

1 **The rapid evolution of alternative splicing in plants**

2 Zhihao Ling, Thomas Brockmöller, Ian T. Baldwin and Shuqing Xu\*

3

4 E-mail addresses: ZL: [zling@ice.mpg.de](mailto:zling@ice.mpg.de), TB: [tbrockmoeller@ice.mpg.de](mailto:tbrockmoeller@ice.mpg.de), ITB:

5 [baldwin@ice.mpg.de](mailto:baldwin@ice.mpg.de), SX: [sxu@ice.mpg.de](mailto:sxu@ice.mpg.de)

6 Running title: rapid evolution of alternative splicing in plants.

7

8 \* Correspondence: Shuqing Xu, Department of Molecular Ecology, Max Planck Institute for  
9 Chemical Ecology, Hans-Knöll-Straße 8 D-07745 Jena Germany.

10 E-mail: [sxu@ice.mpg.de](mailto:sxu@ice.mpg.de)

11 Phone: +49 (0)3641 57 1122

12 The authors declare no conflict of interest.

13

14 **One sentence summary:** Changes of intron located splicing regulators contributed to the rapid  
15 evolution of alternative splicing in plants.

**List of author contributions:** Z.L and S.X. designed the research. Z.L. T.B. and S.X. performed the experiments and analyzed the data. Z.L and S.X. wrote the paper with contributions of all authors. I.T.B provided part of the sequencing data and complemented writing.

**Financial Sources:** Max Planck Society (all), Swiss National Science Foundation (project number: PEBZP3-142886 to SX), Marie Curie Intra-European Fellowship (IEF) (Project Number: 328935 to SX), European Research Council advanced grant ClockworkGreen (Project number: 293926 to ITB)

**Corresponding Author:** Shuqing Xu, [sxu@ice.mpg.de](mailto:sxu@ice.mpg.de)

16 **Abstract**

17 Alternative pre-mRNA splicing (AS) is prevalent among all plants and is involved in many  
18 interactions with environmental stresses. However, the evolutionary patterns and underlying  
19 mechanisms of AS in plants remain unclear. By analyzing the transcriptomes of six plant species,  
20 we revealed that AS diverged rapidly among closely related species, largely due to the gains and  
21 losses of AS events among orthologous genes. Furthermore, AS that generates transcripts  
22 containing premature termination codons (PTC), although only representing a small fraction of  
23 the total AS, are more conserved than those that generate non-PTC containing transcripts,  
24 suggesting that AS coupled with nonsense-mediated decay (NMD) might play an important role  
25 in regulating mRNA levels post-transcriptionally. With a machine learning approach we  
26 analyzed the key determinants of AS to understand the mechanisms underlying its rapid  
27 divergence. Among the studied species, the presence/absence of alternative splicing site (SS)  
28 within the junction, the distance between the authentic SS and the nearest alternative SS, the size  
29 of exon-exon junctions were the major determinants for both alternative 5' donor site and  
30 3' acceptor site, suggesting a relatively conserved AS mechanism. Comparative analysis further  
31 demonstrated that variations of the identified AS determinants, mostly are located in introns,  
32 significantly contributed to the AS turnover among closely related species in both Solanaceae  
33 and Brassicaceae taxa. These new mechanistic insights into the evolution of AS in plants  
34 highlight the importance of post-transcriptional regulation in mediating plant-environment  
35 interactions.

## 36 INTRODUCTION

37 Due to their sessile lifestyle, plants have evolved various mechanisms to respond to  
38 environmental stresses. Alternative splicing (AS), a mechanism by which different mature RNAs  
39 are formed by removing different introns or using different splice sites (SS) from the same pre-  
40 mRNA, is known to be important for stress-induced responses in plants (Mastrangelo et al., 2012;  
41 Staiger and Brown, 2013). Both biotic and abiotic stresses such as herbivores (Ling et al., 2015),  
42 pathogens (Howard et al., 2013), cold (Leviatan et al., 2013) and salt (Ding et al., 2014) can all  
43 induce genome-wide changes in AS in plants. The environment-induced AS changes in turn can  
44 affect phenotypic traits of plants and may contribute to their adaptations to different stresses  
45 (Mastrangelo et al., 2012; Staiger and Brown, 2013). For example, low temperature-induced AS  
46 changes of flowering regulator genes affect flowering time and floral development in *A. thaliana*  
47 (Severing et al., 2012; Rosloski et al., 2013). The strong association between AS and  
48 environmental stimuli suggests that AS is involved in adaptation processes and thus evolved  
49 rapidly.

50 Two main functions of AS have been postulated: (i) to expand proteome diversity when  
51 different transcript isoforms are translated into different proteins (with different subcellular  
52 localizations, stability, enzyme activity etc.) (Kazan, 2003; Reddy, 2007; Barbazuk et al., 2008);  
53 (ii) to regulate gene expression by generating transcripts harboring premature termination codons  
54 (PTC) that are recognized by the nonsense-mediated decay (NMD) machinery and degraded  
55 (Chang et al., 2007; Kalyna et al., 2012; Kervestin and Jacobson, 2012). Although initially  
56 considered to be transcriptional noise, several AS events that introduce PTCs have been found to  
57 be highly conserved in animals (Ni et al., 2007; Lareau and Brenner, 2015) and plants (Iida and  
58 Go, 2006; Kalyna et al., 2006; Darracq and Adams, 2013), suggesting that that the combination

59 of AS with NMD might play an important role in controlling mRNA levels post-transcriptionally.  
60 However, it is unclear whether NMD-coupled AS is more conserved than the AS that generates  
61 transcripts without PTC at a genome-wide level.

62 The evolution of AS in plants, compared to that in vertebrates, remains largely unclear.  
63 Studies that compared organ-specific transcriptomes from different vertebrate species spanning  
64 ~350 million years of evolution showed that AS complexity differs dramatically among  
65 vertebrate lineages, and AS evolved much faster than gene expression has (Barbosa-Morais et al.,  
66 2012; Merkin et al., 2012). For example, within 6 million years, the splicing profiles of an organ  
67 are more similar to other organs of the same species than the same organ in other species, while  
68 the expression profiles of the same organ are similar to the organ in other species (Barbosa-  
69 Morais et al., 2012; Merkin et al., 2012). In plants, largely due to the lack of comprehensive  
70 transcriptomic data, such comparative analysis remains unavailable. However, several  
71 indications suggest that AS in plants and vertebrates may share evolutionary pattern of the rapid  
72 divergence. For example, only 16.4% of AS between maize and rice, and 5.4% between *Brassica*  
73 and *Arabidopsis* are conserved (Severing et al., 2009; Darracq and Adams, 2013). A more recent  
74 study further showed that only 2.8% of genes showed conserve AS between two species of mung  
75 beans, *Vigna radiate* and *V. angularis* (Satyawati et al., 2016). Furthermore, large changes in AS  
76 also exist between different ecotypes of the same species (Streitner et al., 2012). However, such  
77 low conservations of AS found among species could also be due to several other confounding  
78 effects. For example, it is also known that the levels of gene expression, which are highly  
79 associated with AS, also diverge rapidly in plants (Yang and Wang, 2013). As a consequence, it  
80 remains unclear whether the low observed levels of AS conservation results simply from the  
81 rapid expression changes between species. Furthermore, AS detection of is highly dependent on

82 sequencing depth and the tissue types used for generating transcriptomic data (Xu et al., 2002;  
83 Ellis et al., 2012; Ling et al., 2015). Therefore, it is necessary to systematically control for  
84 different confounding effects in order understand the evolutionary pattern of AS in plants.

85 From a mechanistic perspective, divergence of AS among species is contributed by  
86 factors that affect the exon-intron splicing process, which is mediated by the spliceosome. While  
87 the recognition processes of exonic and intronic regions are directed by sequence features of the  
88 pre-mRNA in animals, how the spliceosome removes introns and ligates exons is poorly  
89 understood in plants. In metazoans, it is known that four crucial signals are required for accurate  
90 splicing: (i) 5' splice sites (SS), which contains a GU dinucleotide at the intron start surrounded  
91 by a piece of longer consensus sequences of lower conservation, (ii) 3' SS, which includes an  
92 AG at the 3' end surrounded by similar sequences of 5' SS, (iii) a polypyrimidine tract (PPT) and  
93 (iv) a branch site (BS) sequence located ~17-40 nt upstream of the 3' SS (Black, 2003; Fu and  
94 Ares, 2014). In plants, similar sequence features with a small difference at specific positions  
95 were found, except for the requirement of a BS (Reddy, 2007). In addition, a UA-rich tract in  
96 introns has also been found to be important for efficient splicing in plants (Lewandowska et al.,  
97 2004; Simpson et al., 2004; Baek et al., 2008). In animals, the regulation of splicing also depends  
98 largely on *cis* signals and *trans*-acting splicing factors (SFs) that can recognize the signals  
99 (Barbosa-Morais et al., 2012; Merkin et al., 2012). Among different SFs, serine/arginine-rich  
100 (SR) proteins are from an important SFs family that has been shown to be involved in AS  
101 regulation (Lopato et al., 1999; Gao et al., 2004; Wang and Brendel, 2004; Reddy, 2007; Reddy  
102 and Shad Ali, 2011). In addition, many splicing regulatory elements (SREs) and RNA-binding  
103 proteins (RBPs) have been identified in animals, and the interactions among these SREs in the  
104 pre-mRNA and RBPs were found either to promote or suppress the use of a particular splice sites

105 (Licatalosi et al., 2008; Chen and Manley, 2009; Barash et al., 2010). The number of SR proteins  
106 genes in plants is nearly twice that of number found in non-photosynthetic organisms, although  
107 the number varies among different species (Iida and Go, 2006; Isshiki et al., 2006; Richardson et  
108 al., 2011). To date, more than 200 RBPs and 80 SREs in plants have been identified using  
109 computational approaches (Lorkovic, 2009), however, only a few of these have been functionally  
110 validated (Yoshimura et al., 2002; Pertea et al., 2007; Schoning et al., 2008; Thomas et al., 2012).

111 In mammals, the emergence of AS originated from constitutive splicing with the fixation  
112 of SREs and the creation of alternative competing SS (Koren et al., 2007; Lev-Maor et al., 2007).  
113 Distinctive features that distinguish alternatively spliced exons/introns from constitutively  
114 spliced exons/introns can be used to accurately predict the specific AS type (Koren et al., 2007;  
115 Braunschweig et al., 2014). Furthermore, other factors including secondary and tertiary RNA  
116 structures, chromatin remodeling, insertion of transposable elements (TEs) and gene duplication  
117 (GD) may also be involved in regulating AS (Liu et al., 1995; Sorek et al., 2002; Donahue et al.,  
118 2006; Su et al., 2006; Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009; Warf and Berglund,  
119 2010; Lambert et al., 2015). However, the extent to which changes in these factors contributed to  
120 the evolutionary history of AS in vertebrates remains largely unclear. Recently, a study using  
121 millions of synthetic mini-genes with degenerated subsequences demonstrated that the likelihood  
122 of AS decreases exponentially with increasing distance between constitutive and newly  
123 introduced alternative SS (Rosenberg et al., 2015), suggesting that sequence changes between  
124 constitutive and alternative SS might contribute to the rapid changes of AS among species. In  
125 plants, however, the detailed mechanisms that affect AS remain largely unclear (Reddy et al.,  
126 2013). Although it has been proposed that changes in chromatin features such as DNA  
127 methylation, histone marks as well as RNA structural features, and SREs are important in

128 regulating AS in plants, experimental evidence remains largely lacking (Reddy et al., 2013). A  
129 recent study shows that DNA methylation could affect AS in rice (Wang et al., 2016), indicating  
130 changes in DNA methylation can contribute to the variations of AS among species, however, this  
131 hypothesis has not been thoroughly tested.

