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, splicing is a critical step of gene expression that produces transcripts competent for protein synthesis by removing introns from pre-mRNA. The spliceosome defines target splice sites based on sequences located around intron ends and in almost all cases, spliced segments begin with a GT and end with an AG. The key molecules for this function are the spliceosomal RNA (snRNA) and the U2AF heterodimer, which interact with splice signals found at the ends of introns. In chordates, tunicate genomes in the family Fritillariidae (Lohmann, 1915) form a remarkable exception by hosting numerous introns that escape the GT/AG rule, and they represent a material of choice to address the evolution of intron recognition.

A comparison of chordate genomes revealed that the emergence of the unusual splicing code in *Fritillaria borealis* was accompanied by the duplication of its subunits U2AF1 and U2AF2. We found that the conserved duplicates bind to canonical splice signals on RNA, but different binding was seen with the divergent versions U2AF1β and U2AF2β, suggesting they could recognize other splice sites. When expressed in mammalian cells, U2AF2β promotes the removal of introns that have a weak polypyrimidine tract, and we also show that it interacts specifically with AHCYL1, a protein that can regulate the N6-methyladenosine (m6A) modification of snRNA U2 and U6. 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 modifications could play decisive roles during the recognition of non-GT/AG introns. Our study suggests that discrete changes implemented on the spliceosome during the evolution of Fritillariidae have contributed to a dramatic shift in splice site identity.

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

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

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

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

94 **Results**

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

97 Genetic resources for *F. borealis* were initially limited by specimen availability, only
98 permitting to deploy short-read sequencing that produced genome assemblies with an
99 important level of fragmentation[18]. We took advantage of recent developments in long-read
100 sequencing techniques and laboratory culture to improve the genome assembly, and to
101 establish a collection of full-length transcripts. The latest assembly features a smaller
102 number of contigs and higher N50 (**Table S1**), which is useful for aligning full-length cDNA
103 on the *F. borealis* genome. Based on the annotation of >30.000 introns discovered after
104 transcript alignment, we could count only a small fraction of canonical introns (5%), while
105 AG/AC and AG/AT were the most frequent termini seen among non-canonical introns
106 (respectively 18 and 16%) (**Fig1B**). To gain better insight into intron evolution in the
107 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

