# Distinct types of short open reading frames are translated in plant cells

Igor Fesenko[1,*], Ilya Kirov[1], Andrey Kniazev[1], Regina Khazigaleeva[1], Vassili Lazarev[2], Daria Kharlampieva[2], Ekaterina Grafskaia[2], Viktor Zgoda[3], Ivan Butenko[2], Georgy Arapidi[1], Anna Mamaeva[1], Vadim Ivanov[1], Vadim Govorun[1,2].

*[1] Laboratory of Proteomics, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russian Federation; [2] Federal Research and Clinical Centre of Physical-Chemical Medicine, Moscow, Russian Federation; [3]Laboratory of System Biology, Institute of Biomedical Chemistry, Moscow, Russian Federation.*

**Corresponding author(s).**

***** Igor Fesenko, e-mail: fesigor@gmail.com

**Running title:** Genome-wide translation of sORFs in plants

**Keywords:** short open reading frames, plant peptides, LC-MS/MS, evolution, alternative splicing, lncRNA

## ABSTRACT

Genomes contain millions of short (<100 codons) open reading frames (sORFs), which are usually dismissed during gene annotation. Nevertheless, peptides encoded by such sORFs can play important biological roles, and their impact on cellular processes has long been underestimated. Here, we analyzed approximately 70,000 transcribed sORFs in the model plant *Physcomitrella patens* (moss). Several distinct classes of sORFs that differ in terms of their position on transcripts and the level of evolutionary conservation are present in the moss genome. Over 5000 sORFs were conserved in at

least one of ten plant species examined. Mass spectrometry analysis of proteomic and peptidomic datasets suggested that 602 sORFs located on distinct parts of mRNAs and long non-coding RNAs (lncRNAs) are translated, including 74 conservative sORFs. Combined analysis of the translation of the sORFs and the main ORF from a single gene suggested the existence of bi- and poly-cistronic mRNAs with tissue-specific expression. Alternative splicing is likely involved in the excision of translatable sORFs from such transcripts. We identified a group of sORFs homologous to known protein domains and suggested they function as small interfering peptides. Functional analysis of a candidate lncRNA-encoded peptide showed it to be involved in regulating growth and differentiation in moss. The high evolutionary rate and wide translation of sORFs suggest that they may provide a reservoir of potentially active peptides and their importance as a raw material for gene evolution. Our results thus open new avenues for discovering novel, biologically active peptides in the plant kingdom.

## INTRODUCTION

The genomes of nearly all organisms contain hundreds of thousands of short open reading frames (sORFs; <100 codons) whose coding potential has been the subject of recent reviews (Andrews and Rothnagel 2014; Couso 2015; Hellens et al. 2016; Couso and Patraquim 2017). However, gene annotation algorithms are generally not suited for dealing with sORFs because short sequences are unable to obtain high conservation scores, which serve as an indicator of functionality (Ladoukakis et al. 2011). Nevertheless, using various bioinformatic approaches, sORFs with high coding potential have been identified in a range of organisms including fruit flies, mice, yeast and *Arabidopsis thaliana* (Ladoukakis et al. 2011; Hanada et al. 2013; Aspden et al. 2014; Bazzini et al. 2014). The first systematic study of sORFs was conducted on baker's yeast, where 299 previously non-annotated sORFs were identified and tested in genetic experiments (Kastenmayer et al. 2006). Subsequently, 4561 conserved sORFs were identified in the genus *Drosophila*, 401 of which were postulated to be functional, taking into account their syntenic positions, favorable Dn/Ds values and

53 transcriptional evidence (Ladoukakis et al. 2011). In a recent study, Mackowiak and colleagues
54 predicted the presence of 2002 novel conserved sORFs (from 9 to 101 codons) in *H. sapiens*, *M.*
55 *musculus*, *D. rerio*, *D. melanogaster* and *C. elegans* (Mackowiak et al. 2015). The first comprehensive
56 study of sORFs in plants postulated the existence of thousands of sORFs with high coding potential in
57 Arabidopsis (Lease and Walker 2006; Hanada et al. 2007; Hanada et al. 2013), including 49 that
58 induced various morphological changes and had visible phenotypic effects.