132         Because AS regulation is a complex process involving many factors, computational  
133 modeling is a useful tool for identifying key factors and predicting the outcome of splicing.  
134 While the Bayesian neural network (BNN) method was developed for decoding the splicing code  
135 in mammals (Barbosa-Morais et al., 2012), deep learning approaches, which refers to methods  
136 that map data through multiple levels of abstraction, have recently been shown to surpass BNN-  
137 based approaches (Leung et al., 2014; Mamoshina et al., 2016). Furthermore, deep learning  
138 methods are also able to cope with large, heterogeneous and high-dimensional datasets and  
139 problems that include predicting DNA and RNA-binding proteins (Alipanahi et al., 2015) and  
140 AS (Leung et al., 2014; Mamoshina et al., 2016).

141         Here we performed a comparative analysis of the transcriptomes of both closely and  
142 distantly related plant species to explore the evolutionary history of AS in plants. To further  
143 understand the mechanisms underlying the AS evolution in plants, we applied a deep learning  
144 approach to investigate the determinants of AS and their effects on AS evolution. Specifically,  
145 we aimed to address the following questions in plants: 1) What are the evolutionary pattern of  
146 AS? 2) Are the AS events that are coupled with NMD more conserved than regular AS events? 3)  
147 Which factors are important in determining AS? 4) Which factors have contributed to the rapid  
148 turnover of AS between closely related plant species?

149

## 150 **RESULTS**

## 151 **Genome-wide AS patterns are species-specific in plants**

152 To provide an overview of AS evolution among different plant families, we studied the  
153 genome-wide AS in *Arabidopsis thaliana*, soybean (*Glycine max*), tomato (*Solanum*  
154 *lycopersicum*) and wild tobacco (*Nicotiana attenuata*), for which comparable transcriptomic data  
155 sets are available from the same tissues (roots, leaves and flowers) and represent a wide-range of  
156 eudicots. The overall distributions of different AS types within each species are consistent with  
157 previous studies. In all investigated species, intron retention (IR) and alternative 3' acceptor site  
158 (AltA) are the two major AS types (Figure S1) (Aoki et al., 2010; Marquez et al., 2012; Shen et  
159 al., 2014; Ling et al., 2015).

160 To investigate the evolutionary patterns of AS, we compared AS profiles across selected  
161 tissues and species. Because sequencing depth is known to strongly affect AS detection, we  
162 randomly subsampled 16 million (the lowest depth among all samples) uniquely mapped reads  
163 from each sample to standardize for the heterogeneity of sequencing depths. All downstream  
164 comparative analyses were based on this subsampled dataset. Clustering analyses using percent  
165 spliced index (PSI) that measures the qualitative differences of AS among samples showed that  
166 different tissues of the same species are more similar to each other than the same tissue from  
167 different species (Figure 1A). This pattern is consistent with the hypothesis that AS evolves  
168 rapidly in plants. Using qualitative measures of AS that consider the presence or absence of AS  
169 (binary) for all one-to-one orthologous genes, the same species-specific clustering pattern was  
170 found (Figure 1B). Furthermore, consistent results were also obtained when each type of AS was  
171 analyzed separately (Figure S2).

172 To further investigate the evolutionary patterns of AS among closely related species, we  
173 analyzed a recently published transcriptome dataset from three Brassicaceae species (*A. thaliana*,

174 *Arabidopsis lyrata* and *Capsella rubella*), each of which have comparable transcriptome data  
175 from two tissues (root and shoot) and two treatments (control and cold treated) (Seymour et al.,  
176 2014). Using both quantitative (PSI) and qualitative measures (binary) of AS, a similar species-  
177 specific clustering pattern was observed (Figure 1C and D). Interestingly, within same species  
178 and same tissue, samples exposed to cold stress clustered together regarding levels of PSI, a  
179 result which is consistent with previous studies that demonstrate that stresses can induce  
180 genome-wide AS responses (Li et al., 2013; Ding et al., 2014; Ling et al., 2015).

181

182 **Genome-wide AS regulations diverge faster than gene expression among closely related**  
183 **species.**

184 Species-specific clustering patterns were also reported at the level of gene expression  
185 (GE) among *A. thaliana*, rice and maize (Yang and Wang, 2013). To examine whether species-  
186 specific AS clustering results from gene expression divergences, we compared the divergence  
187 patterns of AS and GE among transcriptomes of different species. Comparisons among species  
188 from different plant families showed that both GE and AS cluster in species-specific patterns  
189 (Figure 1A and B, S3A and B). However, when species from the same plant family are compared,  
190 such as tomato and *N. attenuata* (Solanaceae), the species-specific AS pattern remained (Figure  
191 1A and B), but the GE data clustered in tissues-specific pattern (Figure S3A and B). This  
192 suggests that the expression profiles of the same tissues from different species are more similar  
193 to each other than the expression patterns from different tissues of the same species. This result  
194 was also found in the expression profiles of tissue samples from the three Brassicaceae species,  
195 among which the expression profiles of shoots and roots from different species were clearly  
196 separated (Figure S3C and D). These results indicate that AS evolves faster than GE in plants,

197 the pattern of which is similar to that in animals (Barbosa-Morais et al., 2012; Merkin et al.,  
198 2012).

199

### 200 **Rapid gains and losses of AS among different species**

201 Species-specific clustering of AS pattern suggests low level of AS conservation among  
202 species. Overall, among 3,857 one-to-one orthologous genes among the four dicots that have AS  
203 in at least one species, only ~7% of them have AS in all four species, while ~41% of them have  
204 species-specific AS. The rapid change of AS could result from the rapid loss or gain of EEJ  
205 between species. To exam the conservation of EEJs, we compared the EEJ structures among  
206 pairwise orthologous genes. In total, 60% of EEJs are conserved in at least two species (while  
207 only ~12% for AS), and the analysis based on AS events from the most conserved EEJs (found  
208 in all four species) showed that 92% of these are species-specific (Figure S4A). A similar pattern  
209 was found based on the analysis of EEJs that are conserved in at least two species, in which only  
210 10 AS events (0.25%, out of 4,015) were found conserved among all four species (Figure S4B).  
211 We also performed the same analysis within the three Brassicaceae species and found 69% of  
212 total EEJs are conserved in at least two species (while only 27% of AS are conserved) and 72%  
213 of AS events that were found at the EEJs shared among all three species were species-specific  
214 (Figure S5A). Furthermore, only ~8% of AS events (1,476 out of 19,170) are conserved among  
215 all three species (Figure S5B). Together, the results between divergent species and closely  
216 related species consistently suggest that AS diverge rapidly in plants.

217 To investigate the transition spectrum of AS at the conserved EEJs between each species  
218 pairs, we calculated the transitions among different types of AS. Among the four dicots, while  
219 the transitions among different AS types are rare, the gain/loss of AS was found to be the most

220 abundant transition type among all comparisons (Figure 2A-F). Among different AS types, AltA  
221 and ES are the most and least conserved AS, respectively. Similar patterns were observed among  
222 three closely related species in Brassicaceae (Figure S6A-C). These results suggest that the  
223 species-specific AS pattern is largely not due to the rapid changes of EEJs among species, but  
224 rather the rapid species-specific gains and losses of AS.

225

226 **The group of AS that generate PTC-containing alternative transcripts is more conserved**  
227 **than others**

228 Previous studies suggest that many pre-mRNAs underwent unproductive AS, which  
229 generates transcripts with in-frame PTCs that are coupled with nonsense-mediated decay (NMD)  
230 in plants (Kalyna et al., 2012; Drechsel et al., 2013). To investigate whether unproductive AS  
231 can affect the AS conservation and contribute to the rapid loss/gain of AS among different plant  
232 species, we separated the AS into two groups: (1) PTC+ AS and (2) PTC- AS (details see  
233 Materials and Methods). Overall, the portion of PTC+ AS ranges from 9% - 15% among the four  
234 dicots (Figure S7), suggesting that only a small portion of AS generated PTC-containing  
235 transcripts. Comparing the levels of conservation between tomato and *N. attenuata*, we found the  
236 PTC+ AS is significantly more conserved than PTC- AS ( $P < 0.02$ , Figure 3A). For example,  
237 among nine PTC+ AS of *N. attenuata* which are both conserved and have PTC information in  
238 tomato, eight of them (89%) also generated PTC+ transcripts in tomato.

239 To further investigate the level of conservations of PTC+ AS, we extended our analysis  
240 by adding the transcriptome data of a very ancient plant species, the spreading earth moss  
241 (*Physcomitrella patens*). We focused on the 10 most highly conserved AS events found in all  
242 four dicot plants (Figure S4B) and checked their presence in moss. In total, we found six AS

243 events that were also present in moss, indicating these AS events might have evolved since land  
244 plants and played essential functions in plants. Interestingly, two of these ultra-conserved AS  
245 events were from serine/arginine-rich (SR) genes (*RS2Z33*-like and *RS40*-like), which are part of  
246 RNA splicing machinery and the *RS2Z33*-like gene also has AS in rice (Iida and Go, 2006).  
247 Analyzing the protein coding potential of the transcripts generated by these six ultra-conserved  
248 AS events showed that five resulted in PTC+ transcripts. For example, the AS events of *RS2Z33*-  
249 like and *RS40*-like genes result in PTC+ alternative transcripts in all five species and are likely  
250 the targets of NMD (Figure 3B). To further investigate whether these PTC+ transcripts are  
251 regulated by NMD, we analyzed the available transcriptome data from *A. thaliana* wild type  
252 (WT) and NMD-deficient (*lba1* and *upf3-1* double mutant) plants (Drechsel et al., 2013). Among  
253 all five PTC+ transcripts in *A. thaliana*, three showed significantly higher expression in NMD-  
254 deficient plants (Drechsel et al., 2013) ( $P < 7e-06$ ), including *RS2Z33*-like and *RS40*-like genes  
255 (Figure 3B). Together, these results suggest that AS coupled with PTC is more conserved than  
256 regular AS and some of these AS-PTC pairs may play essential roles in plants.

257

### 258 **Mechanisms involved in determining AS are overall conserved among different plant** 259 **species**

260 To further understand the mechanisms that contributed to the rapid turnover of AS among  
261 species, it is necessary to identify the key features of AS in plants, which was largely unknown  
262 (Reddy et al., 2013). Because splicing is often mediated by SS, we were interested in whether the  
263 SS were different between constitutively and alternatively spliced junctions. Comparisons of the  
264 SS and their surrounding 12 bp sequences between constitutively and alternatively spliced  
265 junctions revealed that their SS are overall very similar (Figure S8). Furthermore, we separately

266 identified sequence motifs (12-mer) that are enriched in 5' and 3' splice sites (SS) compared to  
267 random sequences and found that these identified motifs are also highly conserved among  
268 studied species (Figure S9).

269         When alternative SS is present, the distance between the regular and the nearest  
270 alternative SS and inter-GT/AG, splicing junction size and the strength of the alternative SS are  
271 also important for the regulation of different AS types (Gopal et al., 2005; Kandul and Noor,  
272 2009; Braunschweig et al., 2014; Rosenberg et al., 2015). For the different AS types, we  
273 compared these features from both constitutively and alternatively spliced junctions. Because  
274 exon skipping (ES) events are rare in all species, we only studied the three most abundant AS  
275 types (AltD, AltA and IR). The results showed that for a given junction, while the likelihood of  
276 both AltD and AltA decreases with the distance between authentic and alternative SS as well as  
277 the distance between authentic SS and the nearest internal GT/AG, the likelihood of both AltD  
278 and AltA increases with junction size (Figure 4A and B). Interestingly, although the likelihood of  
279 IR in smaller junctions appears larger than in large junctions, no significant correlation with  
280 junction size was found (Figure S10A). Both 5' and 3' SS of the junction with IR are  
281 significantly weaker than constitutive junction (Figure S10B).

