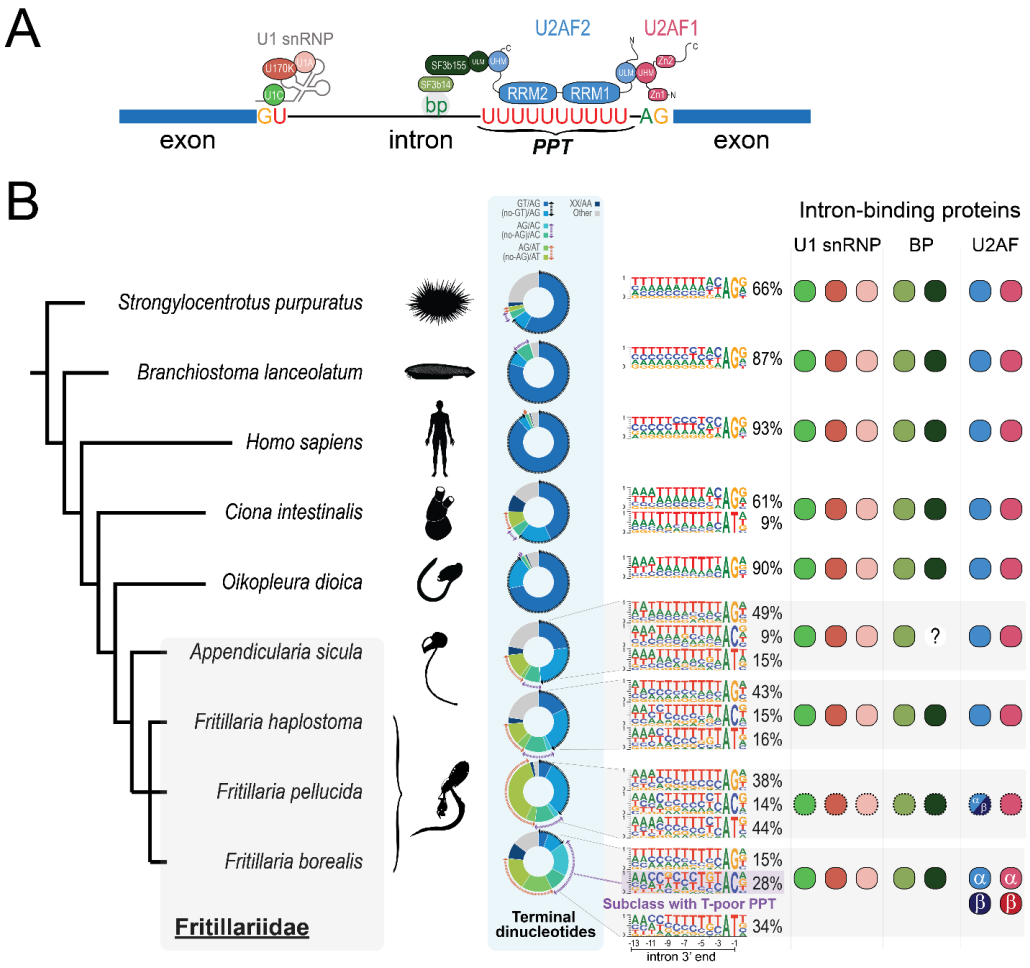
776

777

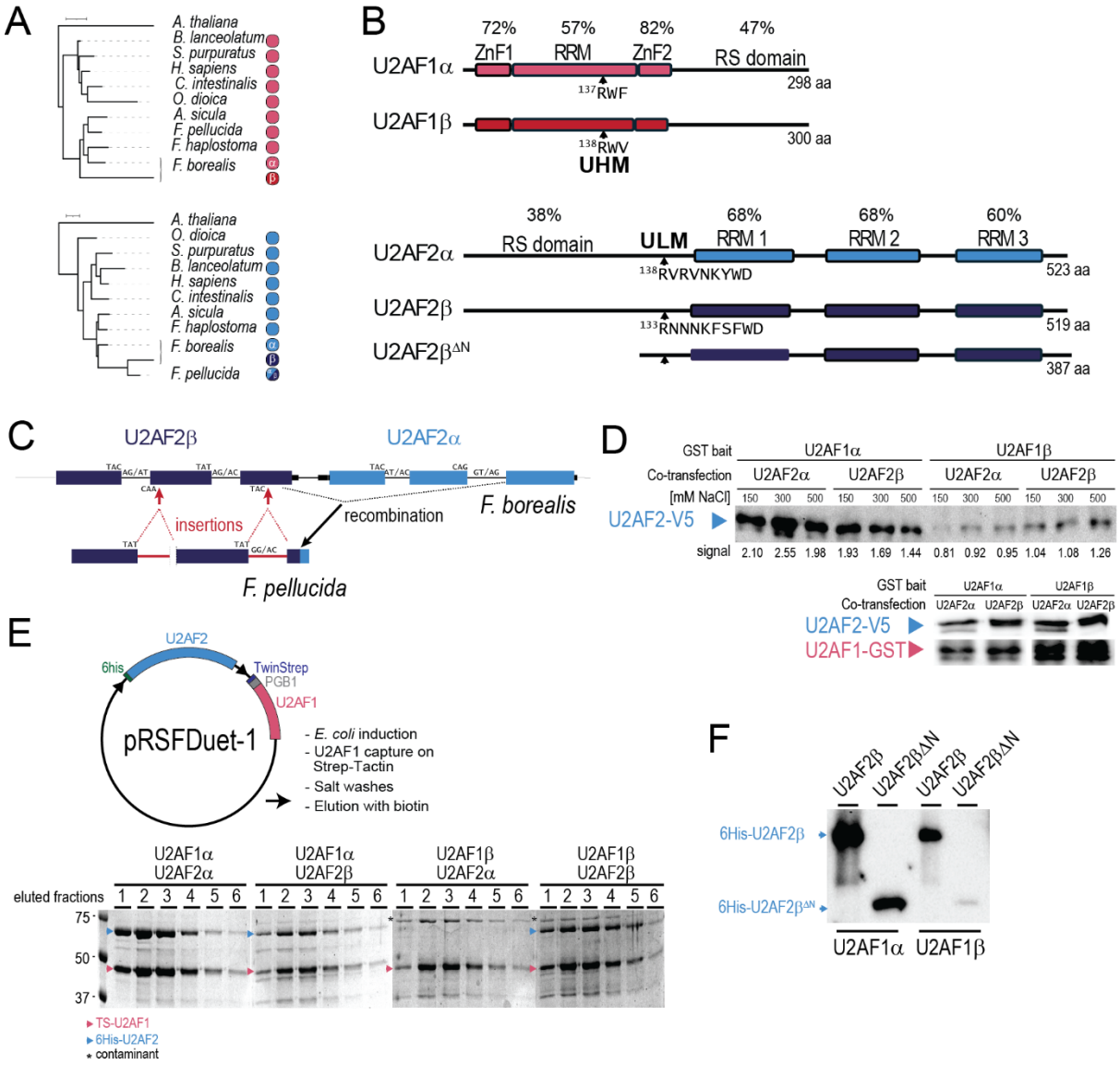
778

779

Figures


A
U1 snRNP
U2AF2
U2AF1
SF3b155
SF3b14
bp
RRM2
RRM1
GU
UUUUUUUUUU
AG
exon
intron
PPT
exon
B
Intron-binding proteins
U1 snRNP
BP
U2AF
Strongylocentrotus purpuratus
Branchiostoma lanceolatum
Homo sapiens
Ciona intestinalis
Oikopleura dioica
Appendicularia sicula
Fritillaria haplostoma
Fritillaria pellucida
Fritillaria borealis
Fritillariidae
66%
87%
93%
61%
9%
90%
49%
9%
15%
43%
15%
16%
38%
14%
44%
15%
28%
34%
Subclass with T-poor PPT
Terminal dinucleotides
intron 3' end