59 Recent studies have pointed to the important roles of sORF-encoded peptides (SEPs) in
60 cells (Magny et al. 2013; Nelson et al. 2016; D'Lima et al. 2017; Huang et al. 2017; Matsumoto et al.
61 2017). However, unraveling the roles of SEPs is a challenging task, as is their detection at the
62 biochemical level. In animals, SEPs are known play important roles in a diverse range of cellular
63 processes (Kondo et al. 2010; Magny et al. 2013). By contrast, only a few functional SEPs have been
64 reported in plants, including POLARIS (PLS; 36 amino acids), EARLY NODULIN GENE 40 (ENOD40;
65 12, 13, 24 or 27 amino acids), ROTUNDIFOLIA FOUR (ROT4; 53 amino acids), KISS OF DEATH (KOD;
66 25 amino acids), BRICK1 (BRK1; 84 amino acids), Zm-908p11 (97 amino acids) and Zm-401p10 (89
67 amino acids) (Andrews and Rothnagel 2014; Tavormina et al. 2015). These SEPs help modulate root
68 growth and leaf vascular patterning (Chilley et al. 2006), symbiotic nodule development (Djordjevic
69 et al. 2015), polar cell proliferation in lateral organs and leaf morphogenesis (Narita et al. 2004), and
70 programmed cell death (apoptosis) (Blanvillain et al. 2011).

71 To date, functional sORFs have been found in a variety of transcripts, including
72 untranslated regions of mRNA (5′ leader and 3′ trailer sequences), lncRNAs, and microRNA
73 transcripts (pri-miRNAs) (Andrews and Rothnagel 2014; Laing et al. 2015; Lauressergues et al. 2015;
74 Couso and Patraquim 2017). Evidence for the transcription of potentially functional sORFs has been
75 obtained in *Populus deltoides*, *Phaseolus vulgaris*, *Medicago truncatula*, *Glycine max* and *Lotus*
76 *japonicus* (Guillen et al. 2013). The transcription of sORFs can be regulated by stress conditions and
77 depends on the developmental stage of the plant (De Coninck et al. 2013; Hanada et al. 2013;
78 Rasheed et al. 2016). Indeed, sORFs might represent an important source of advanced traits required
79 under stress conditions. During stress, genomes undergo widespread transcription to produce a
80 diverse range of RNAs (Kim et al. 2010; Mazin et al. 2014); therefore, a large portion of sORFs

81 becomes accessible to the translation machine for peptide production. Stress conditions can lead to
82 the transcription of sORFs located in genomic regions that are usually non-coding (Giannakakis et al.
83 2015). Such sORFs appear to serve as raw materials for the birth and subsequent evolution of new
84 protein-coding genes (Couso and Patraquim 2017).

85 The transcription of an sORF does not necessarily indicate that it fulfills any biological
86 role, as opposed to being a component of the so-called translational noise (Guttman et al. 2013).
87 According to ribosomal profiling data, thousands of lncRNAs display high ribosomal occupancy in
88 regions containing sORFs in mammals (Ingolia et al. 2011; Aspden et al. 2014; Bazzini et al. 2014).
89 However, lncRNAs can have the same ribosome profiling patterns as canonical non-coding RNAs
90 (e.g., rRNA) that are known not to be translated, implying that these lncRNAs are unlikely to produce
91 functional peptides (Guttman et al. 2013). In addition, identification of SEPs via mass spectrometry
92 analyses has found many fewer peptides than predicted sORFs (Slavoff et al. 2013; Aspden et al.
93 2014). Thus, the abundance, lifetime and other features of SEPs are generally unclear.