282         Furthermore, the presence/absence of UA-rich tract, polypyrimidine (PPT) tract, branch  
283 site (BS) is also known to be associated with 3' splicing recognition in eukaryotes  
284 (Lewandowska et al., 2004; Fu and Ares, 2014). We compared the frequency of AltA and IR  
285 between junctions of the AS gene with and without the presence of UA, PPT tract and BS within  
286 100 bp upstream of 3' SS. We found the frequencies of both AltA and IR are significantly higher  
287 in junctions without UA and PPT than junctions with them, while the presence of BS had no  
288 significant effect (Figure S11).

289 *Cis*-regulatory elements, including enhancers and silencers near SS are also important for  
290 the regulations of splicing. To identify these candidate regulatory elements, we performed *de*  
291 *novo* hexamer motif enrichment analysis by comparing 50 bp sequences from 5' and 3' sides of  
292 both donor and acceptor sites between alternatively spliced and constitutively spliced junctions.  
293 The results showed that most of the putative enhancer motifs for alternatively spliced junctions  
294 are highly similar to the identified SS. In addition, we also identified several putative silencer  
295 motifs (range from 5-10 for AltD and 10-18 for AltA in the five species), which were  
296 significantly more enriched in constitutively spliced junctions than alternatively spliced junctions  
297 (Figure S12 A and B).

298 To evaluate whether these identified features represent the AS determinants, we used a  
299 machine learning approach and modeled the different types of AS in each of the studied species.  
300 The rationale for this approach is that if the features we identified as representative of the key AS  
301 determinants were accurate, we would be able to predict whether an exon-intron junction is  
302 constitutively or alternatively spliced based on their quantitative or qualitative information. For  
303 this, we combined all of extracted featured mentioned above. In addition, we also extracted  
304 information on whether the alternative SS would introduce a frameshift, which may result in  
305 premature terminate code (PTC) or open reading frames, the number of reads that support the  
306 junction, which represent levels of expression that is known to be associated with AS, as well as  
307 the presence and absence of the identified *cis*-motifs. Using this information, our model achieved  
308 high precision and specificity for both AltD and AltA in all five species (Figure 4C and D, S13A  
309 and B), which suggests that the identified features can provide sufficient information to  
310 discriminate AS junctions from constitutively spliced junctions. However, for IR, the extracted

311 features were not predictively useful (the average AUC is 0.54), indicating that additional  
312 undetected factors have contributed to the determination of IR.

313         This modeling approach further provides indicative information on the relative  
314 importance of each feature to the prediction model. The results showed that for AltD, the  
315 distance between the authentic SS and the nearest alternative 5' SS or inter GT, the junction size  
316 and presence/absence of 5' additional SS in the intron are among the top important features for  
317 the prediction in all species (Supplemental file 1). In addition, the frame shifts introduced by the  
318 nearest alternative 5' SS and nearest GT were also important contributors to the model  
319 (Supplemental file 2). For AltA, the distance to the nearest inter-AG dinucleotide is the top  
320 feature for the prediction among all five species. Interestingly, all of the identified putative  
321 silencers only had a marginal role for the predictions of both AltD and AltA (Supplemental file  
322 2). Together, these results showed that the mechanisms regulating AltD and AltA are likely  
323 overall conserved among the studied species.

324

### 325 **Changes in AS determinants contributed to the rapid turnover of AS in plants**

326         The relatively conserved AS regulation mechanisms among studied species provide a  
327 foundation for investigating the mechanisms that contributed to a rapid turnover of AS among  
328 closely related plant species. We hypothesized that the changes in the identified AS determinants  
329 among species resulted in a rapid divergence of AS in plants. To test this, we associated the  
330 changes of the identified AS determinants and AS conservation among closely related species.  
331 Because we did not find determinants for IR, we only focused on the evolution of AltA and AltD.

332         Variation in the distance between authentic SS and alternative SS or inter-GT/AG were  
333 negatively associated with AS conservation: the levels of AS conservation decreased with

334 increasing distance in all three pairs of comparisons (Figure 5A and B), for both AltD and AltA.  
335 In addition, the changes in the reading frame introduced by the alternative SS also significantly  
336 decreased the conservations of both AltA and AltD (Figure 5C and D). The similar pattern was  
337 also found for the distance between authentic SS and the nearest inter-GT/AG (Figure 5E-H).

338 Variation in the *cis*-regulatory elements (CREs) UA-tract, PPT and BS (Figure S11 and  
339 Supplemental file 2) significantly reduced the conservation for AltA (Figure S14B), but did not  
340 affect the conservation of AltD among species (Figure S14A). This result is consistent with the  
341 functional roles of these CREs in regulating AltA.

342 To further systematically analyze different factors that might affect the conservation of  
343 AS, we constructed an AS evolution model for each closely related species pair using a deep  
344 learning method. In addition to the key AS determinants identified in this study, we also included  
345 several other features that were previously hypothesized to be important for AS conservation  
346 between species in the model, such as changes in copy numbers (role of gene duplications),  
347 transposable element (TE) insertion within the junction, GC-content and sequence similarity of  
348 SS. For the AltD, all three models between species pairs achieved significantly better prediction  
349 than by chance (highest *P*-value =  $3e-44$ ), with an average precision of 0.63 and specificity of  
350 0.82. In all three pairwise comparison models, the distance changes between authentic and  
351 nearest alternative 5' SS or inter-GT/AG are among the top five important features (Figure S15A  
352 and Supplemental file 1). For AltA, all three models achieved a precision and specificity  
353 (average 0.70 and 0.85, respectively) that was significantly higher than by chance (highest *P*-  
354 value =  $3e-145$ ). In all three models, distance changes between authentic SS and the nearest  
355 inter-AG or alternative 3' SS and the changes on CREs (UA and PPT tracts) represent the top

356 five most important features that contributed to the model predictions (Figure S15B and  
357 Supplemental file 1).

358 Interestingly, we found TE insertions to also be an important factor that reduced the  
359 conservation of both AltD and AltA between *N. attenuata* and tomato but not between any of  
360 two Brassicaceae species (Figure S15A and B). This is likely due to the difference of TE  
361 abundance between *N. attenuata* (~63%) and tomato (~81%), values which are much higher than  
362 the differences between *A. thaliana* (~23%) and *A. lyrata* (28%) (Hu et al., 2011; Tomato  
363 Genome, 2012). Furthermore, we also analyzed the impact of DNA methylation changes  
364 between *A. thaliana* and *A. lyrata* using data from (Seymour et al., 2014) and found no  
365 significant effects (Figure S15).

366

## 367 **DISCUSSION**

368 Here we showed that alternative splicing (AS) diverges more rapidly than does gene  
369 expression (GE) and the rapid gain and loss of AS resulted in lineage-specific AS profiles in  
370 plants. Although AS events that introduce premature termination codons (PTC), represent only a  
371 small proportion of the total AS events, they are more conserved than AS events that do not  
372 introduce PTC (Figure 3A). Consistently, several AS events that generate PTC-containing  
373 transcripts were ultra-conserved among highly divergent plants. To understand the mechanisms  
374 that resulted in a rapid turnover of AS between closely related species, we identified several key  
375 determinants for both alternative donor (AltD) and alternative acceptor (AltA) splicing, and  
376 found that the change of these key determinants between species significantly contributed to the  
377 rapid gain and loss of AS in plants.

378           In this analysis, we observed a dominant species-specific pattern of AS among different  
379 species, suggesting that AS in plants diverges rapidly (Figure 1). Such rapid evolution of AS in  
380 plants is similar to the pattern found among vertebrate species that span ~350 million years of  
381 evolution (Barbosa-Morais et al., 2012; Merkin et al., 2012), in which AS is largely segregated  
382 by species, while GE is segregated by tissue types (Barbosa-Morais et al., 2012; Merkin et al.,  
383 2012). This indicates that the rapid evolution of AS might be universal among eukaryotes.  
384 Interesting, in plants, the evolution of GE appears to be faster than in vertebrates, as the tissue-  
385 dominant clustering of GE was only observed among closely related species, but not among  
386 species from different families (Figure S3). This pattern is consistent with a previous study  
387 which showed the overall GE of three highly divergent species, including both monocots and  
388 dicots (diverge ~200 million years ago), are grouped according to species rather than organs  
389 (Yang and Wang, 2013). In vertebrates, some tissues, such as brain, testis, heart and muscle still  
390 showed a strong tissue-specific splicing signature, despite the dominant species-specific splicing  
391 background (Barbosa-Morais et al., 2012; Merkin et al., 2012). Although all three tissues (root,  
392 leaves and flowers) used in our study did not show such strong tissue-specific splicing signatures,  
393 some other plant tissues might. For example, the transcriptomes of sexual tissues are  
394 substantially different from those of vegetative tissues, and anthers harbor the most diverged  
395 specialized metabolomes (Yang and Wang, 2013; Li et al., 2016). Future studies that include  
396 transcriptome data of much more fine-scaled tissue samples will provide new insights on this  
397 aspect.

398           We found that the AS resulted in transcripts with PTC, which is likely coupled with  
399 nonsense-mediated decay (NMD) for degradation is more conserved than the AS that do not  
400 generate PTC-containing transcripts in plants (Figure 3A). Consistently, among six ultra-

401 conserved AS events across different plant species including the spreading earth moss, five  
402 produced PTC+ transcripts, indicating that PTC+ AS might be more important than it was  
403 previously thought. Previous studies showed that all human serine/arginine-rich (SR) genes and  
404 some SR genes in plants produce AS resulted in PTC+ transcripts (Kalyna et al., 2006; Lareau et  
405 al., 2007; Palusa and Reddy, 2010). Furthermore, the junction regions that contain PTC+ AS in  
406 numerous splicing factors (SFs) are ultra-conserved between different kingdoms and the loss of  
407 the ancient PTC+ AS in paralogs through gene duplications were repeatedly replaced by newly  
408 created distinct unproductive splicing (Lareau and Brenner, 2015). In the same line, our results  
409 showed that many of these PTC+ AS are likely functionally important and are consistent with the  
410 hypothesis that the unproductive splicing coupled with NMD can be a functional process that  
411 controls the abundance of active proteins at a post-transcriptional level.

412         Among all five plant species, the distance between the 5'/3' nearest alternative splice  
413 sites (SS) and the authentic SS is the main determinant that distinguishes AltD/AltA from  
414 constitutive splicing (Figure 4 and Supplemental file 1). For a given spliced junction, the  
415 likelihood of AS decreases with increased distance between the authentic and nearest alternative  
416 SS (Figure 4A and B). Interestingly, similar patterns were also found in mammals, in which, the  
417 closer the alternative SS was to the authentic SS, the more likely it was used for AS (Dou et al.,  
418 2006; Rosenberg et al., 2015). Interestingly, the frequency of AltA also decreases with the  
419 increased distance between the authentic SS and nearest inter-AG dinucleotide. This result is  
420 consistent with the pattern found in humans in that only closely located AGs (< 6 nt) can  
421 efficiently compete with the authentic SS and the distance between branch site (BS) and the first  
422 downstream AG can affect the 3' SS selection (Chiara et al., 1997; Chua and Reed, 2001).  
423 Although, the BS in plants is not well studied and BS was not identified in ~30% of junctions

424 (Reddy, 2007), the similar effects of inter-AG distance on AltA in both plants and mammals  
425 indicates that the mechanisms of generating AS, at least for AltA, might be similar between these  
426 two kingdoms.