<u>Figure 1:</u> Phylogenomics of the 3'ss. **A)** Molecules involved in splice site recognition. In humans, the 5' end of the intron is recognized by the U1 snRNP, a complex formed with snRNA, Sm proteins (not shown) and specific proteins (U1A, U1C, U1-70K). The branchpoint nucleotide (bp) is recognized by the SF3b complex (formed with subunits SF3b14 and SF3b155). The U2AF complex recognizes the 3' end of the intron. It includes the small subunit U2AF1, that binds the 3'ss, and the large subunit U2AF2 that interacts with the PPT. Protein-protein interaction modules mediate interaction between U2AF subunits, and between U2AF and SF3b. UHM, U2AF Homology Motif; ULM, U2AF Ligand Motif. **B)** Evolution of intron classes. Species were placed on the tree based on published phylogenies[18]. The central panel shows the frequency of intron classes in different genomes, defined with their terminal dinucleotides. Statistics for *S. purpuratus*, *B. lanceolatum*, *H. sapiens* and *C. intestinalis* introns were retrieved from Minor Intron Database[61]. Logos show nucleotide composition at the 3' end of main intron classes, along with their frequency. The right-hand panel represents gene copies identified for RNPs involved in splice site recognition. Each copy is represented by a dot, full border shows evidence gained with transcripts and genomic DNA, dashed border shows evidence from genomic DNA only. We could not find a SF3b155 gene in *A. sicula*, possibly due to incomplete genomes and transcriptomes. Greek alphabet letters represent *F. borealis* U2AF paralogues, and hybrid U2AF2 found in *F. pellucida*.



823

824

825

826

827

828

829

830

831

832 **<u>Figure 2:</u>** The *F. borealis* U2AF proteins. **A)** Maximum-Likelihood phylogeny of U2AF1(top)
833 and U2AF2 (bottom) orthologues. Scales correspond to 0.1 substitution per site. **B)**
834 Domain organization and protein-protein interaction modules in *F. borealis* paralogues.
835 Percentage identity between paralogue domains is shown. **C)** Duplicates of U2AF2 form a
836 gene cluster in *F. borealis*. The *F. pellucida* U2AF2 gene has features of U2AF2α and
837 U2AF2β, showing a secondary reduction of the cluster by recombination between
838 duplicates. Rearrangements in *F. pellucida* also include loss and gain of introns, and a
839 reduction of the N-terminal RS domain. **D)** GST-pulldown assays in HEK293T cells. U2AF1-
840 GST baits were co-expressed with U2AF2 tagged with a V5 epitopes. Left panel, after
841 capturing U2AF1 on GSH-sepharose beads, associated U2AF2 was detected with an
842 antibody against V5 (blue arrow). GST capture and washes were performed under
843 increasing salt concentration. Right panel, expression of transfected U2AF1 (pink arrow)
844 and U2AF2 (blue arrow) in GST-pulldown inputs. **E)** Purification of U2AF heterodimers. Left
845 panel, representation of the construct used for tandem expression of U2AF subunits in *E.*
846 *coli* and outline of the purification procedure. Right panel, Coomassie-stained SDS-PAGE
847 showing U2AF subunits eluted from Strep-Tactin resin. Dimer formation failed only for the
848 U2AF1β/ U2AF2α combination. **F)** U2AF1β of the RS domain abolishes dimer formation
849 between U2AF1β and U2AF2β. Proteins eluted from Strep-Tactin resin were detected with
850 an antibody against 6his, after testing dimer formation between U2AF1 paralogues and
851 versions of U2AF1β.