94 In this study, we performed comprehensive analysis of sORFs with canonical AUG start
95 codons and high coding potential in the genome of the model moss *P. patens*. To identify candidate
96 functional sORFs, we developed an integrated pipeline, including analysis of transcriptomic,
97 proteomic and peptidomic data. We classified the sORFs based on their locations on transcripts and
98 analyzed their features, such as evolutionary conservation, peptide-coding potential and possible
99 functions. We determined that plant genomes contain hundreds of translatable sORFs, including
100 those located in alternative frames in protein-coding genes. The speed of evolution depended on the
101 type of sORF, with CDS-sORFs and lncRNA-ORFs under strong positive selection while uORFs and
102 dORFs had a greater chance of being fixed in the genome. Moreover, the presence of some sORFs in
103 the transcriptome depended on alternative splicing events. We also identified more than 200 sORFs
104 sharing homology with known proteins, implying that they function as small interfering peptides.
105 Finally, we selected a candidate lncRNA-encoded peptide for further analysis and provide evidence
106 for its biological function.

## RESULTS

### Discovery and classification of potential coding sORFs in the moss genome

Our approach is summarized in Figure 1A. At the first stage of analysis, we used the sORFfinder tool (Hanada et al. 2010) to identify single-exon sORFs starting with an AUG start codon and less than 300 bp long. This approach resulted in the identification of 638,439 sORFs with high coding potential (CI index) in all regions of the *P. patens* genome.