427         While the deep learning model for AltA achieved high precision and specificity among  
428 five species (AUC > 0.9), the models for AltD performed less well than those for AltA, although  
429 still performing much better than by chance (AUC < 0.8, Figure 4C and D, S13). This indicates  
430 that additional determinants that contribute to the regulations of AltD were not detected by our  
431 method. It is known that the mechanisms involved in AltD are more complex than AltA. For  
432 example, in both human and mouse, while both the presence and quantity of exon splicing  
433 enhancer (ESE) and exon splicing silencer (ESS) are important for generating AltD (Koren et al.,  
434 2007), AltA is mainly affected by the competition of closely located AG dinucleotide by a  
435 scanning mechanism for the downstream sequence of the BS-polypyrimidine tract (PPT) (Smith  
436 et al., 1989; Smith et al., 1993; Chiara et al., 1997; Chua and Reed, 2001). These results suggest  
437 that splicing regulatory elements (SREs) may play more important roles in the proper selection  
438 of alternative SS in AltD than AltA. This may explain why the junction size contributed more in  
439 the AltD model than in the AltA model (Supplemental file 1), since larger junction size might  
440 increase the likelihood of introducing intronic SREs. In our attempts to identify SREs, although a  
441 few candidate sequence motifs were identified using the enrichment analysis, none of them  
442 significantly contributed to the model predictions. Two non-exclusive possibilities may explain  
443 this failure. First, the identified motifs are not involved in splicing regulation processes, although  
444 their density was significantly different between constitutively and alternatively spliced junctions.  
445 Second, they might be involved in tissue-specific regulations of AS, which likely did not  
446 contribute to the overall AltD prediction based on all three tissues. Future studies using different

447 approaches to investigate the alteration of AS by introducing millions of random hexamers into  
448 specific regions of a gene junction in a plant then measuring the consequences of splicing, may  
449 allow us to more reliably detect splicing regulators of AltD in plants.

450         Although we found both junction size and SS for IR junctions are different from the  
451 constitutively spliced junctions (Figure S10), the identified features did not improve the AS  
452 prediction from that occurring by chance, indicating some other key determinants for IR remain  
453 missing in the model. As the expression level of IR is usually low and therefore requires high  
454 sequencing depth for their detection (Figure S16), it is likely that the sequencing depth of the  
455 transcriptome data used in this study was not sufficient to detect all of the IR junctions.  
456 Therefore, many true IR junctions may not have been considered as IRs in our analysis, which  
457 reduced prediction precision and power.

458         For both AltA and AltD, their rapid evolution between closely related species were  
459 mainly due to variations in the key sequence determinants near the SS (Figure 5, S14 and S15)  
460 and the key sequence determinants such as distance to authentic SS and *cis*-elements (BS, PPT,  
461 UA-rich tract for AltA) are all located within intronic regions. Intron sequences diverge rapidly  
462 (Mattick, 1994; Hare and Palumbi, 2003), therefore, the process of which likely have contributed  
463 to the rapid gains and losses of AS among different lineages to produce species-specific AS  
464 profiles in plants. For example, a decreased distance between alternative SS and authentic SS as  
465 a result of a short deletion of the intron sequence could lead to a gain of AS at the junction, and  
466 as consequence is likely to be shared among different tissues. Consistently, in vertebrates, the  
467 mutations that affect intronic splicing regulatory elements (SREs) were shown to be the main  
468 factor that resulted in the dominant species-specific splicing pattern (Merkin et al., 2012).  
469 However, our data can not exclude the possibility that the species-specific trans-factors, such as

470 SR protein family, which have distinct numbers of homologues among species (Figure S17) (Iida  
471 and Go, 2006; Isshiki et al., 2006; Ling et al., 2015), may have also contributed to the divergence  
472 of AS among different species (Ast, 2004; Barbosa-Morais et al., 2012).

473 We also investigated other factors that were hypothesized to affect AS evolution, such as  
474 gene duplication (GD), DNA methylation and transposable element (TE) insertion (Sorek et al.,  
475 2002; Su et al., 2006; Flores et al., 2012). However, with the exception of TE insertions, the  
476 effects of which were found to be species-specific, most of the tested factors did not show  
477 significant effects on the levels of AS conservation between closely related species (Figure S15  
478 and Supplemental file 2). The species-specific effects of TE on the AS conservation were likely  
479 due to the different abundance of TE insertions in the genomes of different species (Hu et al.,  
480 2011; Tomato Genome, 2012; Slotte et al., 2013; Sierro et al., 2014), suggesting genomic  
481 composition of each species might also affect the evolutionary alteration of AS.

482

## 483 **CONCLUSIONS**

484 We found that the AS profiles diverged rapidly in plants, which is largely due to rapid  
485 gains and losses of AS in each lineage, while a group of AS that generate PTC-containing  
486 transcripts is highly conserved among very distantly related plants. The alteration of a few key  
487 sequence determinants of AltA and AltD, all located in the intron region, contributed to the rapid  
488 divergence of AS among closely related plant species. These results provide mechanistic insights  
489 into the evolution of AS in plants and highlight the role of post-transcriptional regulation of a  
490 plant's responses to environment interactions.

491

## 492 **MATERIALS AND METHODS**

## 493 **Read mapping, transcripts assembly and abundance estimation**

494           The raw sequence reads were trimmed using AdapterRemoval (v1.1) (Lindgreen, 2012)  
495 with parameters “--collapse --trimns --trimqualities 2 --minlength 36”. The trimmed reads from  
496 each species were then aligned to the respective reference genome using Tophat2 (v2.0.6)  
497 (Trapnell et al., 2009), with maximum and minimum intron size set to 50,000 and 41 bp,  
498 respectively. The numbers of uniquely mapped reads and splice junctions (SJs) mapped reads  
499 were then counted using SAMtools (v0.1.19) (Li et al., 2009) by searching “50” in the MAPQ  
500 string and “\*N\*” flag in the CIGAR string of the resulting BAM files. The uniquely mapped  
501 reads from each sample were sub-sampled into same sequencing depth (16 million) using  
502 SAMtools (v0.1.19) (Li et al., 2009). The mapping information and IDs of all download datasets  
503 deposited in Sequence Read Archive (SRA) (<http://www.ncbi.nlm.nih.gov/sra>) were listed in  
504 Table S1 and S2.

505           The transcripts of each species were assembled using Cufflinks (v2.2.0) (Trapnell et al.,  
506 2012) with the genome annotation as the reference. The open reading frame (ORF) of each  
507 transcript was analyzed using TransDecoder from TRINITY (v2.1.0) (Grabherr et al., 2011). To  
508 estimate the expression level of genes/transcripts, all trimmed reads were re-mapped to the  
509 assembled transcripts using RSEM (v1.2.8) (Grabherr et al., 2011). Transcripts per million (TPM)  
510 was calculated for each gene/transcript (Wagner et al., 2012). Only genes with TPM greater than  
511 five in at least one sample were considered as an expressed gene.

512

## 513 **AS detection**

514           All AS analysis were based on splicing junctions obtained from the BAM files produced  
515 by Tophat2. To remove the false positive junctions that were likely due to non-specific or

516 erroneous alignments, all original junctions were filtered based on overhang sizes greater than 13  
517 bp, as suggested in (Ling et al., 2015). All filtered junctions were then used for AS identification  
518 and annotation using JUNCBASE v0.6 (Graveley et al., 2011). Due to the relatively low  
519 sequence depth of each individual sample of Brassicaceae RNA-seq data (Seymour et al., 2014)  
520 (Table S2), we merged the BAM files of each three replicates together and random subsampled  
521 17 million (the lowest depth among all merged samples) unique mapped reads from each merged  
522 to avoid the heterogeneity of sequencing depth.

523         The percent spliced index (PSI) of each AS event, which represents the relative ratio of  
524 two different isoforms generated by the AS was calculated in each sample.  $PSI = (\text{number of}$   
525  $\text{reads to inclusion isoform}) / (\text{number of reads to inclusion isoform} + \text{number of reads to}$   
526  $\text{exclusion isoform})$  as suggested in (Graveley et al., 2011). To avoid false-positives, PSI was  
527 calculated only for AS events that had a total read count equal or greater than ten.

528

### 529 **Identification of conserved exon-exon junctions (EEJs) and AS**

530         We separately extracted 100 bp sequence from the flanking upstream exon and  
531 downstream exon of each junction that have mapped read to support, and combine each side of  
532 exon sequence (in total 200 bp sequence) to represent the EEJ. The sequences of all EEJs from  
533 two species were searched against with each other using TBLASTX (v.2.2.25) (Altschul et al.,  
534 1990) to find homologous relationships (Figure S18). A python script was used to filter the  
535 TBLASTX results based on the following requirements: (1) The gene pair contain the EEJs must  
536 be the orthologous gene pair between the two species; (2) the EEJ sequences between two  
537 species must be the best reciprocal blast hit based on the bit score; (3) at least 3 amino acid (aa)

538 from both the flanking upstream exon and downstream exon sequence were aligned and (4)  
539 alignment coverage  $\geq 60$  bp, (5) E-value  $< 1E-3$ .

540 We only consider an AS event to be conserved if the same type of AS was found on the  
541 conserved EEJs between two plant species.

542

### 543 **Identification of AS that generates premature termination codons (PTC)**

544 The junctions related to each AS event were mapped back to assembled transcripts; only  
545 AS which related to junctions that mapped to two unique transcripts (had no structural difference  
546 except the AS region) were retained to avoid the situation where the sequence differences of the  
547 two transcripts resulted from multiple AS events. The transcript was considered to have a PTC if  
548 the stop codon of the longest ORF is at least 50 nucleotides upstream of an exon-exon boundary  
549 (the 50 nucleotides rule) (Nagy and Maquat, 1998). The PTC-generating AS events are defined  
550 as only one of the resulting transcripts contain PTC.

551

### 552 **One-to-one orthologous gene identifications and gene family size estimation**

553 One-to-one orthologous gene pairs were predicted based on pair-wise sequence  
554 similarities between species of the corresponding dataset. First, we calculated the sequence  
555 similarities between all protein-coding genes using BLASTP for the selected species and filtered  
556 the results based on E-value less than  $1E-6$ . Second, we selected the groups of genes that  
557 represent the best reciprocal hits that are shared among all species from the corresponding  
558 dataset.

559 For calculating the gene family size, we first defined gene families among different  
560 species by using a similarity-based approach. To do so, we used all genes that were predicted

561 from the respective genomes of each species. In brief, all-vs-all BLASTP was used to compare  
562 the sequence similarity of all protein coding genes, and the results were filtered based on the  
563 following criteria: E-value less than 1E-20; match length greater than 60 amino acids; sequence  
564 coverage greater than 60% and identity greater than 50%. All BLASTP results that remained  
565 after filtering were clustered into gene families using the Markov cluster algorithm (mcl)  
566 (Enright et al., 2002). The gene family size for a species is represented by the number of genes of  
567 this species within the corresponding gene family.

568

### 569 **Correlation and clustering**

570 For the pairwise comparison of AS, Spearman correlation and binary distance was  
571 applied to the PSI data ( $0.05 < \text{PSI} < 0.95$  in at least one sample) and binary data (all genes that  
572 had no AS in all four species were excluded), respectively. A non-parametric correlation was  
573 selected for PSI level because of its bimodal nature distribution (0 and 100). For the pairwise  
574 comparison of GE, Pearson correlation was applied to  $\log_2(\text{TPM}+1)$  of expressed genes to avoid  
575 infinite values.

576 The R package “pvcluster” was used for clustering of samples with 1,000 bootstrap  
577 replications. When we clustered and performed principal-component analysis (PCA) of gene  
578 expression, the TPM values were normalized by GC% (EDASeq package in R) and TMM (the  
579 trimmed mean of M-values).

580

### 581 **Identification of possible alternative splice sites (SS) and regulatory sequences**

582 The 5' and 3' splice site including 5 bp up and downstream sequences of all EEJs were  
583 used as the positive dataset, while the sequences extracted using the same method for all inter-

584 GT (for 5' splice site) and inter-AG (for 3' splice site) within junction regions were used as  
585 background dataset. The putative SS motifs (6-mer) of both 5' and 3' SS were separately  
586 identified using Homer V3.12 (Heinz et al., 2010) and only motifs present in at least 5% of total  
587 positive sequences and  $P$ -value $<1E-20$  were kept. The appearance of putative SS was identified  
588 using scanMotifGenomeWide, a Perl script included in the Homer toolkits and only sequence  
589 regions with match score  $>2$  were kept.