852

853

854

855

856

857

858

859

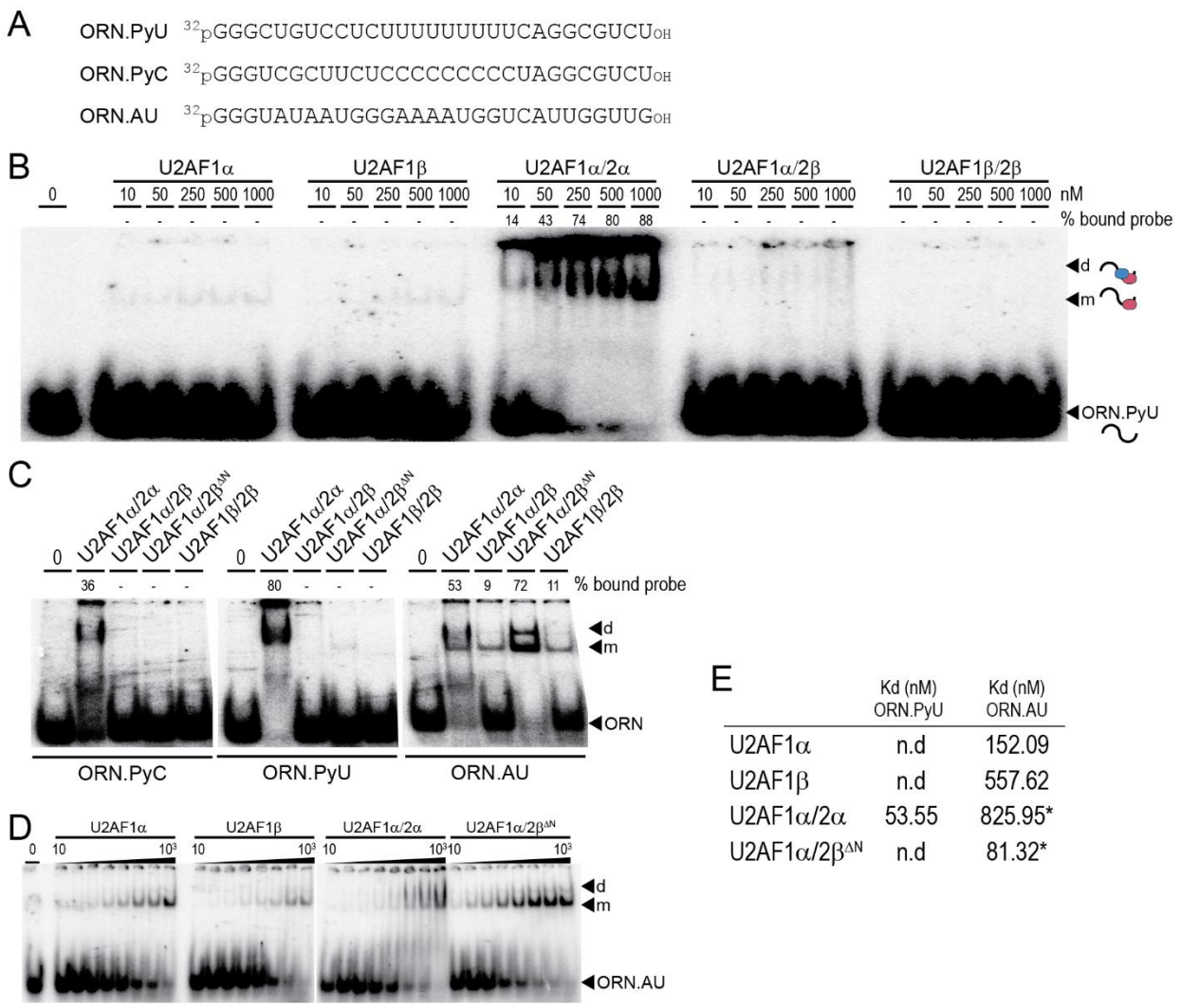


Figure 3: U2AF binding on RNA tested with EMSA. **A)** Representation of RNA probes used for EMSA. **B)** EMSA with ORN.PyU and purified U2AF. From left to right, experimental series were performed with U2AF1 only, or with different U2AF dimers. Bound probe measurement is reported only for values superior to 5%. Drawings on the right represent, from bottom to top: unbound probe, probe bound to one U2AF subunit and probe bound to two U2AF subunits. **C)** Comparison of U2AF binding on different probes. For each assay, 330 nM U2AF dimer was used. **D)** EMSA with ORN.AU. **E)** U2AF dissociation constants determined for ORN.PyU and ORN.AU. Values marked with a star correspond to the average of two titrations.

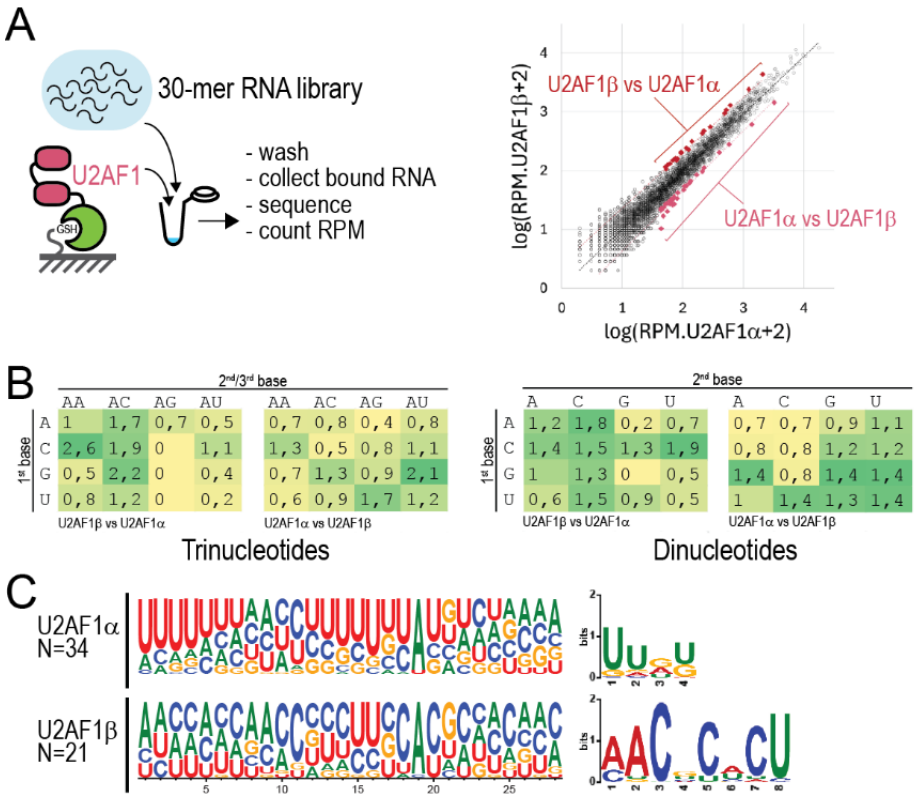
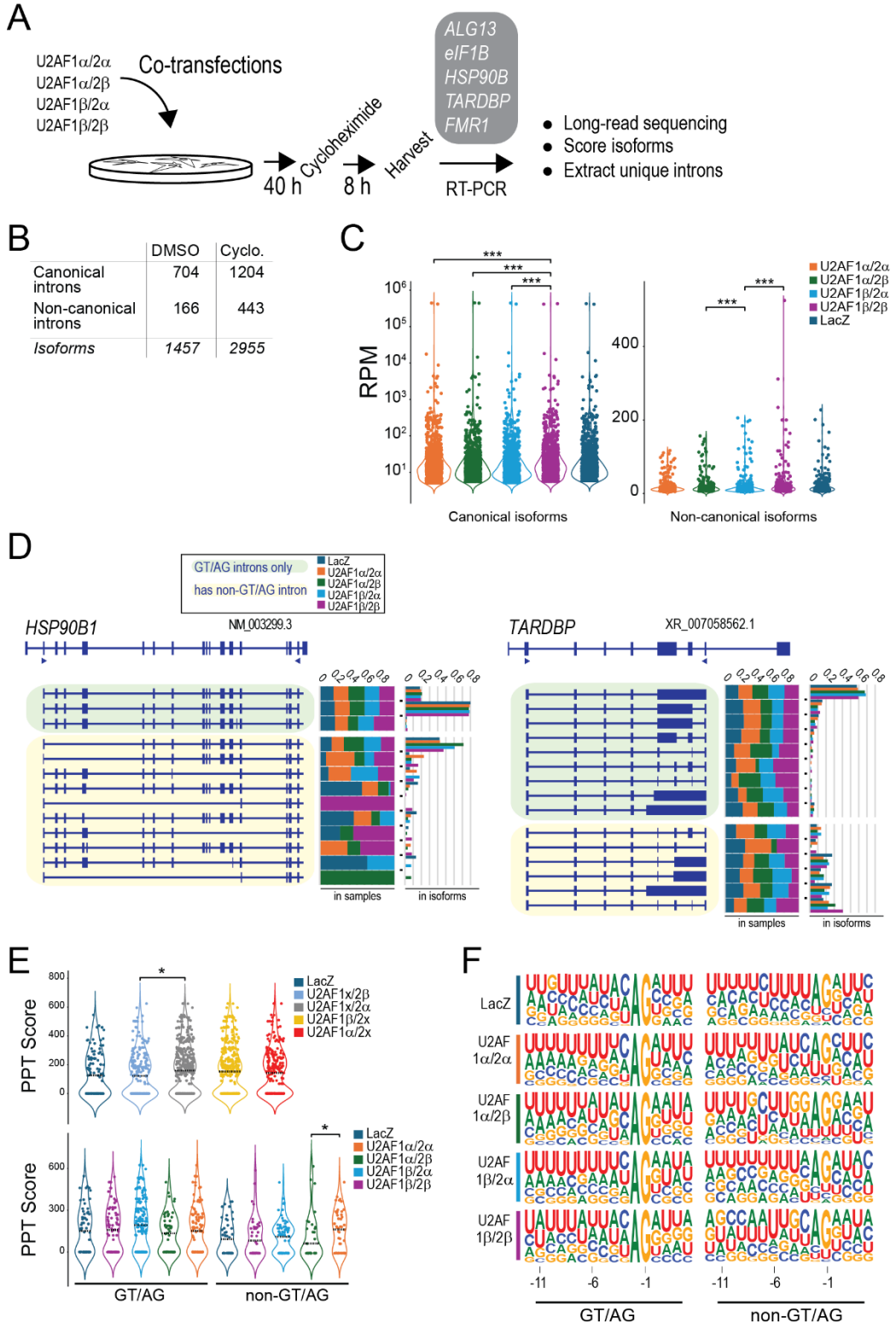