362 **Discussion**

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

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

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

395 *Approaching the new functions of U2AF*

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

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

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

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

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

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

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

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

470 **Acknowledgments**

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

481 **Material and Methods**

482 *Animal samples*

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

488 *Genome sequencing*

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

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

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

510 *Transcriptome sequencing*

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

519 *Genome assembly*

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

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

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

537 *Gene and intron annotations*

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

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

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

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

572 *Homology search*

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

581 *Expression and purification of U2AF heterodimers*

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

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

606 *Electrophoretic mobility shift assays*

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

625 *RNA-compete assays*

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

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

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

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

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

668 *Mammalian cell expression and transfection*

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

674 *Long-read sequencing of cDNA*

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

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

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

703 *Transcriptome analysis of U2AF2-expressing cells*

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

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

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

724 *Identification of U2AF2 partners*

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

740 *Generation of AHCYL1 KO cells*

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

749 *Detection of N6-methyladenosine in snRNA*

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

761 *Mapping of m6A in U1 snRNA*

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

772

773

774

775

776

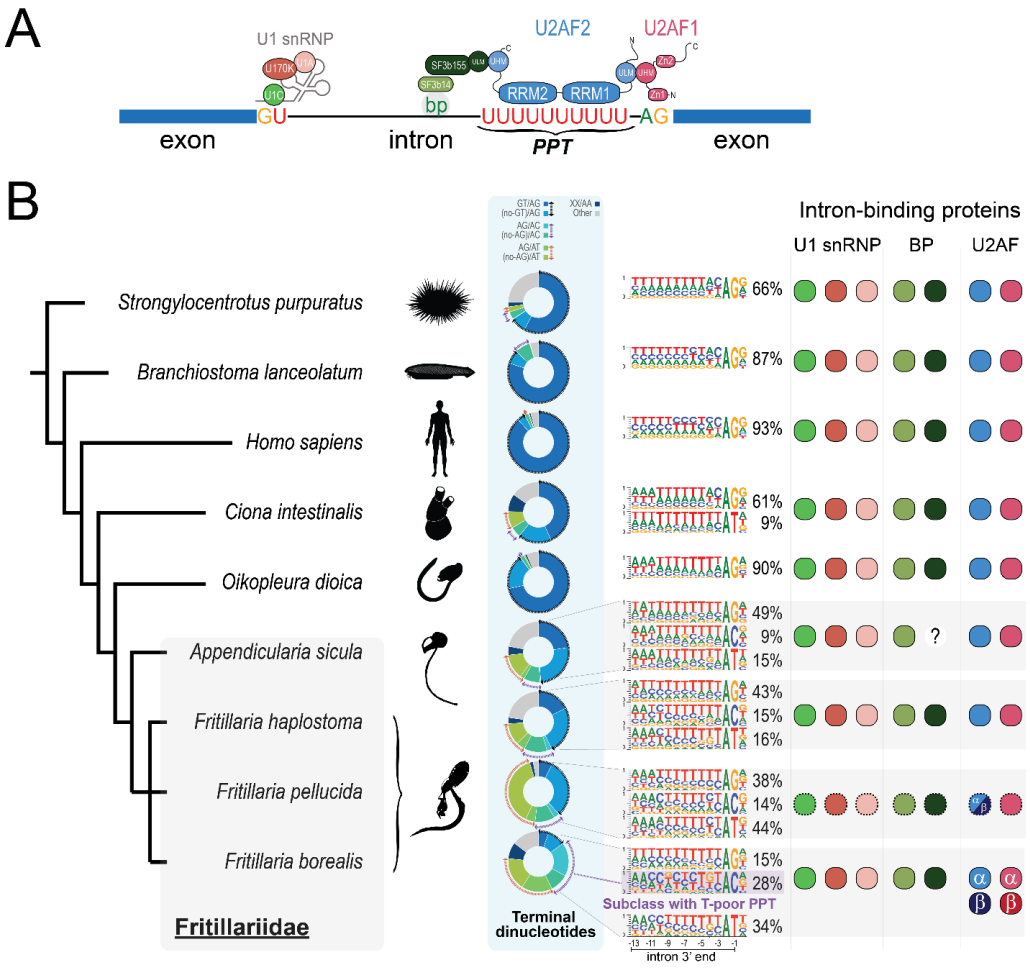
777

778

779

780

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

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

811

812

813

814

815

816

817

818

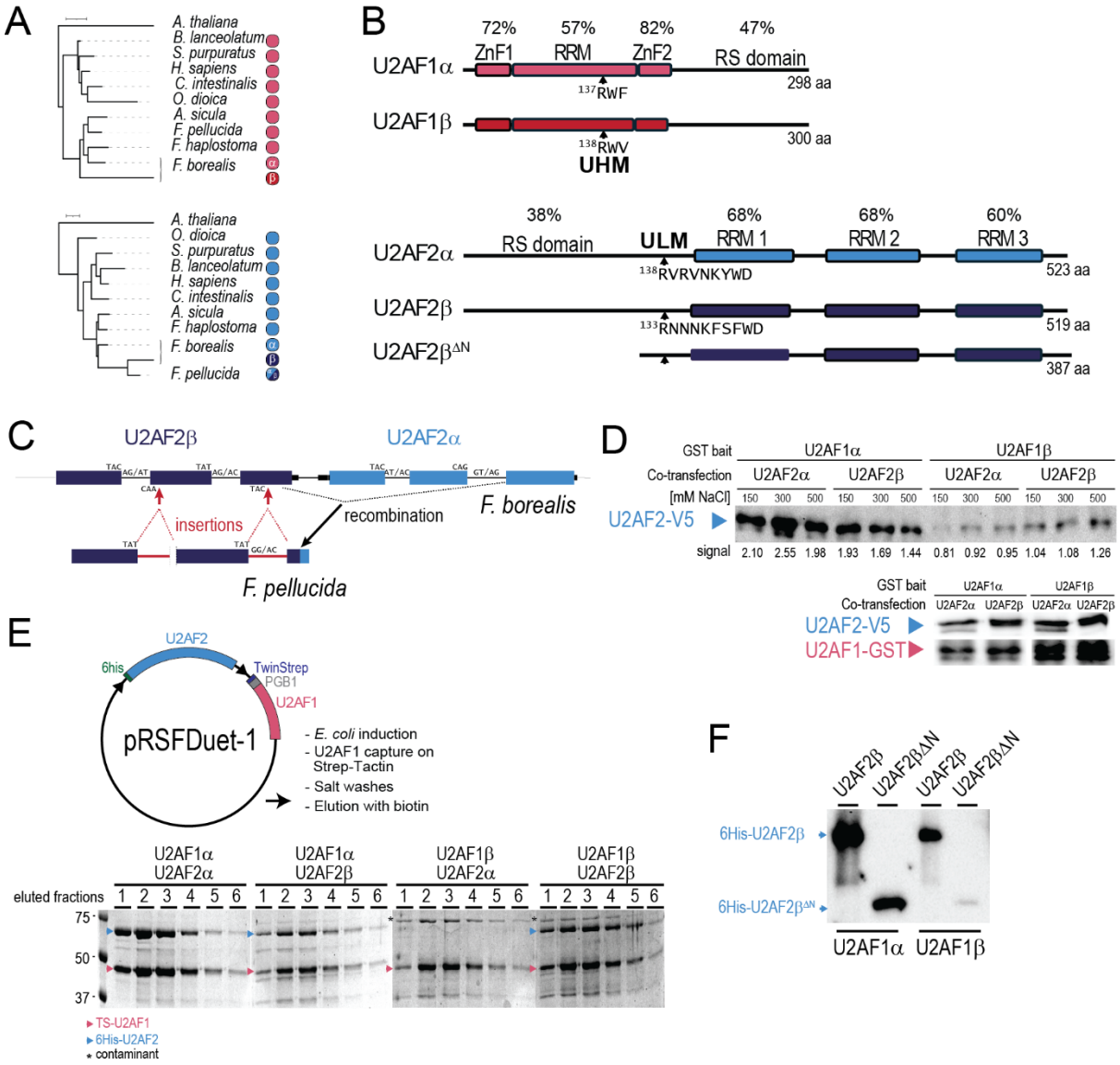
819

820

821

822

823



824

825

826

827

828

829

830

831

832

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

853

854

855

856

857

858

859

860

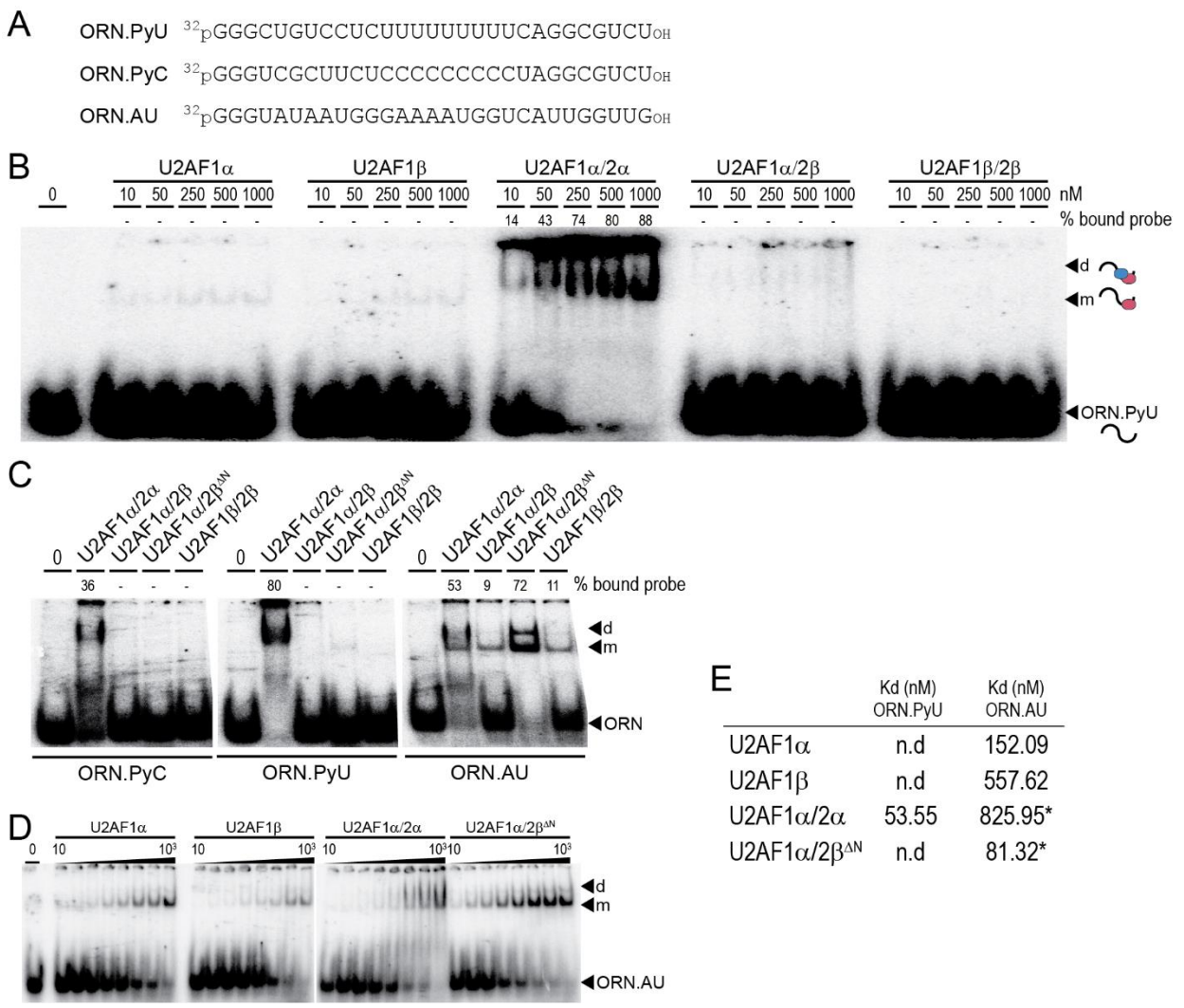


Figure 3: U2AF binding on RNA tested with EMSA. **A)** Representation of RNA probes used for EMSA. **B)** EMSA with ORN.PyU and purified U2AF. From left to right, experimental series were performed with U2AF1 only, or with different U2AF dimers. Bound probe measurement is reported only for values superior 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.