590 Homer was also used to identify the putative regulatory intronic and exonic sequence  
591 motifs (6-mer) of AltD, AltA and IR. The 50 bp up and downstream sequence of 5' SS was  
592 regarded as exonic and intronic sequence and vice versa for 3' SS. For AltD and AltA, the  
593 related sequences of EEJs with AS were used as the positive dataset, while 10,000 related  
594 sequences of EEJs without AS by random selection (due to a large number of sequences) were  
595 used as background dataset. The enriched motifs in the positive dataset were regarded as splicing  
596 enhancers, while the enriched motifs in the negative dataset were considered as splicing silencers.  
597 For IR, the related sequences from both splice sites of EEJs with IR were used as the positive  
598 dataset and the same sequences from EEJs without IR were used as background dataset. The  
599 conserved motifs between species were identified using compareMotifs, a Perl script included in  
600 the Homer toolkits and only one mismatch was allowed. To identify PPT, UA-rich tracts and  
601 branch site (BS) of each EEJ, we used the Perl scripts from Szczesnia et al. (Schwartz et al.,  
602 2008; Szczesniak et al., 2013) To estimate the effect of each putative sequence motif, PPT and  
603 UA-tracts, we calculated the AS frequency of EEJs containing or not containing the motif/tract  
604 (Rosenberg et al., 2015). Then for each motif/tract, the  $\log_2$  odds ratio (effect size) with and  
605 without the motif/tract were calculated to quantify to what extent the presence of the motif/tract  
606 increases or decreases the AS frequency compare to its absence:

$$Effect\ Size = \log_2 \frac{p(AS|motif)/(1 - p(AS|motif))}{p(AS|-motif)/(1 - p(AS|-motif))}$$

607

## 608 **Deciphering the splicing codes and AS conservation using deep learning algorithm**

609 To investigate which sequence determinants contributed to the AS in plants, we  
610 constructed multi-layer feed-forward artificial neural networks using H<sub>2</sub>O's deep learning  
611 algorithm ('h2o' package) in R 3.0.2 (R Development Core Team 2013). For each AS type, a  
612 matrix was created based on the information of all EEJs that contain the AS (only that single  
613 event) and other EEJs within the same gene. The AS status (either AS or constitutive) was  
614 considered as output and the features that were known to be associated with splicing recognition  
615 and regulation in eukaryotes (Lewandowska et al., 2004; Kandul and Noor, 2009; Rosenberg et  
616 al., 2015) (listed in Supplemental file 1) were used as input for training the model. To reduce the  
617 background noise, we removed the EEJs which were supported by less than five reads on  
618 average. In addition, because the number of constitutively spliced EEJs in all cases is much  
619 larger than alternatively spliced EEJs, we randomly selected the same number of constitutive  
620 spliced EEJs as alternative spliced EEJs and combined them together with all alternative spliced  
621 EEJs as the full dataset (50% precision by chance). To train and test the deep neural networks  
622 (DNN), the full dataset was randomly split, which 60% of data were used for training, 20% used  
623 for validation and the other 20% was reserved for testing. We trained for a fixed number (10,000)  
624 of epochs or stopped the training once the top 10 model were within 1% of improvement, and  
625 selected the hyper-parameters that gave the optimal AUC performance on the validation data.  
626 The model was then retrained using these selected hyper-parameters with the full dataset.

627 Using the similar approach, we constructed the model for AS conservation. For each AS  
628 type, a matrix was created based on the information of all orthologous EEJ pairs between two

629 species that contain the AS in at least one species. To reduce the background noise, any EEJ with  
630 multiple AS types, low number of support reads (less than five) or orthologous EEJ pair have  
631 different AS types were removed. The conservation levels (conserved, lost or gained in the other  
632 species) were used as the output of the model and the difference of features that were known to  
633 be important to AS and AS conservation (Su et al., 2006; Kelley et al., 2014; Li et al., 2014;  
634 Lambert et al., 2015; Rosenberg et al., 2015) (listed in Supplemental file 2) between two species  
635 were used as input to train the model. Yass v1.15 (Noe and Kucherov, 2005) was used to align  
636 the SS' flanking sequences (combined 50 bp upstream and downstream sequences of 5'/3' splice  
637 site, 100 bp in total) of each orthologous EEJ pair, the similarity was calculated as: (length of  
638 alignment - number of gaps - number of mismatches) / (total sequence length). To reduce the  
639 bias from different transition types in the dataset (much higher proportion of loss/gain than  
640 conserved AS), the data used to train the model was selected as the ratio of 1:1:1 for conserved,  
641 lost and raised situations (33.3% precision by chance). Due to the rather small sample size of  
642 conserved AS, the model based on the same original data may differ as the randomly selected  
643 data of AS lost/raised were different each time. Therefore, the model construction process was  
644 repeated 10 times and the models that achieved the highest AUC for the complete dataset were  
645 considered.

646

#### 647 **Accession Numbers**

648 The ID of all RNA-seq data that are deposited or downloaded from NCBI short reads  
649 archive (SRA) database for generating the results in this study is listed in Table S1.

650

#### 651 **ACKNOWLEDGMENTS**

652 We thank Danell Seymour and Daniel Koenig for providing the methylation data, Michal  
653 Szczesniak for providing the Perl scripts for finding UA tracts.

654

## 655 **SUPPLEMENTAL MATERIAL**

656 **Supplemental Figure S1.** The distribution of different types of alternative splicing (AS) events  
657 in (A) *A. thaliana*, (B) *G. max*, (C) *S. lycopersicum*, (D) *N. attenuate*.

658 **Supplemental Figure S2.** Species-specific clustering of alternative splicing (AS) among  
659 different plant species.

660 **Supplemental Figure S3.** Conservation of gene expression (GE) in eudicots.

661 **Supplemental Figure S4.** Comparative profiling of conserved junctions and alternative splicing  
662 in eudicots.

663 **Supplemental Figure S5.** Comparative profiling of conserved junction and alternative splicing  
664 (within one-to-one orthologues) in three Brassicaceae species.

665 **Supplemental Figure S6.** The transition spectrum among different types of AS between species  
666 pairs.

667 **Supplemental Figure S7.** The proportion of AS that generates PTC (PTC+) or not (PTC-) in *N.*  
668 *attenuata*, tomato, *A. thaliana* and soybean.

669 **Supplemental Figure S8.** The probability of DNA bases surrounding SS with different AS types  
670 compared to regular SS in five plant species.

671 **Supplemental Figure S9.** The complete linkage hierarchical clustering of SS motifs among  
672 different plant species.

673 **Supplemental Figure S10.** The determinants of intron retention (IR) in plants.

674 **Supplemental Figure S11.** The effect of UA-rich, polypyrimidine tract (PPT) and branch site (BS) on  
675 alternative acceptor (AltA) and intron retention (IR) in plants.

676 **Supplemental Figure S12.** The effect size of conserved 6-mer motifs and features identified in  
677 (A) alternative 5' donor site (AltD) and (B) alternative 3' acceptor site (AltA) between species  
678 pairs in Solanaceae and Brassicaceae.

679 **Supplemental Figure S13.** The area under the curve (AUC) of deep learning models using  
680 different key features of (A) alternative 5' donor (AltD) and (B) alternative 3' acceptor (AltA) in  
681 tomato, *A. lyrata* and *C. rubella*.

682 **Supplemental Figure S14.** Differences in the *cis*-regulatory elements affect the turnover rate of  
683 (A) AltD and (B) AltA between plant species.

684 **Supplemental Figure S15.** Factors that affect the rapid turnover of AS between plant species.

685 **Supplemental Figure S16.** The composition of AS types detected in leaf and root samples of *N.*  
686 *attenuata* with different sequencing depth.

687 **Supplemental Figure S17.** Phylogenetic tree of SR and SR-like genes in moss and four eudicots.

688 **Supplemental Figure S18.** The diagram showing the process of identifying conserved exon-  
689 exon junctions (EEJs) between two species.

690 **Supplemental Table S1.** RNA-seq coverage and alignment statistics of the four eudicots

691 **Supplemental Table S2.** RNA-seq coverage and alignment statistics of Brassicaceae

692 **Supplemental file 1:** This file contains two excel sheets that show the relative importance of  
693 each factor that contributes to the splicing model of AltD and AltA in five different species. SS:  
694 splicing site; BS: branch site; PPT: polypyrimidine tract. The top five contributors to each model  
695 were highlighted in red.

696 **Supplemental file 2:** This file contains two excel sheets that show the relative importance of  
697 each factor that contributes to the AS conservation of AltD and AltA among closely related

698 species. SS: splicing site; BS: branch site; PPT: polypyrimidine tract. The top five contributors to  
699 each model were highlighted in red.

700

## 701 LITERATURE CITED

- 702 Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and  
703 RNA-binding proteins by deep learning. *Nat Biotechnol* 33: 831
- 704 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol*  
705 *Biol* 215: 403-410
- 706 Aoki K, Yano K, Suzuki A, Kawamura S, Sakurai N, Suda K, Kurabayashi A, Suzuki T, Tsugane T,  
707 Watanabe M, Ooga K, Torii M, Narita T, Shin IT, Kohara Y, Yamamoto N, Takahashi H,  
708 Watanabe Y, Egusa M, Kodama M, Ichinose Y, Kikuchi M, Fukushima S, Okabe A, Arie T, Sato  
709 Y, Yazawa K, Satoh S, Omura T, Ezura H, Shibata D (2010) Large-scale analysis of full-length  
710 cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the  
711 Solanaceae genomics. *BMC Genomics* 11: 210
- 712 Ast G (2004) How did alternative splicing evolve? *Nat Rev Genet* 5: 773-782
- 713 Baek JM, Han P, Iandolino A, Cook DR (2008) Characterization and comparison of intron structure and  
714 alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, *Arabidopsis* and rice.  
715 *Plant Mol Biol* 67: 499-510
- 716 Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the  
717 splicing code. *Nature* 465: 53-59
- 718 Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants:  
719 opportunities and challenges. *Genome Res* 18: 1381-1392
- 720 Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt  
721 S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ  
722 (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338:  
723 1587-1593
- 724 Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-  
725 336
- 726 Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T,  
727 Frey B, Irimia M, Blencowe BJ (2014) Widespread intron retention in mammals functionally  
728 tunes transcriptomes. *Genome Res* 24: 1774-1786
- 729 Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway.  
730 *Annu Rev Biochem* 76: 51-74
- 731 Chen M, Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and  
732 genomics approaches. *Nature reviews: Molecular cell biology* 10: 741-754
- 733 Chiara MD, Palandjian L, Feld Kramer R, Reed R (1997) Evidence that U5 snRNP recognizes the 3'  
734 splice site for catalytic step II in mammals. *EMBO J* 16: 4746-4759
- 735 Chua K, Reed R (2001) An upstream AG determines whether a downstream AG is selected during  
736 catalytic step II of splicing. *Mol Cell Biol* 21: 1509-1514
- 737 Darracq A, Adams KL (2013) Features of evolutionarily conserved alternative splicing events between  
738 *Brassica* and *Arabidopsis*. *New Phytol* 199: 252-263
- 739 Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L (2014) Genome-wide analysis of alternative splicing of  
740 pre-mRNA under salt stress in *Arabidopsis*. *BMC Genomics* 15: 431
- 741 Donahue CP, Muratore C, Wu JY, Kosik KS, Wolfe MS (2006) Stabilization of the tau exon 10 stem loop  
742 alters pre-mRNA splicing. *J Biol Chem* 281: 23302-23306