Figure 4: RNA-compete experiments with U2AF1. **A)** On the left, schematic drawing of approach and on the right, read count plot showing differential enrichment of library oligos in RNA recovered with U2AF1α (x-axis) and with U2AF1β (y-axis). Colored diamonds represent groups of high affinity oligos selected for sequence analysis. **B)** Abundance of 3- and 2-mer motifs in the sequence of high affinity oligos. Values correspond to the average frequency in collected oligos relative to the average frequency measured in the first library. **C)** High affinity oligos collected with U2AF1α or U2AF1β were used to establish sequence logos (left side) and compared pairwise with STREME to reveal over-represented motifs (right panel).


A
U2AF1α/2α
U2AF1α/2β
U2AF1β/2α
U2AF1β/2β
Co-transfections
40 h
Cycloheximide
8 h
Harvest
ALG13
eIF1B
HSP90B
TARDBP
FMR1
RT-PCR
Long-read sequencing
Score isoforms
Extract unique introns
B
DMSO
Cyclo.
Canonical introns 704 1204
Non-canonical introns 166 443
Isoforms 1457 2955
C
RPM
Canonical isoforms
Non-canonical isoforms
LacZ
D
GT/AG introns only
has non-GT/AG intron
HSP90B1
NM_003299.3
TARDBP
XR_007058562.1
in samples
in isoforms
E
PPT Score
GT/AG
non-GT/AG
F
LacZ
U2AF 1α/2α
U2AF 1α/2β
U2AF 1β/2α
U2AF 1β/2β
GT/AG
non-GT/AG

890

891

892 **<u>Figure 5:</u>** Targeted cDNA sequencing reveals different impact of U2AF paralogues on
893 splicing. **A)** Experimental approach, showing co-transfections of U2AF subunits, NMD
894 inhibition with cycloheximide, RT-PCR targets and isoform analysis. **B)** Table showing the
895 total number of introns and isoforms discovered in the cDNA of either untreated cells, or
896 cells treated with cycloheximide. **C)** Scatter plots showing read count of isoforms
897 discovered in treated cells that are produced only by canonical splicing events (left), and
898 for isoforms produced by the splicing of at least one non-canonical intron (right). Wilcoxon
899 signed-rank test values are shown with stars (*, >0.05; **, >0.01; ***, >0.005). **D)** Isoform
900 diversity in two target genes. For each gene, we represent the reference isoform on top,
901 and the best expressed isoforms in treated cells, classified as canonical or non-canonical
902 splicing isoforms. Arrowheads show primers used for RT-PCR. Isoform proportion is
903 expressed either as frequency in samples, or frequency among the best expressed
904 isoforms. **E)** Scoring PPT strength. The upper plot displays PPT scores of all introns spliced
905 in treated cells, for control sample expressing LacZ, and for the two co-transfections made
906 with either U2AF2β, U2AF2α, U2AF1β or U2AF1α. The lower plot displays PPT scores for
907 distinct U2AF combinations. Dashed lines show the dataset median. **F)** Sequence logos
908 produced with the 3' end sequence of introns obtained from sample-specific splicing
909 events.