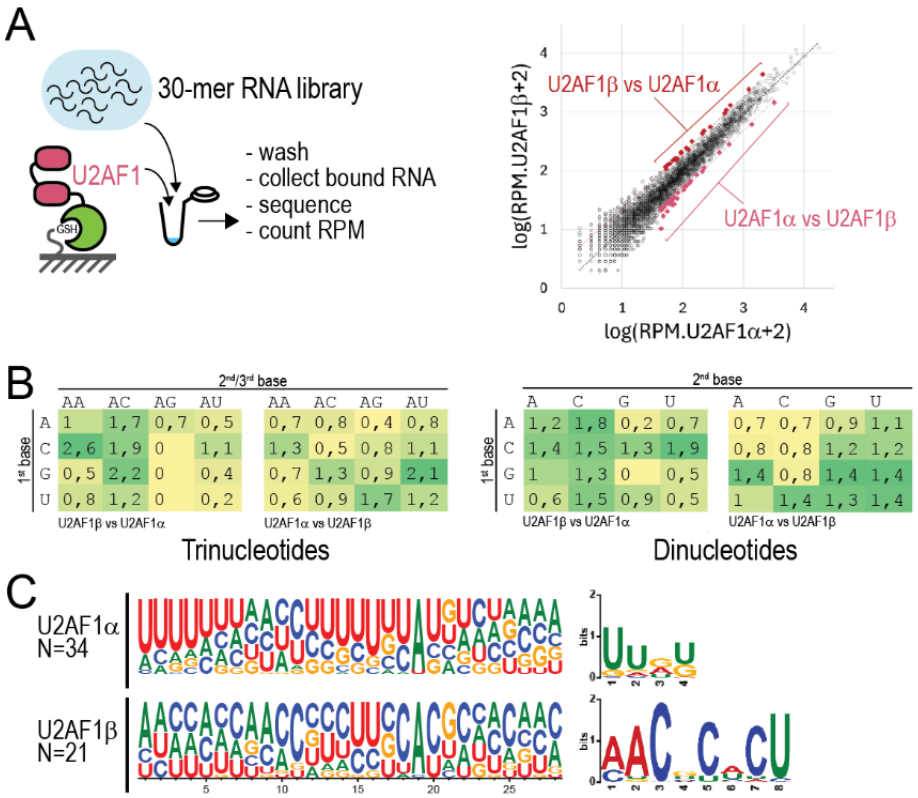
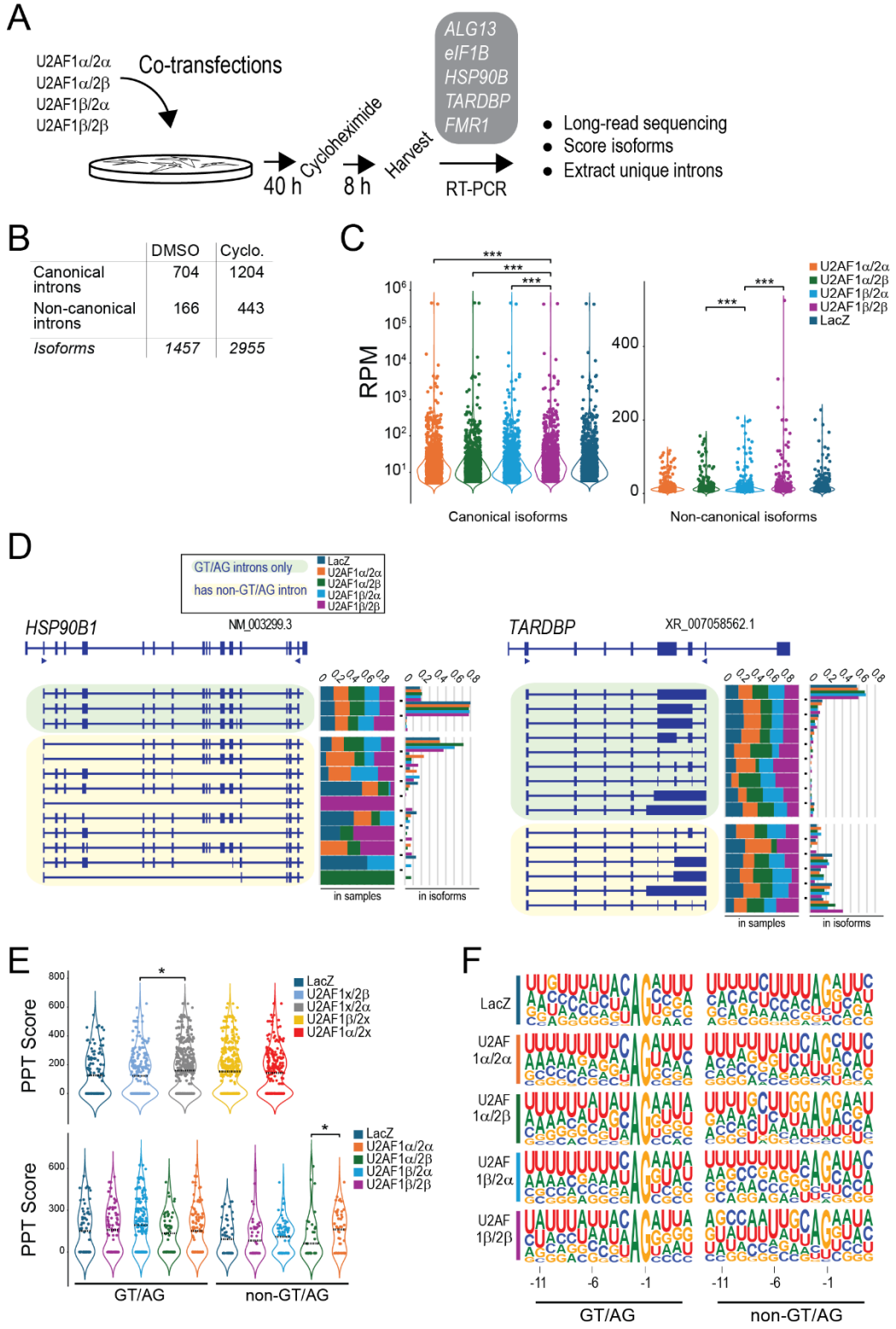


Figure 4: RNA-compete experiments with U2AF1. **A)** On the left, schematic drawing of approach and on the right, read count plot showing differential enrichment of library oligos in RNA recovered with U2AF1α (x-axis) and with U2AF1β (y-axis). Colored diamonds represent groups of high affinity oligos 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

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

911

912

913

914

915

916

917

918

919

920

921

922

923

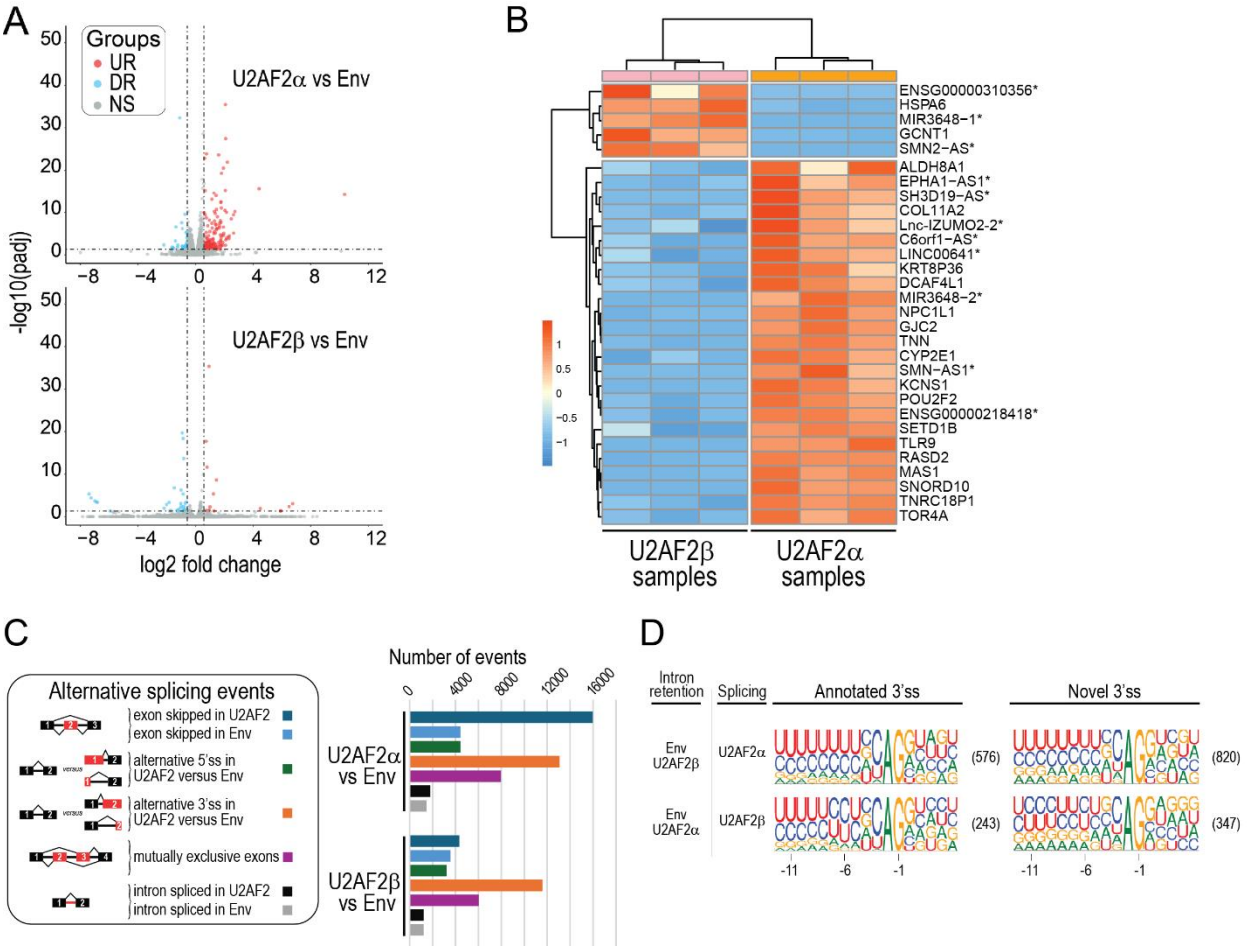
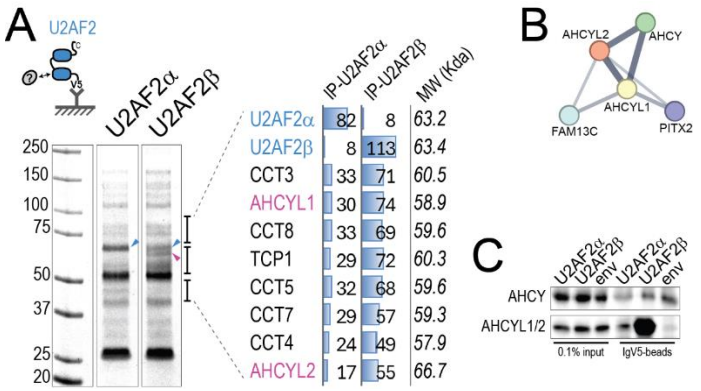