- 743 Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ (2006) Genomic splice-site analysis reveals frequent  
744 alternative splicing close to the dominant splice site. *RNA* 12: 2047-2056
- 745 Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, Ratsch G, Wachter A (2013)  
746 Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant  
747 of the *Arabidopsis* steady state transcriptome. *Plant Cell* 25: 3726-3742
- 748 Ellis JD, Barrios-Rodriguez M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim  
749 PM, Wrana JL, Blencowe BJ (2012) Tissue-specific alternative splicing remodels protein-protein  
750 interaction networks. *Mol Cell* 46: 884-892
- 751 Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of  
752 protein families. *Nucleic Acids Res* 30: 1575-1584
- 753 Flores K, Wolschin F, Corneveaux JJ, Allen AN, Huentelman MJ, Amdam GV (2012) Genome-wide  
754 association between DNA methylation and alternative splicing in an invertebrate. *BMC*  
755 *Genomics* 13: 480
- 756 Fu XD, Ares M, Jr. (2014) Context-dependent control of alternative splicing by RNA-binding proteins.  
757 *Nat Rev Genet* 15: 689-701
- 758 Gao H, Gordon-Kamm WJ, Lyznik LA (2004) ASF/SF2-like maize pre-mRNA splicing factors affect  
759 splice site utilization and their transcripts are alternatively spliced. *Gene* 339: 25-37
- 760 Gopal S, Awadalla S, Gaasterland T, Cross GA (2005) A computational investigation of kinetoplastid  
761 trans-splicing. *Genome Biol* 6: R95
- 762 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury  
763 R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW,  
764 Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly  
765 from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644-652
- 766 Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ,  
767 Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH,  
768 Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads  
769 B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH,  
770 Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P,  
771 Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE (2011) The developmental  
772 transcriptome of *Drosophila melanogaster*. *Nature* 471: 473-479
- 773 Hare MP, Palumbi SR (2003) High intron sequence conservation across three mammalian orders suggests  
774 functional constraints. *Mol Biol Evol* 20: 969-978
- 775 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK  
776 (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory  
777 elements required for macrophage and B cell identities. *Mol Cell* 38: 576-589
- 778 Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, He L, Winter-Sederoff H, Gassmann W,  
779 Veronese P, Heber S (2013) High-throughput RNA sequencing of pseudomonas-infected  
780 *Arabidopsis* reveals hidden transcriptome complexity and novel splice variants. *PLoS One* 8:  
781 e74183
- 782 Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J,  
783 Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottillar RP, Salamov AA, Schneeberger K,  
784 Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J,  
785 Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011) The  
786 *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43:  
787 476-481
- 788 Iida K, Go M (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in  
789 land plants. *Mol Biol Evol* 23: 1085-1094
- 790 Isshiki M, Tsumoto A, Shimamoto K (2006) The serine/arginine-rich protein family in rice plays  
791 important roles in constitutive and alternative splicing of pre-mRNA. *Plant Cell* 18: 146-158
- 792 Kalyna M, Lopato S, Voronin V, Barta A (2006) Evolutionary conservation and regulation of particular  
793 alternative splicing events in plant SR proteins. *Nucleic Acids Res* 34: 4395-4405

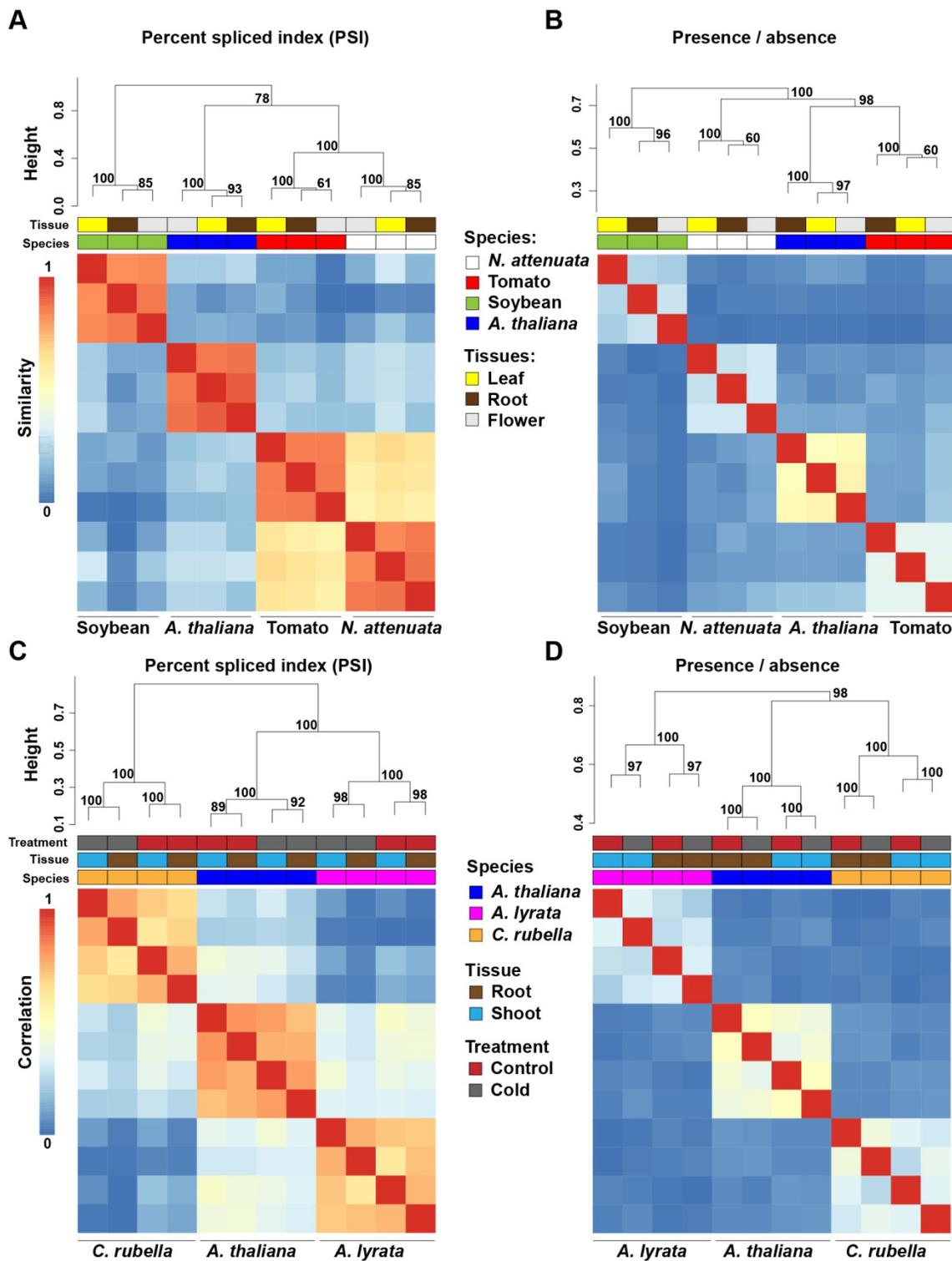
- 794 Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, Marshall J, Fuller J, Cardle  
795 L, McNicol J, Dinh HQ, Barta A, Brown JW (2012) Alternative splicing and nonsense-mediated  
796 decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res* 40:  
797 2454-2469
- 798 Kandul NP, Noor MA (2009) Large introns in relation to alternative splicing and gene evolution: a case  
799 study of *Drosophila bruno-3*. *BMC Genet* 10: 67
- 800 Kazan K (2003) Alternative splicing and proteome diversity in plants: the tip of the iceberg has just  
801 emerged. *Trends Plant Sci* 8: 468-471
- 802 Kelley DR, Hendrickson DG, Tenen D, Rinn JL (2014) Transposable elements modulate human RNA  
803 abundance and splicing via specific RNA-protein interactions. *Genome Biol* 15: 537
- 804 Kervestin S, Jacobson A (2012) NMD: a multifaceted response to premature translational termination.  
805 *Nat Rev Mol Cell Biol* 13: 700-712
- 806 Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin  
807 marking of introns and expressed exons by H3K36me3. *Nat Genet* 41: 376-381
- 808 Koren E, Lev-Maor G, Ast G (2007) The emergence of alternative 3' and 5' splice site exons from  
809 constitutive exons. *PLoS Comput Biol* 3: e95
- 810 Lambert MJ, Cochran WO, Wilde BM, Olsen KG, Cooper CD (2015) Evidence for widespread  
811 subfunctionalization of splice forms in vertebrate genomes. *Genome Res* 25: 624-632
- 812 Lareau LF, Brenner SE (2015) Regulation of splicing factors by alternative splicing and NMD is  
813 conserved between kingdoms yet evolutionarily flexible. *Mol Biol Evol* 32: 1072-1079
- 814 Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE (2007) Unproductive splicing of SR genes  
815 associated with highly conserved and ultraconserved DNA elements. *Nature* 446: 926-929
- 816 Leung MK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code.  
817 *Bioinformatics* 30: i121-129
- 818 Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast  
819 G (2007) The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* 3:  
820 e203
- 821 Leviatan N, Alkan N, Leshkowitz D, Fluhr R (2013) Genome-wide survey of cold stress regulated  
822 alternative splicing in *Arabidopsis thaliana* with tiling microarray. *PLoS One* 8: e66511
- 823 Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin CF, Makalowski W,  
824 Brown JW, Jarmolowski A (2004) Determinants of plant U12-dependent intron splicing  
825 efficiency. *Plant Cell* 16: 1340-1352
- 826 Li D, Heiling S, Baldwin IT, Gaquerel E (2016) Illuminating a plant's tissue-specific metabolic diversity  
827 using computational metabolomics and information theory. *Proc Natl Acad Sci U S A*
- 828 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome  
829 Project Data Processing S (2009) The sequence alignment/map format and SAMtools.  
830 *Bioinformatics* 25: 2078-2079
- 831 Li Q, Xiao G, Zhu YX (2014) Single-nucleotide resolution mapping of the *Gossypium raimondii*  
832 transcriptome reveals a new mechanism for alternative splicing of introns. *Mol Plant* 7: 829-840
- 833 Li W, Lin WD, Ray P, Lan P, Schmidt W (2013) Genome-wide detection of condition-sensitive  
834 alternative splicing in *Arabidopsis* roots. *Plant Physiol* 162: 1750-1763
- 835 Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X,  
836 Darnell JC, Darnell RB (2008) HITS-CLIP yields genome-wide insights into brain alternative  
837 RNA processing. *Nature* 456: 464-469
- 838 Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res*  
839 *Notes* 5: 337
- 840 Ling Z, Zhou W, Baldwin IT, Xu S (2015) Insect herbivory elicits genome-wide alternative splicing  
841 responses in *Nicotiana attenuata*. *Plant J* 84: 228-243
- 842 Liu HX, Goodall GJ, Kole R, Filipowicz W (1995) Effects of secondary structure on pre-mRNA splicing:  
843 hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in *Nicotiana*  
844 *plumbaginifolia*. *EMBO J* 14: 377-388