910

911

912

913

914

915

916

917

918

919

920

921

922

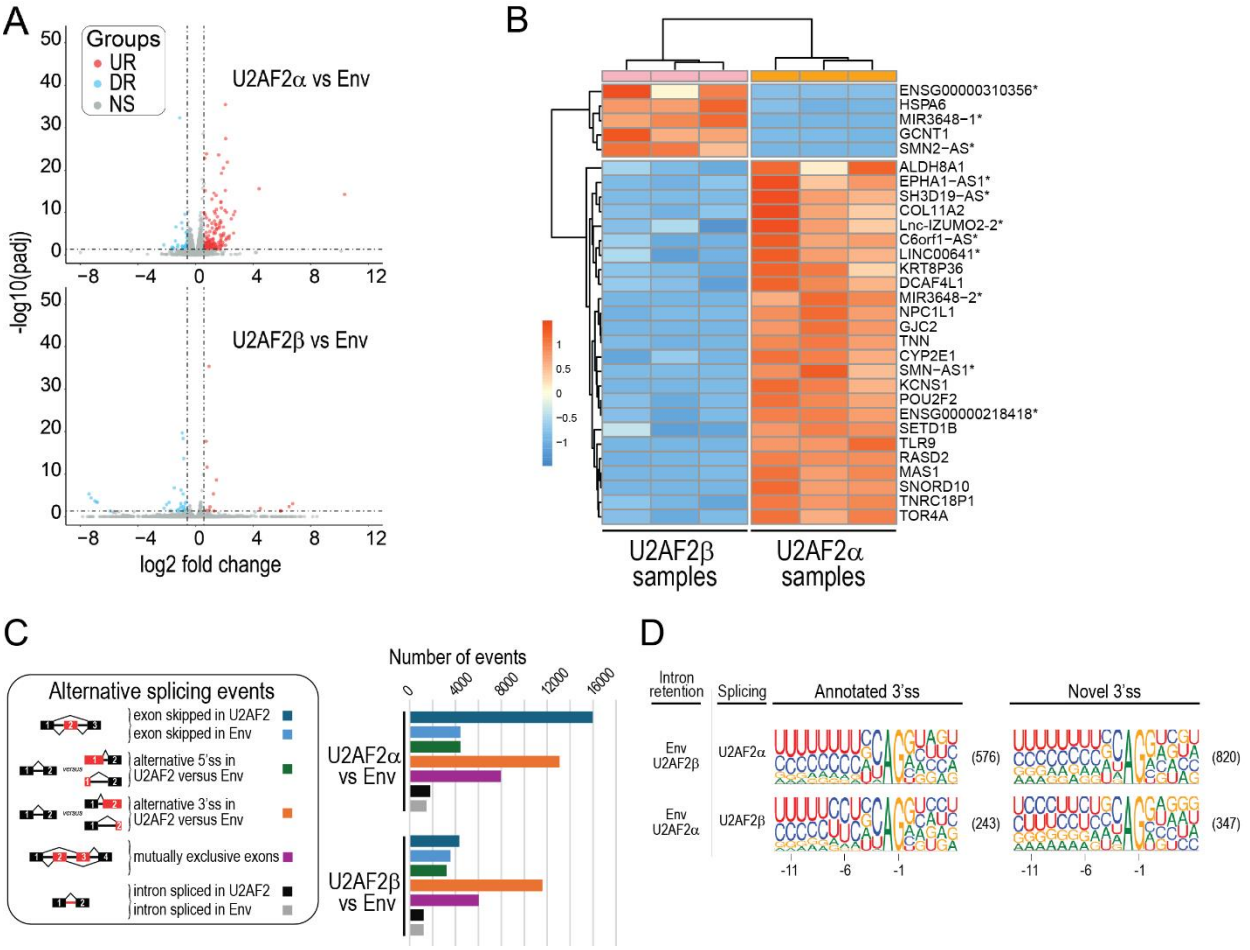


Figure 6: Transcriptome analysis of cells expressing *F. borealis* U2AF2. **A)** Volcano plots showing DEGs found by comparing the transcriptome of cells transfected with either U2AF2 paralogues, or control cells expressing an envelope protein. UR, upregulated genes; DR, down-regulated genes; NS, non-significant change. **B)** Top 30 DEGs between cells transfected with U2AF2α and cells transfected with U2AF2β. Stars show non-protein coding genes. **C)** AS events detected with rMATS, classified by type of events. The events include annotated and novel splice sites. **D)** Expression of U2AF2 paralogues is associated with the processing of distinct 3'ss. Sequence logos were produced with 3'ss from introns spliced specifically in cells transfected with U2AF2α or U2AF2β.

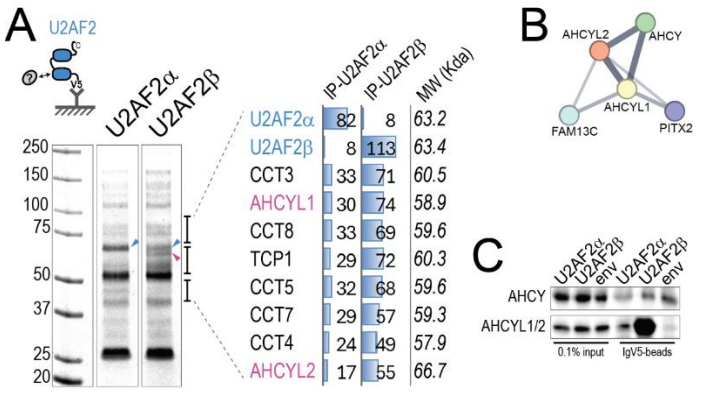


Figure 7: The U2AF2β-AHCYL1 interaction. **A)** Co-immunoprecipitation experiment. We expressed U2AF2 paralogues with a V5 tag in mammalian cells, and after immobilizing protein complexes on antibody-coupled beads, we examined the associated proteins on Coomassie-stained gel. Brackets show the areas analyzed with mass spectrometry. On the right-hand side, the most abundant proteins found in the analysis are listed with the number of corresponding peptides detected, and their predicted size. Blue and Pink arrows show bands corresponding to U2AF2 baits and AHCYL proteins, respectively. Note that AHCYL1 and AHCYL2 share 91.8% identity over 80% of their protein sequence. **B)** AHCYL protein interaction network retrieved from the STRING database[62]. Lines show physical interactions, with thickness proportional to data support. **C)** Validation of the interaction between U2AF2 and AHCY-like proteins. The western blot shows the result of a co-immunoprecipitation experiment using U2AF2-V5 as bait. Lanes 1-3: cell extracts expressing tagged baits, either U2AF2 paralogues, or *env* protein control. Lanes 4-6: material bound on beads after immunoprecipitation. Samples were tested with an AHCY antibody (top) or an antibody recognizing both AHCYL1 and AHCYL2 (bottom).

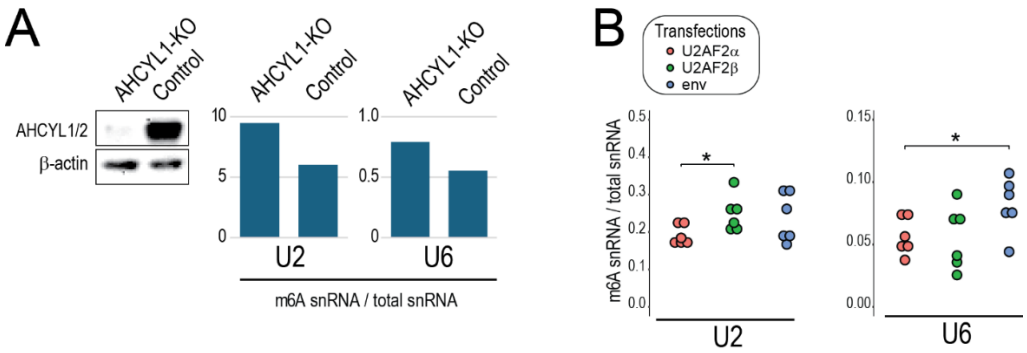


Figure 8: U2AF2 and AHCYL1 reduce m6A levels in snRNA. **A)** Impact of AHCYL1 knock-out. Western blot shows loss of AHCYL1 expression after inactivating the *ahcyl1* gene in HEK293T cells. Graphs show m6A levels measured in U2 or U6 snRNA, compared to the total snRNA amount. **B)** Impact of U2AF2 paralogues. The levels of m6A in snRNA were measured as in **A)**, in total RNA from cells transfected with *F. borealis* U2AF2 or env.