Figure 6: Transcriptome analysis of cells expressing *F. borealis* U2AF2. **A)** Volcano plots showing DEGs found by comparing the transcriptome of cells transfected with either U2AF2 paralogues, or control cells expressing an envelope protein. UR, upregulated genes; DR, down-regulated genes; NS, non-significant 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.



935

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

951

952

953

954

955

956

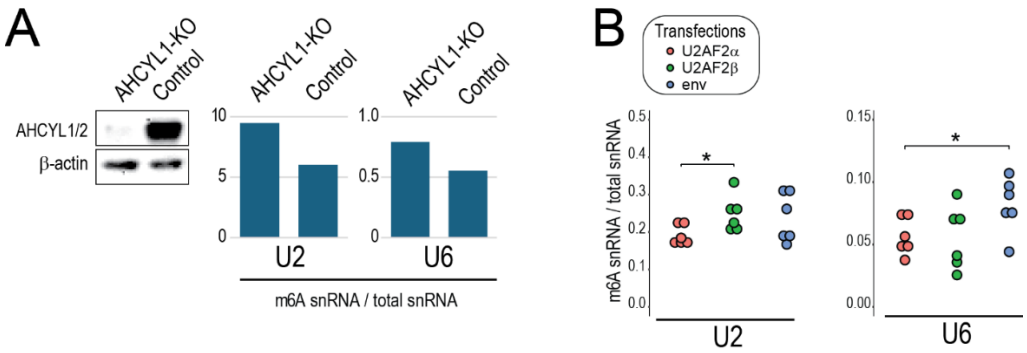


Figure 8: U2AF2 and AHCYL1 reduce m6A levels in snRNA. **A)** Impact of AHCYL1 knock-out. Western blot shows loss of AHCYL1 expression after inactivating the *ahcyl1* gene in HEK293T cells. Graphs show m6A levels measured in U2 or U6 snRNA, compared to the total snRNA amount. **B)** Impact of U2AF2 paralogues.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