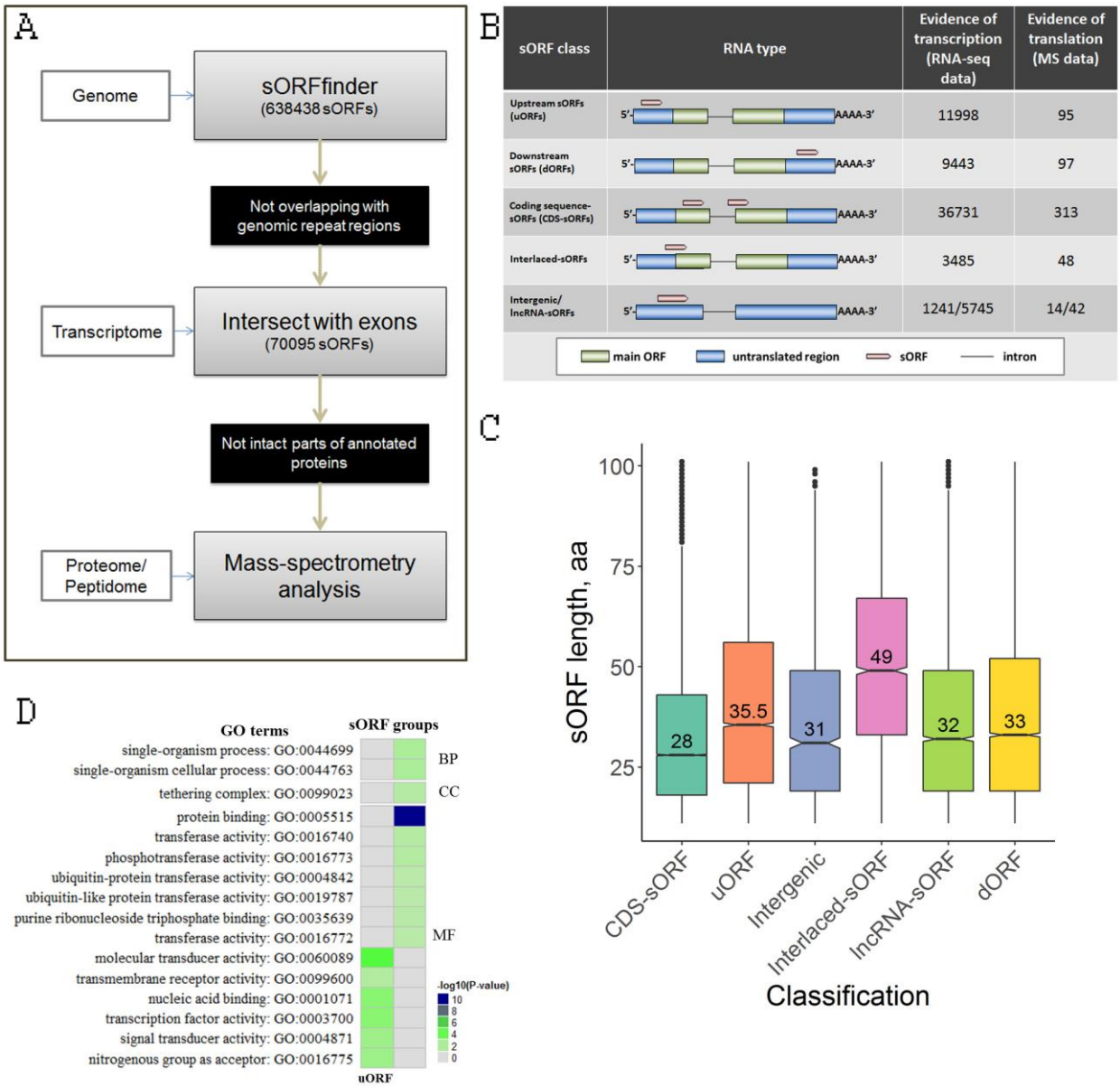



**Figure 1. Several distinct types of sORFs are present in the moss genome. A** – Pipeline used in this study to identify coding sORFs; **B** – Proposed classification of sORFs according to the types of encoding transcripts: upstream ORFs (uORFs) and downstream ORFs (dORFs) in the untranslated

117 regions (UTRs) of canonical mRNAs; CDS-sORFs, which overlap with protein-coding sequences in
118 non-canonical +2 or +3 reading frames or are truncated versions of proteins generated by alternative
119 splicing; interlaced-sORFs, which overlap both the protein-coding sequence and UTR on the same
120 transcript; lncRNA-sORFs and intergenic sORFs, which are located on short non-protein coding
121 transcripts.; **C** – Boxplot of the length distribution of sORFs in different groups; **D** – The results of GO
122 enrichment analysis for genes possessing uORFs and CDS-sORFs. BP, CC and MF represent "Biological
123 process", "Cellular component" and "Molecular function", respectively.
124
125 We selected 70,095 unique sORFs located on transcripts annotated in the moss genome
126 (phytozome.jgi.doe.gov) and/or our dataset (Fesenko et al. 2015) for further analysis, as well as
127 those on lncRNAs from two databases -CantataDB (Szczesniak et al. 2016) and GreenC (Paytuvi
128 Gallart et al. 2016); sORFs located in repetitive regions were discarded (Supplemental Table S1).
129 These selected sORFs, which were 33 to 303 bp long, were located on 33,981 transcripts (22,969
130 genes), with up to 28 sORFs per transcript (Supplemental Figure S1A).
131 We then classified the sORFs based on their location on the transcript: 63,109 "genic-sORFs"
132 (located on annotated transcripts, but not on lncRNA), 1241 "intergenic-sORFs" (located on
133 transcripts from our dataset and not annotated in the current version of the genome) and 5745
134 "lncRNA-sORFs" (located on lncRNAs from CantataDB (Szczesniak et al. 2016), GreenC (Paytuvi
135 Gallart et al. 2016) or our data set (Fesenko et al. 2017); Figure 1B). The genic-sORFs include 11,998
136 upstream ORFs (uORFs; for 5'-UTR location), 9443 downstream ORFs (dORFs; for 3'-UTR location),
137 36,731 coding sequence-sORFs (CDS-sORFs; sORFs overlapping with major ORFs in non-canonical +2
138 and +3 reading frames) and 3485 interlaced-sORFs (overlapping with both the CDS and 5'-UTR or
139 CDS and 3'-UTR on the same transcript) (Figure 1B, Supplemental Figure S1B).
140 As expected based on the sORFfinder search strategy (Hanada et al. 2010), the sORF set was
141 enriched in CDS-sORFs (52%, Fisher's exact test, P-value = 1.736392e-285), whereas dORFs, uORFs
142 and interlaced-sORFs were underrepresented (Fisher's exact test, P-value < 4.792689e-88)
143 compared to a random exonic fragments (REF) set, which was used as a negative control.