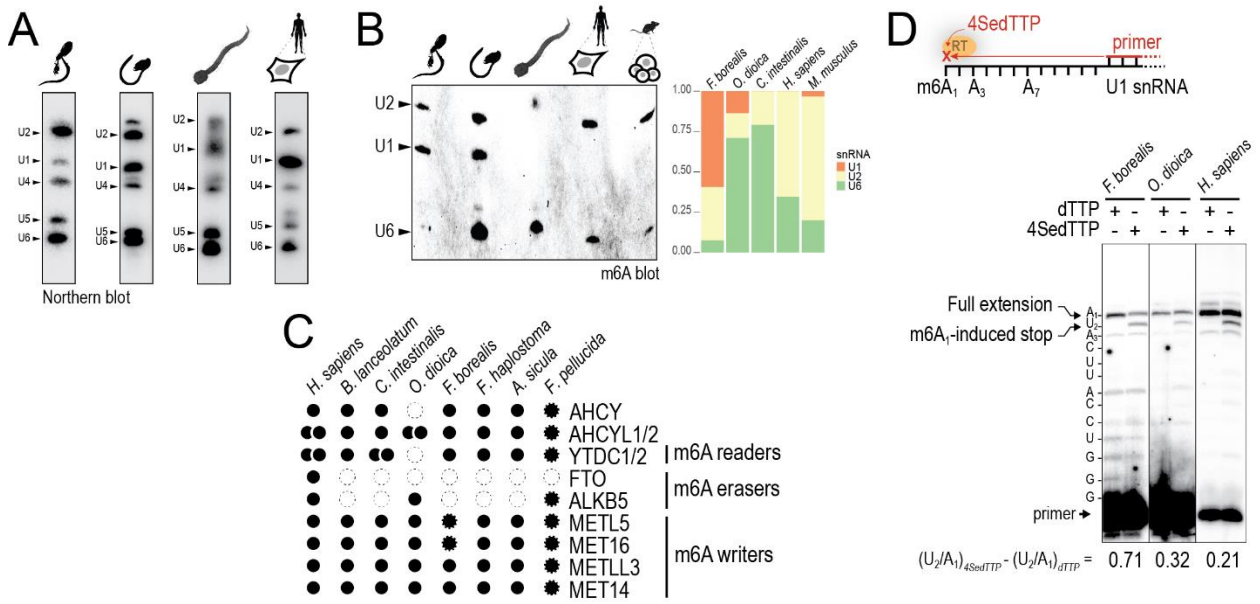


Figure 9: Divergent m6A patterns in *F. borealis* snRNA. **A)** Conservation of U2-type snRNAs. Northern blots show snRNA signals in total RNA extracted from (panels from left to right): maturing adults of *F. borealis* and *O. dioica*, *C. intestinalis* larvae and HEK293T cells. **B)** Detection of m6A. The immunoblot shows m6A signals in total RNA from the same origin as in **A)**. The last lane corresponds to total RNA from mouse ES cells. Bands were identified by comparing the migration profile with Northern blot experiments. The graph represents the proportion of signal intensity for U1, U2 and U6 snRNA. **C)** Conservation of m6A regulatory genes. We examined for each organism the number of copies of different genes involved in the SAM metabolism and its regulation (AHCY, AHCYL1/2), or in the deposition of m6A. Copies are represented by dots, full border indicates evidence gained with transcripts and genomic DNA, dashed border indicates evidence from genomic DNA only. Empty dots represent gene losses. **D)** Mapping m6A modification in the 5'end of U1 snRNA. Top, representation of the experimental approach showing that the presence of m6A on the RNA template induces reverse transcription (RT) stop when the thymidine analogue 4SedTTP is incorporated. Bottom, result of a primer extension experiment conducted on total RNA from three distinct species. We measured for each condition the ratio of cDNA stopped at position U_2 of U1 snRNA relative to full-length cDNA. Values at the bottom of lanes correspond to the ratio measured in the presence of 4SedTTP minus the ratio measured with dTTP.

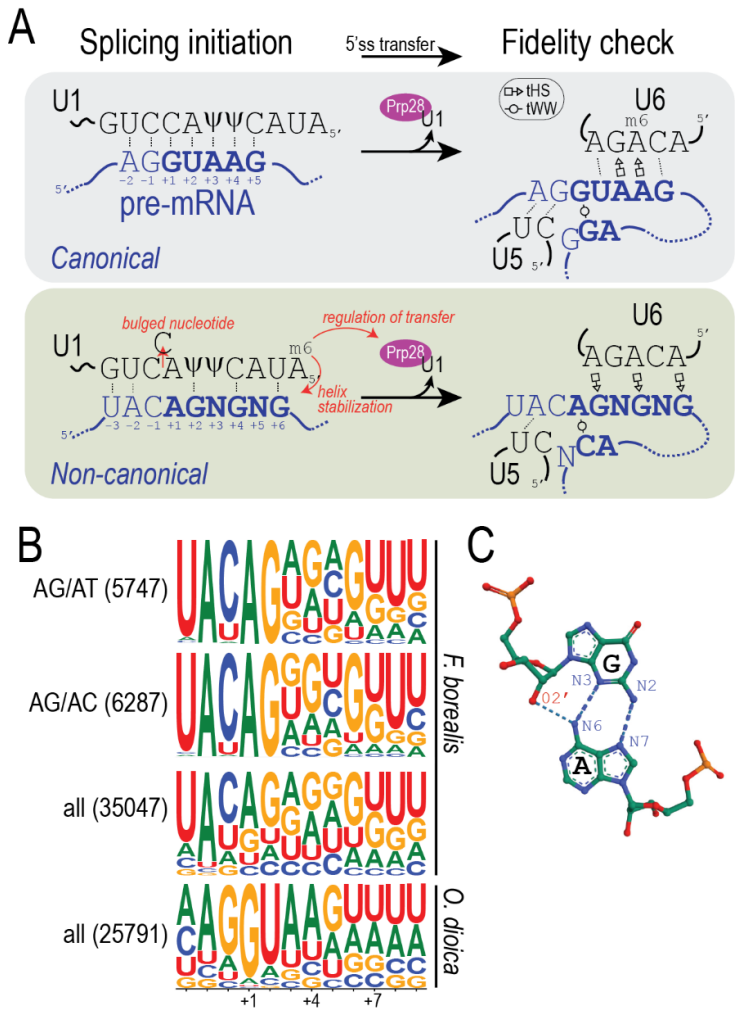


Figure 10: Function of m6A modification during non-canonical splicing. **A)** The RNA network of 5'ss recognition. The top panel is a model based on Artemyeva-Isman & Porter[7] representing base-pair interactions established between the 5'ss (intronic residues shown in bold) and U1 snRNA during the initiation of splicing, when the U1 snRNP scans for exon/intron boundaries; and between the 5'ss and U6 snRNA after the pre-mRNA is transferred to the tri.snRNP. The assembly of an RNA network involving U5 and U6 snRNA on one hand, and intron termini on the other hand, constitutes a fidelity check that ensures splicing accuracy. Dashed lines represent Watson-Crick interactions, symbols represent non-Watson-Crick pairs. The bottom panel proposes a model of interactions established on a non-canonical 5'ss representing the major classes of *F. borealis* introns. Here, the binding of U1 snRNA may involve shifted register and stabilization by m6A$_1$. **B)** Sequence logos of 5'ss showing overrepresentation of guanidine in *F. borealis* introns. **C)** Example of tHS interaction between the sugar edge of G_{2657} and the Hoogsteen edge of A_{2664} in the Sarcin/Ricin Domain from E. coli 23 S rRNA (PDB 3DVZ)[63].