- 845 Lopato S, Gattoni R, Fabini G, Stevenin J, Barta A (1999) A novel family of plant splicing factors with a  
846 Zn knuckle motif: examination of RNA binding and splicing activities. *Plant Mol Biol* 39: 761-  
847 773
- 848 Lorkovic ZJ (2009) Role of plant RNA-binding proteins in development, stress response and genome  
849 organization. *Trends Plant Sci* 14: 229-236
- 850 Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine.  
851 *Molecular Pharmaceutics* 13: 1445-1454
- 852 Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased  
853 complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res* 22: 1184-1195
- 854 Mastrangelo AM, Marone D, Laido G, De Leonardis AM, De Vita P (2012) Alternative splicing:  
855 enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci* 185-186: 40-49
- 856 Mattick JS (1994) Introns: evolution and function. *Curr Opin Genet Dev* 4: 823-831
- 857 Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in  
858 Mammalian tissues. *Science* 338: 1593-1599
- 859 Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when  
860 nonsense affects RNA abundance. *Trends Biochem Sci* 23: 198-199
- 861 Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M  
862 (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by  
863 alternative splicing and nonsense-mediated decay. *Gene Dev* 21: 708-718
- 864 Noe L, Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res*  
865 33: W540-543
- 866 Palusa SG, Reddy AS (2010) Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine  
867 (SR) genes with nonsense-mediated decay. *New Phytol* 185: 83-89
- 868 Perteu M, Mount SM, Salzberg SL (2007) A computational survey of candidate exonic splicing enhancer  
869 motifs in the model plant *Arabidopsis thaliana*. *BMC Bioinformatics* 8: 159
- 870 Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev*  
871 *Plant Biol* 58: 267-294
- 872 Reddy AS, Marquez Y, Kalyna M, Barta A (2013) Complexity of the alternative splicing landscape in  
873 plants. *Plant Cell* 25: 3657-3683
- 874 Reddy AS, Shad Ali G (2011) Plant serine/arginine-rich proteins: roles in precursor messenger RNA  
875 splicing, plant development, and stress responses. *Wiley Interdiscip Rev RNA* 2: 875-889
- 876 Richardson DN, Rogers MF, Labadorf A, Ben-Hur A, Guo H, Paterson AH, Reddy AS (2011)  
877 Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-  
878 family classification and extent of alternative splicing. *PLoS One* 6: e24542
- 879 Rosenberg AB, Patwardhan RP, Shendure J, Seelig G (2015) Learning the sequence determinants of  
880 alternative splicing from millions of random sequences. *Cell* 163: 698-711
- 881 Rosloski SM, Singh A, Jali SS, Balasubramanian S, Weigel D, Grbic V (2013) Functional analysis of  
882 splice variant expression of MADS AFFECTING FLOWERING 2 of *Arabidopsis thaliana*. *Plant*  
883 *Mol Biol* 81: 57-69
- 884 Satyawan D, Kim MY, Lee SH (2016) Stochastic alternative splicing is prevalent in mungbean (*Vigna*  
885 *radiata*). *Plant Biotechnol J*
- 886 Schoning JC, Streitner C, Meyer IM, Gao Y, Staiger D (2008) Reciprocal regulation of glycine-rich  
887 RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-  
888 mediated decay in *Arabidopsis*. *Nucleic Acids Res* 36: 6977-6987
- 889 Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. *Nat Struct*  
890 *Mol Biol* 16: 990-995
- 891 Schwartz SH, Silva J, Burstein D, Pupko T, Eyraas E, Ast G (2008) Large-scale comparative analysis of  
892 splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* 18: 88-103
- 893 Severing EI, van Dijk AD, Morabito G, Busscher-Lange J, Immink RG, van Ham RC (2012) Predicting  
894 the impact of alternative splicing on plant MADS domain protein function. *PLoS One* 7: e30524

- 895 Severing EI, van Dijk ADJ, Stiekema WJ, van Ham RC (2009) Comparative analysis indicates that  
896 alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC*  
897 *Genomics* 10
- 898 Seymour DK, Koenig D, Hagemann J, Becker C, Weigel D (2014) Evolution of DNA methylation patterns  
899 in the *Brassicaceae* is driven by differences in genome organization. *PLoS Genet* 10: e1004785
- 900 Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL, Tian Z (2014) Global  
901 dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 26: 996-1008
- 902 Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV (2014)  
903 The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun*  
904 5: 3833
- 905 Simpson CG, Jennings SN, Clark GP, Thow G, Brown JW (2004) Dual functionality of a plant U-rich  
906 intronic sequence element. *Plant J* 37: 82-91
- 907 Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS,  
908 Newman LK, Wang W, Mandakova T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S,  
909 Brandvain Y, Coop G, Andolfatto P, Hu TT, Blanchette M, Clark RM, Quesneville H, Nordborg  
910 M, Gaut BS, Lysak MA, Jenkins J, Grimwood J, Chapman J, Prochnik S, Shu S, Rokhsar D,  
911 Schmutz J, Weigel D, Wright SI (2013) The *Capsella rubella* genome and the genomic  
912 consequences of rapid mating system evolution. *Nat Genet* 45: 831-835
- 913 Smith CW, Chu TT, Nadal-Ginard B (1993) Scanning and competition between AGs are involved in 3'  
914 splice site selection in mammalian introns. *Mol Cell Biol* 13: 4939-4952
- 915 Smith CW, Porro EB, Patton JG, Nadal-Ginard B (1989) Scanning from an independently specified  
916 branch point defines the 3' splice site of mammalian introns. *Nature* 342: 243-247
- 917 Sorek R, Ast G, Graur D (2002) *Alu*-containing exons are alternatively spliced. *Genome Res* 12: 1060-  
918 1067
- 919 Staiger D, Brown JW (2013) Alternative splicing at the intersection of biological timing, development,  
920 and stress responses. *Plant Cell* 25: 3640-3656
- 921 Streitner C, Koster T, Simpson CG, Shaw P, Danisman S, Brown JW, Staiger D (2012) An hnRNP-like  
922 RNA-binding protein affects alternative splicing by in vivo interaction with transcripts in  
923 *Arabidopsis thaliana*. *Nucleic Acids Res* 40: 11240-11255
- 924 Su Z, Wang J, Yu J, Huang X, Gu X (2006) Evolution of alternative splicing after gene duplication.  
925 *Genome Res* 16: 182-189
- 926 Szczesniak MW, Kabza M, Pokrzywa R, Gudys A, Makalowska I (2013) ERISdb: a database of plant  
927 splice sites and splicing signals. *Plant Cell Physiol* 54: e10
- 928 Thomas J, Palusa SG, Prasad KV, Ali GS, Surabhi GK, Ben-Hur A, Abdel-Ghany SE, Reddy AS (2012)  
929 Identification of an intronic splicing regulatory element involved in auto-regulation of alternative  
930 splicing of *SCL33* pre-mRNA. *Plant J* 72: 935-946
- 931 Tomato Genome C (2012) The tomato genome sequence provides insights into fleshy fruit evolution.  
932 *Nature* 485: 635-641
- 933 Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq.  
934 *Bioinformatics* 25: 1105-1111
- 935 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter  
936 L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with  
937 TopHat and Cufflinks. *Nat Protoc* 7: 562-578
- 938 Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM  
939 measure is inconsistent among samples. *Theory Biosci* 131: 281-285
- 940 Wang BB, Brendel V (2004) The ASRG database: identification and survey of *Arabidopsis thaliana*  
941 genes involved in pre-mRNA splicing. *Genome Biol* 5: R102
- 942 Wang XT, Hu LJ, Wang XF, Li N, Xu CM, Gong L, Liu B (2016) DNA methylation affects gene  
943 alternative splicing in plants: an example from rice. *Mol Plant* 9: 305-307
- 944 Warf MB, Berglund JA (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem*  
945 *Sci* 35: 169-178

- 946 Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the  
947 human transcriptome. *Nucleic Acids Res* 30: 3754-3766
- 948 Yang R, Wang X (2013) Organ evolution in angiosperms driven by correlated divergences of gene  
949 sequences and expression patterns. *Plant Cell* 25: 71-82
- 950 Yoshimura K, Yabuta Y, Ishikawa T, Shigeoka S (2002) Identification of a *cis* element for tissue-specific  
951 alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. *J Biol Chem*  
952 277: 40623-40632
- 953
- 954

955 **FIGURES and FIGURE LEGENDS**



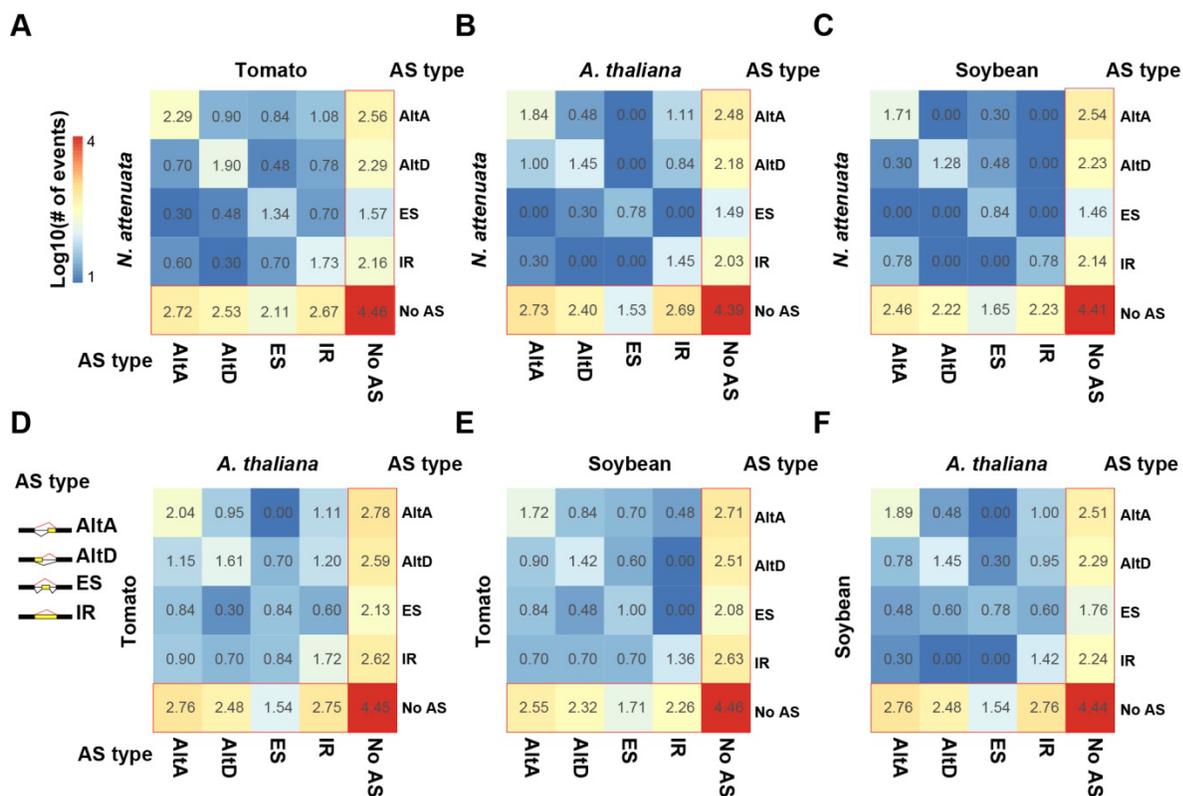
956

957 **Figure 1. Species-specific clustering of alternative splicing (AS) among different plant**

958 **species. (A) and (C), heatmaps depict species-specific clustering based on PSI values among**

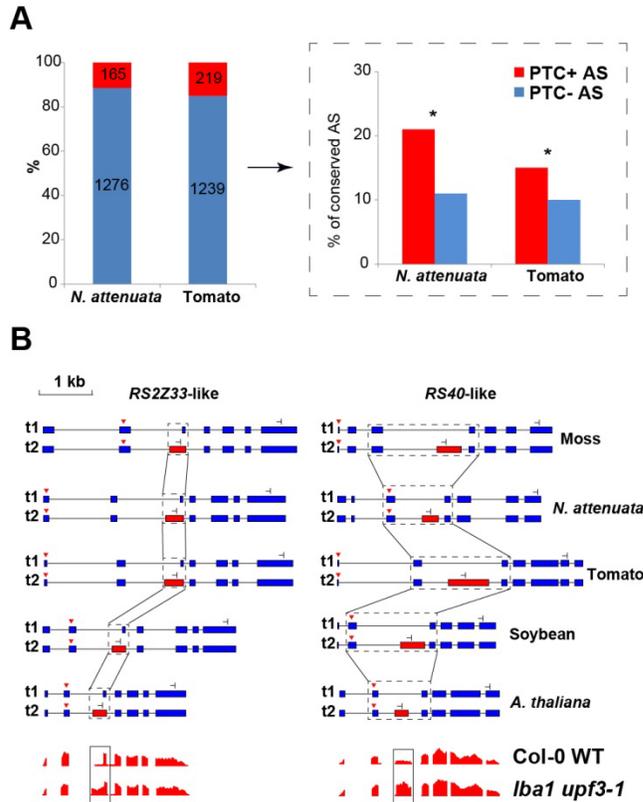
959 four eudicots species (**A**) and three Brassicaceae species (**C**). The clustering is based on  
960 conserved splicing junctions (A and C: n = 502 and 5241, respectively). (**B**) and (**D**), heatmaps  
961 depict species-specific clustering based on presence and absence of AS of the one-to-one  
962 orthologous genes. In total, junctions from 3,857 (**B**) and 6,262 (**D**) orthologous were used for  
963 the clustering. Numbers present in each branch node represent the approximately unbiased  
964 bootstrap value calculated from 1000 bootstrap replications. The color code above each heatmap  
965 represents species, tissue, and treatments.