144 On average, CDS-sORFs (median size of 22 codons) were shorter than uORFs (median size of
145 35 codons; Mann-Whitney P = 2.2e-151) and dORFs (median length 32 codons, Mann-Whitney P =
146 1.03e-43). The median size of interlaced-sORFs was 49 codons, which is significantly longer than
147 other genic-sORFs (Mann-Whitney P = 0.0021) (Figure 1C).

148 We performed comparative GO enrichment analysis of four groups of genic-sORFs (dORF,
149 uORF, CDS-sORF and interlaced). To exclude the possibility that differences between groups could be
150 explained merely by structural differences in genes carrying sORFs (e.g., genes with longer 5'-UTRs
151 have a greater chance of possessing uORFs), we also performed GO enrichment analysis of a set of
152 genes with randomly selected exon fragments (REFs). GO terms that were enriched in both datasets
153 were excluded. The analysis showed significant (adjusted P-value < 0.01) GO enrichment for genes
154 possessing CDS-sORFs and uORFs. The patterns of GO enrichment differed between the two groups
155 of genes: set of genes possessing CDS-sORFs were enriched in GO terms associated with protein
156 binding and transferase activity, while genes possessing uORFs were involved in signal transduction
157 and transcriptional regulation (Figure 1D). Such contrasting patterns in functions between genes
158 with different sORF locations allude to the roles of sORFs and/or their peptides in different levels of
159 cellular regulation.