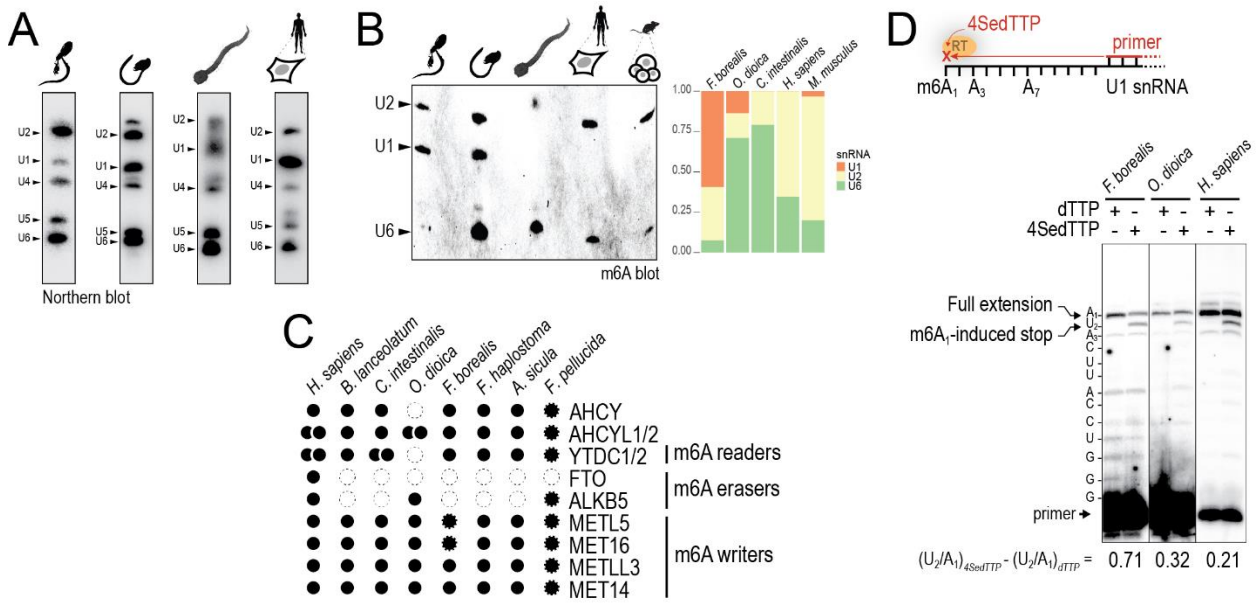


Figure 9: Divergent m6A patterns in *F. borealis* snRNA. **A)** Conservation of U2-type snRNAs. Northern blots show snRNA signals in total RNA extracted from (panels from left to right): maturing adults of *F. borealis* and *O. dioica*, *C. intestinalis* larvae and HEK293T cells. **B)** Detection of 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.

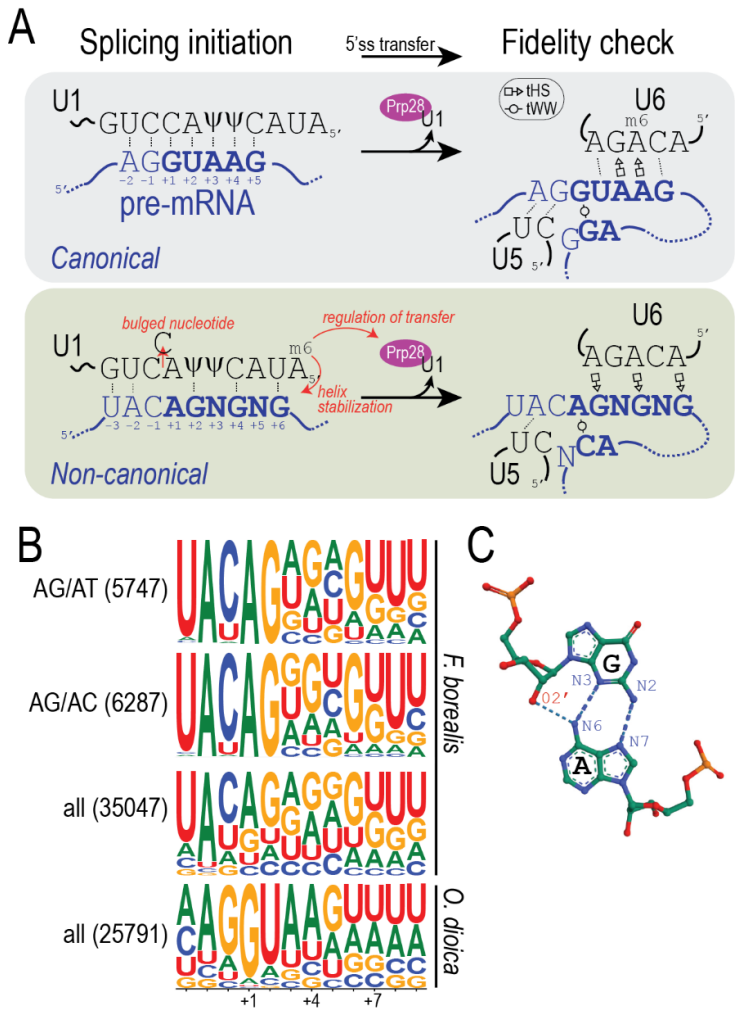


Figure 10: Function of m6A modification during non-canonical splicing. **A)** The RNA network of 5'ss recognition. The top panel is a model based on Artemyeva-Isman & Porter[7] representing base-pair interactions established between the 5'ss (intronic residues shown in bold) and U1 snRNA during the initiation 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].

1015

1016 **References**

1017 **1/**Suzuki, H *et al.* Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH
1018 medulloblastoma. *Nature*, doi:10.1038/s41586-019-1650-0 (2019).
1019 **2/**Glasser, E *et al.* Pre-mRNA splicing factor U2AF2 recognizes distinct conformations of
1020 nucleotide variants at the center of the pre-mRNA splice site signal. *Nucleic Acids Res*,
1021 doi:10.1093/nar/gkac287 (2022).
1022 **3/**Yoshida, H *et al.* A novel 3' splice site recognition by the two zinc fingers in the U2AF
1023 small subunit. *Genes Dev*, doi:10.1101/gad.267104.115 (2015).
1024 **4/**Olthof, AM *et al.* Taxonomy of introns and the evolution of minor introns. *Nucleic Acids*
1025 *Res*, doi:10.1093/nar/gkae550 (2024).
1026 **5/**Erkelenz, S *et al.* Ranking noncanonical 5' splice site usage by genome-wide RNA-seq
1027 analysis and splicing reporter assays. *Genome Res*, doi:10.1101/gr.235861.118 (2018).
1028 **6/**Parker, R & Siliciano, PG. Evidence for an essential non-Watson-Crick interaction
1029 between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature*,
1030 doi:10.1038/361660a0 (1993).
1031 **7/**Artemyeva-Isman, OV & Porter, ACG. U5 snRNA Interactions With Exons Ensure Splicing
1032 Precision. *Front Genet*, doi:10.3389/fgene.2021.676971 (2021).
1033 **8/**Bai, R *et al.* Structure of the Post-catalytic Spliceosome from Saccharomyces cerevisiae.
1034 *Cell*, doi:10.1016/j.cell.2017.10.038 (2017).
1035 **9/**Costa, M *et al.* Crystal structures of a group II intron lariat primed for reverse splicing.
1036 *Science*, doi:10.1126/science.aaf9258 (2016).
1037 **10/**Henriet, S *et al.* Evolution of the U2 Spliceosome for Processing Numerous and Highly
1038 Diverse Non-canonical Introns in the Chordate Fritillaria borealis. *Curr Biol*,
1039 doi:10.1016/j.cub.2019.07.092 (2019).
1040 **11/**Huff, JT *et al.* Mechanism for DNA transposons to generate introns on genomic scales.
1041 *Nature*, doi:10.1038/nature20110 (2016).
1042 **12/**Farhat, S *et al.* Rapid protein evolution, organellar reductions, and invasive intronic
1043 elements in the marine aerobic parasite dinoflagellate Amoebophrya spp. *BMC Biol*,
1044 doi:10.1186/s12915-020-00927-9 (2021).
1045 **13/**Guminska, N *et al.* Order of removal of conventional and nonconventional introns from
1046 nuclear transcripts of Euglena gracilis. *PLoS Genet*, doi:10.1371/journal.pgen.1007761
1047 (2018).
1048 **14/**Marz, M *et al.* Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol*,
1049 doi:10.1007/s00239-008-9149-6 (2008).
1050 **15/**Henriet, S *et al.* Laboratory study of Fritillaria lifecycle reveals key morphogenetic
1051 events leading to genus-specific anatomy. *Front Zool*, doi:10.1186/s12983-022-00471-y
1052 (2022).
1053 **16/**Sales-Lee, J *et al.* Coupling of spliceosome complexity to intron diversity. *Curr Biol*,
1054 doi:10.1016/j.cub.2021.09.004 (2021).
1055 **17/**Stark, MR *et al.* Dramatically reduced spliceosome in Cyanidioschyzon merolae. *Proc*
1056 *Natl Acad Sci U S A*, doi:10.1073/pnas.1416879112 (2015).