966

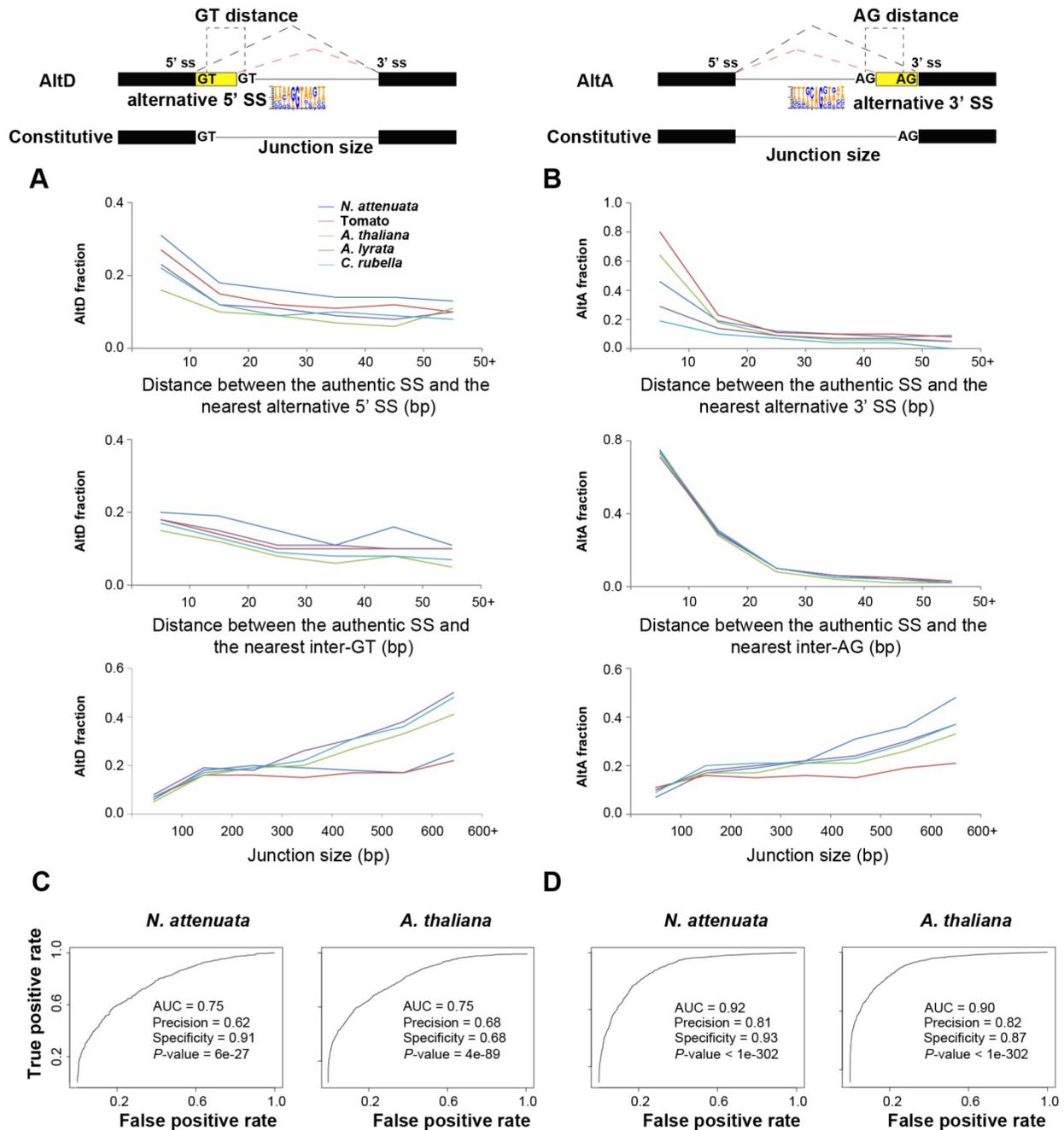


967  
 968 **Figure 2. Transition spectrum of AS between each species pairs. (A)** *N. attenuata* vs tomato,  
 969 **(B)** *N. attenuata* vs *A. thaliana*, **(C)** *N. attenuata* vs soybean, **(D)** tomato vs *A. thaliana*, **(E)**  
 970 tomato vs soybean and **(F)** soybean vs *A. thaliana*. The color of each grid refers to log10  
 971 transformed number of AS events. The transformed values are also shown in the middle of each  
 972 grid. AltA: alternative 3' acceptor site; AltD: alternative 5' donor site; ES: exon skipping; IR:  
 973 intron retention.

974



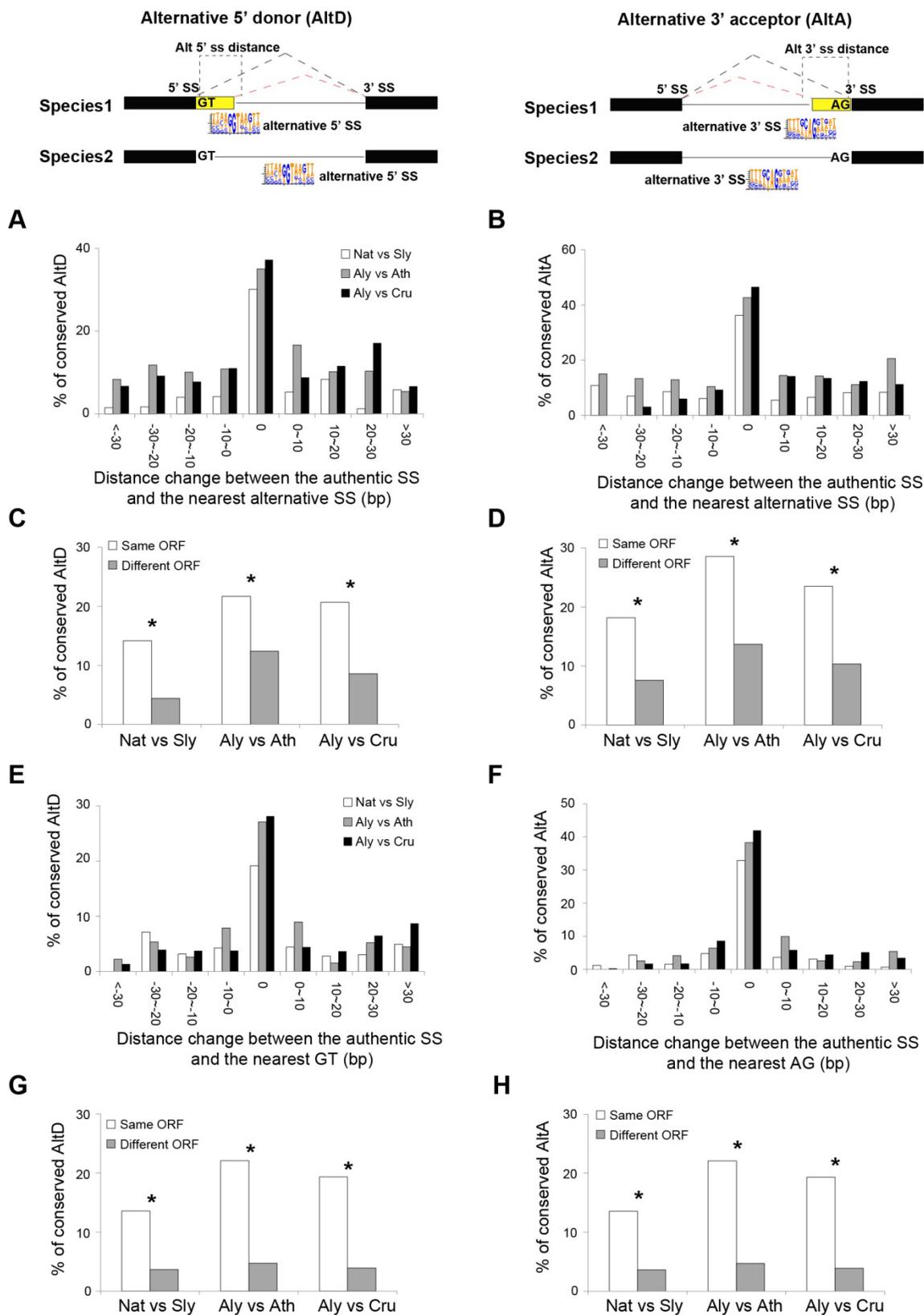
975  
976 **Figure 3. The conservation of AS between PTC+ and PTC- AS.** (A) The number and relative  
977 portions of PTC+/- AS in *N. attenuata* and tomato. The insert indicated by the black arrow  
978 depicts the likelihood of PTC+ and PTC- AS that are conserved between *N. attenuata* and  
979 tomato. Asterisks indicate the significance as determined by Fisher's exact test ( $P < 0.05$ ). (B)  
980 Conserved AS between moss and eudicots in serine/arginine-rich splicing factor *RS2Z33*-like  
981 and *RS40*-like gene. The diagrams of the structure of transcripts generated by the AS in all five  
982 species, the dominant and minor transcripts are represented by t1 and t2, respectively.  
983 Constitutive exons are represented by blue boxes, alternatively spliced exons are represented by  
984 red boxes and introns are represented by black solid lines. The black dotted boxes highlight the  
985 conserved AS and the start and stop codons are shown as red triangles and stop signs,  
986 respectively. The diagrams in the bottom panel showed the relative read coverage of *AtRS2Z33*  
987 and *AtRS40* exons in wild type plant and *lba1 upf3-1* double mutants. The black box highlights  
988 the coverage of the spliced region which is significantly increased in *lba1 upf3-1* double mutants  
989 (The diagrams are modified based on the data shown in  
990 <http://gbrowse.cbio.mskcc.org/gb/gbrowse/NMD201>)



992

993 **Figure 4. The determinants of alternative 5' donor site (AltD) and alternative 3' acceptor**  
 994 **site (AltA) in plants. (A) and (B), the frequencies of AltD/AltA on junctions with different**  
 995 **distance between the authentic SS and nearest alternative SS (5' ss and 3' ss, respectively), and**  
 996 **the distance between authentic SS and the nearest inter GT/AG and junction size. (C) and (D),**  
 997 **the area under the curve (AUC) plot of deep learning models using the key determinants of AltD**

998 and AltA in *N. attenuata* and *A. thaliana*. The model performance including area under the curve  
999 (AUC), accuracy, specificity and significance are also shown.  
1000



1002 **Figure 5. Features that affect the conservation of AltD and AltA between closely related**  
1003 **plant species. (A) and (B),** the portion of conserved AltD/AltA decreases with changes of  
1004 distance between authentic and alternative SS between two species. **(C) and (D),** the percent of  
1005 conserved AltD/AltA in the group that the nearest alternative 5'/3' SS in the two species  
1006 generate the same or different ORF transcripts. **(E) and (F),** the portion of conserved AltD/AltA  
1007 decreases with changes in the distance between authentic SS and nearest inter-GT/AG sites  
1008 between two species. **(G) and (H)** the percent of conserved AltD/AltA in the group that the  
1009 nearest inter-GT/AG in the two species generate transcripts with same or different open reading  
1010 frame (ORF). Nat: *N. attenuata*, Sly: Tomato, Ath: *A. thaliana*, Aly: *A. lyrata*. The asterisks  
1011 indicate the significance as determined by Fisher's exact test ( $P < 0.05$ ).