1014

1015 **References**

1016 **1/**Suzuki, H *et al.* Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH
1017 medulloblastoma. *Nature*, doi:10.1038/s41586-019-1650-0 (2019).
1018 **2/**Glasser, E *et al.* Pre-mRNA splicing factor U2AF2 recognizes distinct conformations of
1019 nucleotide variants at the center of the pre-mRNA splice site signal. *Nucleic Acids Res*,
1020 doi:10.1093/nar/gkac287 (2022).
1021 **3/**Yoshida, H *et al.* A novel 3' splice site recognition by the two zinc fingers in the U2AF
1022 small subunit. *Genes Dev*, doi:10.1101/gad.267104.115 (2015).
1023 **4/**Olthof, AM *et al.* Taxonomy of introns and the evolution of minor introns. *Nucleic Acids*
1024 *Res*, doi:10.1093/nar/gkae550 (2024).
1025 **5/**Erkelenz, S *et al.* Ranking noncanonical 5' splice site usage by genome-wide RNA-seq
1026 analysis and splicing reporter assays. *Genome Res*, doi:10.1101/gr.235861.118 (2018).
1027 **6/**Parker, R & Siliciano, PG. Evidence for an essential non-Watson-Crick interaction
1028 between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature*,
1029 doi:10.1038/361660a0 (1993).
1030 **7/**Artemyeva-Isman, OV & Porter, ACG. U5 snRNA Interactions With Exons Ensure Splicing
1031 Precision. *Front Genet*, doi:10.3389/fgene.2021.676971 (2021).
1032 **8/**Bai, R *et al.* Structure of the Post-catalytic Spliceosome from Saccharomyces cerevisiae.
1033 *Cell*, doi:10.1016/j.cell.2017.10.038 (2017).
1034 **9/**Costa, M *et al.* Crystal structures of a group II intron lariat primed for reverse splicing.
1035 *Science*, doi:10.1126/science.aaf9258 (2016).
1036 **10/**Henriet, S *et al.* Evolution of the U2 Spliceosome for Processing Numerous and Highly
1037 Diverse Non-canonical Introns in the Chordate Fritillaria borealis. *Curr Biol*,
1038 doi:10.1016/j.cub.2019.07.092 (2019).
1039 **11/**Huff, JT *et al.* Mechanism for DNA transposons to generate introns on genomic scales.
1040 *Nature*, doi:10.1038/nature20110 (2016).
1041 **12/**Farhat, S *et al.* Rapid protein evolution, organellar reductions, and invasive intronic
1042 elements in the marine aerobic parasite dinoflagellate Amoebophrya spp. *BMC Biol*,
1043 doi:10.1186/s12915-020-00927-9 (2021).
1044 **13/**Guminska, N *et al.* Order of removal of conventional and nonconventional introns from
1045 nuclear transcripts of Euglena gracilis. *PLoS Genet*, doi:10.1371/journal.pgen.1007761
1046 (2018).
1047 **14/**Marz, M *et al.* Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol*,
1048 doi:10.1007/s00239-008-9149-6 (2008).
1049 **15/**Henriet, S *et al.* Laboratory study of Fritillaria lifecycle reveals key morphogenetic
1050 events leading to genus-specific anatomy. *Front Zool*, doi:10.1186/s12983-022-00471-y
1051 (2022).
1052 **16/**Sales-Lee, J *et al.* Coupling of spliceosome complexity to intron diversity. *Curr Biol*,
1053 doi:10.1016/j.cub.2021.09.004 (2021).
1054 **17/**Stark, MR *et al.* Dramatically reduced spliceosome in Cyanidioschyzon merolae. *Proc*
1055 *Natl Acad Sci U S A*, doi:10.1073/pnas.1416879112 (2015).

1056 **18/**Naville, M *et al.* Massive Changes of Genome Size Driven by Expansions of Non-
1057 autonomous Transposable Elements. *Curr Biol*, doi:10.1016/j.cub.2019.01.080 (2019).
1058 **19/**Henriet, S *et al.* Embryonic expression of endogenous retroviral RNAs in somatic
1059 tissues adjacent to the Oikopleura germline. *Nucleic Acids Res*, doi:10.1093/nar/gkv169
1060 (2015).
1061 **20/**Kielkopf, CL *et al.* A novel peptide recognition mode revealed by the X-ray structure of a
1062 core U2AF35/U2AF65 heterodimer. *Cell*, doi:10.1016/s0092-8674(01)00480-9 (2001).
1063 **21/**Yoshida, H *et al.* Elucidation of the aberrant 3' splice site selection by cancer-
1064 associated mutations on the U2AF1. *Nat Commun*, doi:10.1038/s41467-020-18559-6
1065 (2020).
1066 **22/**Ray, D *et al.* RNAcompete methodology and application to determine sequence
1067 preferences of unconventional RNA-binding proteins. *Methods*,
1068 doi:10.1016/j.ymeth.2016.12.003 (2017).
1069 **23/**Martins de Araújo, M *et al.* Differential 3' splice site recognition of SMN1 and SMN2
1070 transcripts by U2AF and U2 snRNP. *Rna*, doi:10.1261/rna.1273209 (2009).
1071 **24/**Shao, C *et al.* Mechanisms for U2AF to define 3' splice sites and regulate alternative
1072 splicing in the human genome. *Nat Struct Mol Biol*, doi:10.1038/nsmb.2906 (2014).
1073 **25/**David, CJ *et al.* The RNA polymerase II C-terminal domain promotes splicing activation
1074 through recruitment of a U2AF65-Prp19 complex. *Genes Dev*, doi:10.1101/gad.2038011
1075 (2011).
1076 **26/**Budnik, N *et al.* The role of S-adenosylhomocysteine hydrolase-like 1 in cancer.
1077 *Biochim Biophys Acta Mol Cell Res*, doi:10.1016/j.bbamcr.2024.119819 (2024).
1078 **27/**Pinto, R *et al.* The human methyltransferase ZCCHC4 catalyses N6-methyladenosine
1079 modification of 28S ribosomal RNA. *Nucleic Acids Res*, doi:10.1093/nar/gkz1147 (2020).
1080 **28/**Ishigami, Y *et al.* A single m(6)A modification in U6 snRNA diversifies exon sequence at
1081 the 5' splice site. *Nat Commun*, doi:10.1038/s41467-021-23457-6 (2021).
1082 **29/**Pendleton, KE *et al.* The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM
1083 Synthetase Intron Retention. *Cell*, doi:10.1016/j.cell.2017.05.003 (2017).
1084 **30/**Mauer, J *et al.* FTO controls reversible m(6)Am RNA methylation during snRNA
1085 biogenesis. *Nat Chem Biol*, doi:10.1038/s41589-019-0231-8 (2019).
1086 **31/**Flamand, MN *et al.* The Proteins of mRNA Modification: Writers, Readers, and Erasers.
1087 *Annu Rev Biochem*, doi:10.1146/annurev-biochem-052521-035330 (2023).
1088 **32/**Hong, T *et al.* Precise Antibody-Independent m6A Identification via 4SedTTP-Involved
1089 and FTO-Assisted Strategy at Single-Nucleotide Resolution. *J Am Chem Soc*,
1090 doi:10.1021/jacs.7b13633 (2018).
1091 **33/**Martinez, NM *et al.* Pseudouridine synthases modify human pre-mRNA co-
1092 transcriptionally and affect pre-mRNA processing. *Mol Cell*,
1093 doi:10.1016/j.molcel.2021.12.023 (2022).
1094 **34/**Kramarek, M *et al.* Splicing analysis of STAT3 tandem donor suggests non-canonical
1095 binding registers for U1 and U6 snRNAs. *Nucleic Acids Res*, doi:10.1093/nar/gkae147
1096 (2024).
1097 **35/**Roost, C *et al.* Structure and thermodynamics of N6-methyladenosine in RNA: a spring-
1098 loaded base modification. *J Am Chem Soc*, doi:10.1021/ja513080v (2015).