1057 **18/**Naville, M *et al.* Massive Changes of Genome Size Driven by Expansions of Non-
1058 autonomous Transposable Elements. *Curr Biol*, doi:10.1016/j.cub.2019.01.080 (2019).
1059 **19/**Henriet, S *et al.* Embryonic expression of endogenous retroviral RNAs in somatic
1060 tissues adjacent to the Oikopleura germline. *Nucleic Acids Res*, doi:10.1093/nar/gkv169
1061 (2015).
1062 **20/**Kielkopf, CL *et al.* A novel peptide recognition mode revealed by the X-ray structure of a
1063 core U2AF35/U2AF65 heterodimer. *Cell*, doi:10.1016/s0092-8674(01)00480-9 (2001).
1064 **21/**Yoshida, H *et al.* Elucidation of the aberrant 3' splice site selection by cancer-
1065 associated mutations on the U2AF1. *Nat Commun*, doi:10.1038/s41467-020-18559-6
1066 (2020).
1067 **22/**Ray, D *et al.* RNAcompete methodology and application to determine sequence
1068 preferences of unconventional RNA-binding proteins. *Methods*,
1069 doi:10.1016/j.ymeth.2016.12.003 (2017).
1070 **23/**Martins de Araújo, M *et al.* Differential 3' splice site recognition of SMN1 and SMN2
1071 transcripts by U2AF and U2 snRNP. *Rna*, doi:10.1261/rna.1273209 (2009).
1072 **24/**Shao, C *et al.* Mechanisms for U2AF to define 3' splice sites and regulate alternative
1073 splicing in the human genome. *Nat Struct Mol Biol*, doi:10.1038/nsmb.2906 (2014).
1074 **25/**David, CJ *et al.* The RNA polymerase II C-terminal domain promotes splicing activation
1075 through recruitment of a U2AF65-Prp19 complex. *Genes Dev*, doi:10.1101/gad.2038011
1076 (2011).
1077 **26/**Budnik, N *et al.* The role of S-adenosylhomocysteine hydrolase-like 1 in cancer.
1078 *Biochim Biophys Acta Mol Cell Res*, doi:10.1016/j.bbamcr.2024.119819 (2024).
1079 **27/**Pinto, R *et al.* The human methyltransferase ZCCHC4 catalyses N6-methyladenosine
1080 modification of 28S ribosomal RNA. *Nucleic Acids Res*, doi:10.1093/nar/gkz1147 (2020).
1081 **28/**Ishigami, Y *et al.* A single m(6)A modification in U6 snRNA diversifies exon sequence at
1082 the 5' splice site. *Nat Commun*, doi:10.1038/s41467-021-23457-6 (2021).
1083 **29/**Pendleton, KE *et al.* The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM
1084 Synthetase Intron Retention. *Cell*, doi:10.1016/j.cell.2017.05.003 (2017).
1085 **30/**Mauer, J *et al.* FTO controls reversible m(6)Am RNA methylation during snRNA
1086 biogenesis. *Nat Chem Biol*, doi:10.1038/s41589-019-0231-8 (2019).
1087 **31/**Flamand, MN *et al.* The Proteins of mRNA Modification: Writers, Readers, and Erasers.
1088 *Annu Rev Biochem*, doi:10.1146/annurev-biochem-052521-035330 (2023).
1089 **32/**Hong, T *et al.* Precise Antibody-Independent m6A Identification via 4SedTTP-Involved
1090 and FTO-Assisted Strategy at Single-Nucleotide Resolution. *J Am Chem Soc*,
1091 doi:10.1021/jacs.7b13633 (2018).
1092 **33/**Martinez, NM *et al.* Pseudouridine synthases modify human pre-mRNA co-
1093 transcriptionally and affect pre-mRNA processing. *Mol Cell*,
1094 doi:10.1016/j.molcel.2021.12.023 (2022).
1095 **34/**Kramarek, M *et al.* Splicing analysis of STAT3 tandem donor suggests non-canonical
1096 binding registers for U1 and U6 snRNAs. *Nucleic Acids Res*, doi:10.1093/nar/gkae147
1097 (2024).
1098 **35/**Roost, C *et al.* Structure and thermodynamics of N6-methyladenosine in RNA: a spring-
1099 loaded base modification. *J Am Chem Soc*, doi:10.1021/ja513080v (2015).