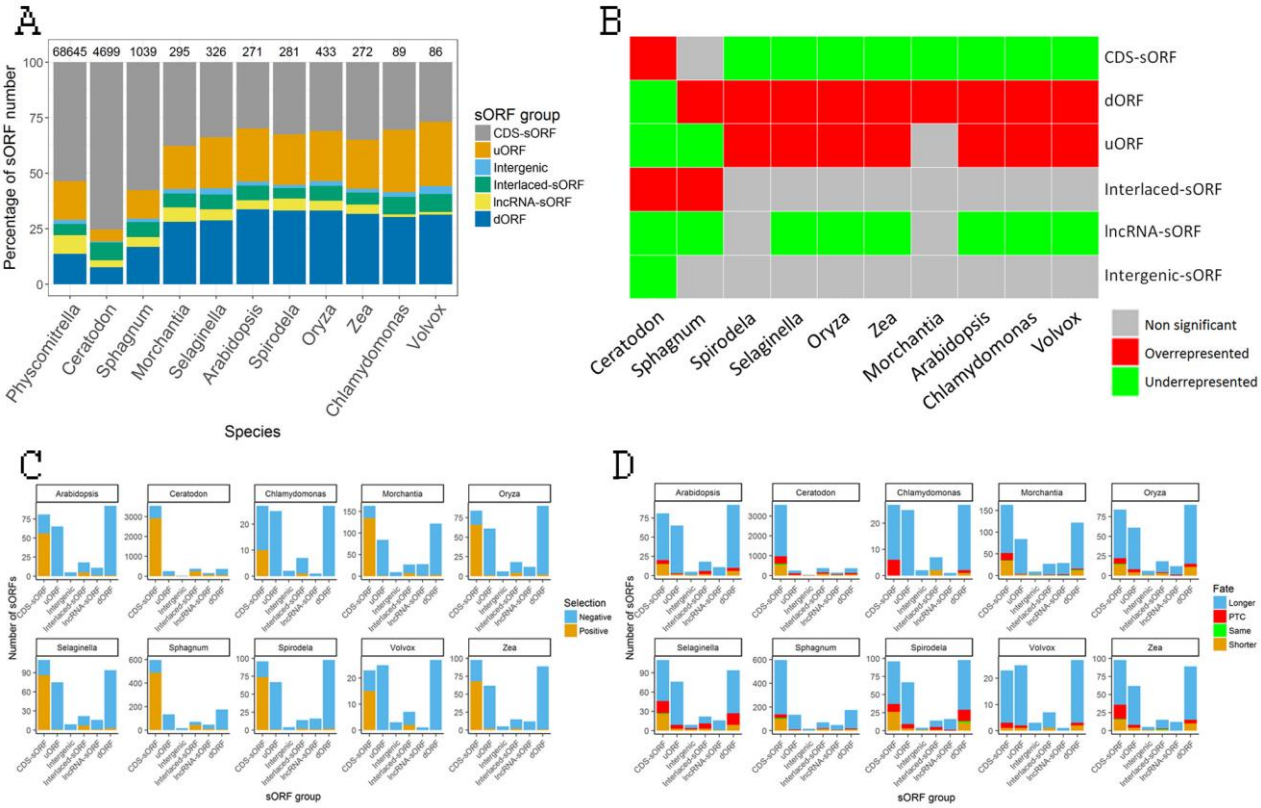
160

161 **Analysis of evolutionary conservation of sORFs**

162 It is widely accepted that evolutionary conservation is a strong indicator of functionality (Ladoukakis
163 et al. 2011). To estimate the number of conservative sORFs in the moss genome and the evolutionary
164 pressure on their amino acid sequences, we performed a tBLASTn (e-value cutoff 0.00001) search of
165 each sORF sequence against the transcriptomes of ten species including those that diverged from *P.*
166 *patens* 177 (*Ceratodon purpureus*), 320 (*Sphagnum fallax*), 493 (*Marchantia polymorpha*), 532
167 (*Arabidopsis thaliana*, *Oryza sativa, Zea mays, Selaginella moellendorffii, Spirodela polyrhiza*) and
168 1160 (*Volvox carteri, Chlamydomonas reinhardtii*) Mya (Supplemental Figure S2).

169 We found 5034 conserved sORFs with detectable homologous sequences in at least one
170 species: 4797 in *C. purpureus*, 1049 in *S. fallax*, 436 in *M. polymorpha*, 328 in *S. moellendorffii*, 297 in
171 *S. polyrhiza*, 275 in *A. thaliana*, 282 in *Z. mays*, 274 in *O. sativa*, 86 in *V. carteri* and 89 in *C. reinhardtii*.

172 The number of conserved sORFs was negatively correlated with the time since divergence, with the
173 fewest homologous sequences found in *V. carteri* and *C. reinhardtii*, which diverged more than 1000
174 Mya from a common ancestor. We found that lncRNA-sORFs were underrepresented among sORFs
175 having homologs in the ten species examined (Figure 2A). We also found significantly fewer uORFs
176 and dORFs in the two closest species, *C. purpureus* and *S. fallax*, whereas CDS-sORFs were
177 significantly overrepresented in these species (Fisher's exact test, P<2.2e-16) (Figure 2B).




178

179 **Figure 2. Analysis of the trends in sORFs evolution.** A – The percentage of each type of
180 sORF among sORFs having homologs in ten plant species. B – Statistical analysis (by Fisher's exact
181 test) of differences between the number of conservative sORFs in each of ten species and the initial
182 dataset; C – Pairwise Ka/Ks ratio distribution for each type of sORF conserved among ten plant
183 species; D – Bar plot comparing the lengths of moss sORFs with the putative lengths of homologous
184 sORFs in ten plant species. Homologous sORF sequences can be expanded (blue) in PTC – premature
185 termination codon in homologous sORFs.

186

187 However, the portion of uORFs and dORFs found in the more distant species was increased
188 relative to the initial dataset compared to CDS-sORFs, causing their significant overrepresentation
189 (Fisher's exact test, P<0.0005). Thus, the relative enrichment of conserved CDS-sORFs and
190 interlaced-sORFs found in the two closest species of *P. patens*, *C. purpureus* and *S. fallax*, resulted
191 from a significant reduction in the number of uORFs and dORFs (Figure 2A).

192 Because upstream sORFs are capable of attenuating translation of the downstream main
193 open reading frame, they can undergo strong selection and be eliminated from UTRs. As a control, we
194 also investigated changes in the proportion of uREFs, dREFs and CDS-REFs in these ten species and
195 obtained opposite results, with significant overrepresentation of CDS-REFs and underrepresentation
196 of dREFs and uREFs in all species (Supplemental Figure S3). To compare this trend with protein
197 coding genes, we selected 158 intronless small proteins (< 100 aa) from the *P. patens* genome
198 annotation. The percentages of sORFs and these proteins showing homology with at least one species
199 were significantly different (7.2% sORFs vs. 86% small proteins), pointing to high genome turnover
200 of sORF sequences.