1099 **36/**Charenton, C *et al.* Mechanism of 5' splice site transfer for human spliceosome
1100 activation. *Science*, doi:10.1126/science.aax3289 (2019).
1101 **37/**Shen, A *et al.* U6 snRNA m6A modification is required for accurate and efficient splicing
1102 of C. elegans and human pre-mRNAs. *Nucleic Acids Res*, doi:10.1093/nar/gkae447 (2024).
1103 **38/**Galej, WP *et al.* Structural studies of the spliceosome: zooming into the heart of the
1104 machine. *Curr Opin Struct Biol*, doi:10.1016/j.sbi.2013.12.002 (2014).
1105 **39/**Leontis, NB *et al.* The non-Watson-Crick base pairs and their associated isostericity
1106 matrices. *Nucleic Acids Res*, doi:10.1093/nar/gkf481 (2002).
1107 **40/**Larue, GE *et al.* Expansion and transformation of the minor spliceosomal system in the
1108 slime mold Physarum polycephalum. *Curr Biol*, doi:10.1016/j.cub.2021.04.050 (2021).
1109 **41/**Voith von Voithenberg, L *et al.* Recognition of the 3' splice site RNA by the U2AF
1110 heterodimer involves a dynamic population shift. *Proc Natl Acad Sci U S A*,
1111 doi:10.1073/pnas.1605873113 (2016).
1112 **42/**Cho, S *et al.* Interaction between the RNA binding domains of Ser-Arg splicing factor 1
1113 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U*
1114 *S A*, doi:10.1073/pnas.1017700108 (2011).
1115 **43/**Denoeud, F *et al.* Plasticity of animal genome architecture unmasked by rapid evolution
1116 of a pelagic tunicate. *Science*, doi:10.1126/science.1194167 (2010).
1117 **44/**Nabeel-Shah, S *et al.* Recruitment of the m(6)A/m6Am demethylase FTO to target RNAs
1118 by the telomeric zinc finger protein ZBTB48. *Genome Biol*, doi:10.1186/s13059-024-03392-
1119 7 (2024).
1120 **45/**Deirdre, A *et al.* Interactions between the terminal bases of mammalian introns are
1121 retained in inosine-containing pre-mRNAs. *EMBO J*, doi:10.1002/j.1460-
1122 2075.1995.tb07326.x (1995).
1123 **46/**Stombaugh, J *et al.* Frequency and isostericity of RNA base pairs. *Nucleic Acids Res*,
1124 doi:10.1093/nar/gkp011 (2009).
1125 **47/**Li, D *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex
1126 metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*,
1127 doi:10.1093/bioinformatics/btv033 (2015).
1128 **48/**Pardo-Palacios, FJ *et al.* SQANTI3: curation of long-read transcriptomes for accurate
1129 identification of known and novel isoforms. *Nat Methods*, doi:10.1038/s41592-024-02229-
1130 2 (2024).
1131 **49/**Dardaillon, J *et al.* ANISEED 2019: 4D exploration of genetic data for an extended range
1132 of tunicates. *Nucleic Acids Res*, doi:10.1093/nar/gkz955 (2020).
1133 **50/**Zhou, P *et al.* A solubility-enhancement tag (SET) for NMR studies of poorly behaving
1134 proteins. *J Biomol NMR*, doi:10.1023/a:1011258906244 (2001).
1135 **51/**Wyman, D *et al.* A technology-agnostic long-read analysis pipeline for transcriptome
1136 discovery and quantification. *bioRxiv*, doi:10.1101/672931 (2020).
1137 **52/**Clark, F & Thanaraj, TA. Categorization and characterization of transcript-confirmed
1138 constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet*,
1139 doi:10.1093/hmg/11.4.451 (2002).
1140 **53/**Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and
1141 deduplication using fastp. *Imeta*, doi:10.1002/imt2.107 (2023).

1142 **54/**Dobin, A *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*,
1143 doi:10.1093/bioinformatics/bts635 (2013).
1144 **55/**Mudge, JM *et al*. GENCODE 2025: reference gene annotation for human and mouse.
1145 *Nucleic Acids Res*, doi:10.1093/nar/gkae1078 (2025).
1146 **56/**Patro, R *et al*. Salmon provides fast and bias-aware quantification of transcript
1147 expression. *Nat Methods*, doi:10.1038/nmeth.4197 (2017).
1148 **57/**Love, MI *et al*. Moderated estimation of fold change and dispersion for RNA-seq data
1149 with DESeq2. *Genome Biol*, doi:10.1186/s13059-014-0550-8 (2014).
1150 **58/**Soneson, C *et al*. Differential analyses for RNA-seq: transcript-level estimates improve
1151 gene-level inferences. *F1000Res*, doi:10.12688/f1000research.7563.2 (2015).
1152 **59/**Xu, S *et al*. Using clusterProfiler to characterize multiomics data. *Nat Protoc*,
1153 doi:10.1038/s41596-024-01020-z (2024).
1154 **60/**Wang, Y *et al*. rMATS-turbo: an efficient and flexible computational tool for alternative
1155 splicing analysis of large-scale RNA-seq data. *Nat Protoc*, doi:10.1038/s41596-023-00944-
1156 2 (2024).
1157 **61/**Olthof, AM *et al*. Minor intron splicing revisited: identification of new minor intron-
1158 containing genes and tissue-dependent retention and alternative splicing of minor introns.
1159 *BMC Genomics*, doi:10.1186/s12864-019-6046-x (2019).
1160 **62/**Szklarczyk, D *et al*. The STRING database in 2025: protein networks with directionality
1161 of regulation. *Nucleic Acids Res*, doi:10.1093/nar/gkae1113 (2025).
1162 **63/**Olieric, V *et al*. A fast selenium derivatization strategy for crystallization and phasing of
1163 RNA structures. *RNA*, doi:10.1261/rna.1499309 (2009).

1164