1100 **36/**Charenton, C *et al.* Mechanism of 5' splice site transfer for human spliceosome
1101 activation. *Science*, doi:10.1126/science.aax3289 (2019).
1102 **37/**Shen, A *et al.* U6 snRNA m6A modification is required for accurate and efficient splicing
1103 of C. elegans and human pre-mRNAs. *Nucleic Acids Res*, doi:10.1093/nar/gkae447 (2024).
1104 **38/**Galej, WP *et al.* Structural studies of the spliceosome: zooming into the heart of the
1105 machine. *Curr Opin Struct Biol*, doi:10.1016/j.sbi.2013.12.002 (2014).
1106 **39/**Leontis, NB *et al.* The non-Watson-Crick base pairs and their associated isostericity
1107 matrices. *Nucleic Acids Res*, doi:10.1093/nar/gkf481 (2002).
1108 **40/**Larue, GE *et al.* Expansion and transformation of the minor spliceosomal system in the
1109 slime mold Physarum polycephalum. *Curr Biol*, doi:10.1016/j.cub.2021.04.050 (2021).
1110 **41/**Voith von Voithenberg, L *et al.* Recognition of the 3' splice site RNA by the U2AF
1111 heterodimer involves a dynamic population shift. *Proc Natl Acad Sci U S A*,
1112 doi:10.1073/pnas.1605873113 (2016).
1113 **42/**Cho, S *et al.* Interaction between the RNA binding domains of Ser-Arg splicing factor 1
1114 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U*
1115 *S A*, doi:10.1073/pnas.1017700108 (2011).
1116 **43/**Denoeud, F *et al.* Plasticity of animal genome architecture unmasked by rapid evolution
1117 of a pelagic tunicate. *Science*, doi:10.1126/science.1194167 (2010).
1118 **44/**Nabeel-Shah, S *et al.* Recruitment of the m(6)A/m6Am demethylase FTO to target RNAs
1119 by the telomeric zinc finger protein ZBTB48. *Genome Biol*, doi:10.1186/s13059-024-03392-
1120 7 (2024).
1121 **45/**Deirdre, A *et al.* Interactions between the terminal bases of mammalian introns are
1122 retained in inosine-containing pre-mRNAs. *EMBO J*, doi:10.1002/j.1460-
1123 2075.1995.tb07326.x (1995).
1124 **46/**Stombaugh, J *et al.* Frequency and isostericity of RNA base pairs. *Nucleic Acids Res*,
1125 doi:10.1093/nar/gkp011 (2009).
1126 **47/**Li, D *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex
1127 metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*,
1128 doi:10.1093/bioinformatics/btv033 (2015).
1129 **48/**Pardo-Palacios, FJ *et al.* SQANTI3: curation of long-read transcriptomes for accurate
1130 identification of known and novel isoforms. *Nat Methods*, doi:10.1038/s41592-024-02229-
1131 2 (2024).
1132 **49/**Dardaillon, J *et al.* ANISEED 2019: 4D exploration of genetic data for an extended range
1133 of tunicates. *Nucleic Acids Res*, doi:10.1093/nar/gkz955 (2020).
1134 **50/**Zhou, P *et al.* A solubility-enhancement tag (SET) for NMR studies of poorly behaving
1135 proteins. *J Biomol NMR*, doi:10.1023/a:1011258906244 (2001).
1136 **51/**Wyman, D *et al.* A technology-agnostic long-read analysis pipeline for transcriptome
1137 discovery and quantification. *bioRxiv*, doi:10.1101/672931 (2020).
1138 **52/**Clark, F & Thanaraj, TA. Categorization and characterization of transcript-confirmed
1139 constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet*,
1140 doi:10.1093/hmg/11.4.451 (2002).
1141 **53/**Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and
1142 deduplication using fastp. *Imeta*, doi:10.1002/imt2.107 (2023).

1143 **54/**Dobin, A *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*,
1144 doi:10.1093/bioinformatics/bts635 (2013).
1145 **55/**Mudge, JM *et al*. GENCODE 2025: reference gene annotation for human and mouse.
1146 *Nucleic Acids Res*, doi:10.1093/nar/gkae1078 (2025).
1147 **56/**Patro, R *et al*. Salmon provides fast and bias-aware quantification of transcript
1148 expression. *Nat Methods*, doi:10.1038/nmeth.4197 (2017).
1149 **57/**Love, MI *et al*. Moderated estimation of fold change and dispersion for RNA-seq data
1150 with DESeq2. *Genome Biol*, doi:10.1186/s13059-014-0550-8 (2014).
1151 **58/**Soneson, C *et al*. Differential analyses for RNA-seq: transcript-level estimates improve
1152 gene-level inferences. *F1000Res*, doi:10.12688/f1000research.7563.2 (2015).
1153 **59/**Xu, S *et al*. Using clusterProfiler to characterize multiomics data. *Nat Protoc*,
1154 doi:10.1038/s41596-024-01020-z (2024).
1155 **60/**Wang, Y *et al*. rMATS-turbo: an efficient and flexible computational tool for alternative
1156 splicing analysis of large-scale RNA-seq data. *Nat Protoc*, doi:10.1038/s41596-023-00944-
1157 2 (2024).
1158 **61/**Olthof, AM *et al*. Minor intron splicing revisited: identification of new minor intron-
1159 containing genes and tissue-dependent retention and alternative splicing of minor introns.
1160 *BMC Genomics*, doi:10.1186/s12864-019-6046-x (2019).
1161 **62/**Szklarczyk, D *et al*. The STRING database in 2025: protein networks with directionality
1162 of regulation. *Nucleic Acids Res*, doi:10.1093/nar/gkae1113 (2025).
1163 **63/**Olieric, V *et al*. A fast selenium derivatization strategy for crystallization and phasing of
1164 RNA structures. *RNA*, doi:10.1261/rna.1499309 (2009).

1165