201 To better understand the large-scale trends of sORF evolution, we examined the differences
202 in selection pressure at the amino acid level between different major groups of sORFs (CDS-sORFs,
203 uORFs, dORFs, lncRNA-sORFs, interlaced-sORFs) using the criterion of Dn/Ds. This analysis showed
204 that the highest portion of sORFs comprised CDS-sORFs, with Dn/Ds ratio > 1, implying ongoing
205 positive selection of sORFs emerging in the CDS of protein-coding genes. This criterion for other
206 sORF groups was < 1 in most cases, pointing to purifying selection for these sequences (Figure 2C).

207 The possible evolution of non-coding portions of the genome into protein-coding genes is a
208 subject of intensive debate (Carvunis et al. 2012; McLysaght and Guerzoni 2015; Couso and
209 Patraquim 2017). The ability of non-coding RNAs bearing sORF sequences to give rise to new genes
210 is a controversial idea whose confirmation is a challenging task. To gain new insight into the process
211 of gene birth, we assessed whether the lengths of homologous sORFs in other species were the same
212 as those in moss or if they tended to change in size. According to our data, putative homologous
213 sORFs tended to differ in length in most cases (Figure 2D). We found that most sORFs expanded

214 during evolution, providing support for the notion that they function as raw materials for selection;
215 however, this point requires further confirmation.
216 Thus, evolutionary analysis demonstrated that the conservation of an sORF on a large
217 evolutionary scale differs from that of randomly selected exon sequences and depends on the
218 location of the sORF, with a greater chance of being fixed for uORFs and dORFs, whereas CDS-sORFs
219 and lncRNA-ORFs are under strong positive selection. A high rate of evolution is the driving force for
220 the exclusion of an sORF from a coding sequence.
221

222 ## Experimental evidence for the translation of sORFs

223 Obtaining evidence for the translation of sORFs is an important step towards identifying functional
224 SEPs. We verified the translation of our predicted sORFs using mass-spectrometry (MS) analysis,
225 which is often considered to be the gold standard for detecting proteins or peptides in a cell. Taking
226 into account the shortage of proteomic methods for identifying small proteins or peptides, in the
227 current study, we used two datasets: the "peptidomic" dataset (endogenous peptides extracted from
228 three types of moss cells: gametophores, protonemata and protoplasts) and the "proteomic" dataset
229 (tryptic peptides generated in a standard proteomic pipeline). All datasets were mapped with
230 MaxQuant against a custom database containing our sORFs together with nuclear, chloroplast and
231 mitochondrial moss protein sequences (see details in the Methods). In total, we confirmed the
232 translation of 602 sORFs: 205 in gametophores, 288 in protonemata and 196 in protoplasts (Figure
233 3A, Supplemental Table S2). The most prominent group of translatable sORFs consisted of CDS-
234 sORFs (306, 51%) (Figure 3B). Interestingly, the translation of 42 sORFs located on lncRNAs was also
235 detected by our analysis.
236 The length of translatable sORFs ranged from 11 to 100 amino acids (aa), which were generally
237 longer than untranslatable sORFs (Mann-Whitney P = 4e-53) (Figure 3C). The length of interlaced-
238 sORFs differed significantly from that of CDS-sORFs and lncRNA-sORFs (Mann-Whitney P = 0.002
239 and Mann-Whitney P = 0.001, respectively) but did not differ from uORFs (Mann-Whitney P = 0.06).
240 We observed that PSMs (peptide spectrum matches) supporting SEP identifications had lower
241 average quality than those mapped to the protein sequences of all datasets (Supplemental Figure

S4A and B). This finding is in agreement with data obtained for the animal kingdom (Slavoff et al. 2013; Mackowiak et al. 2015).

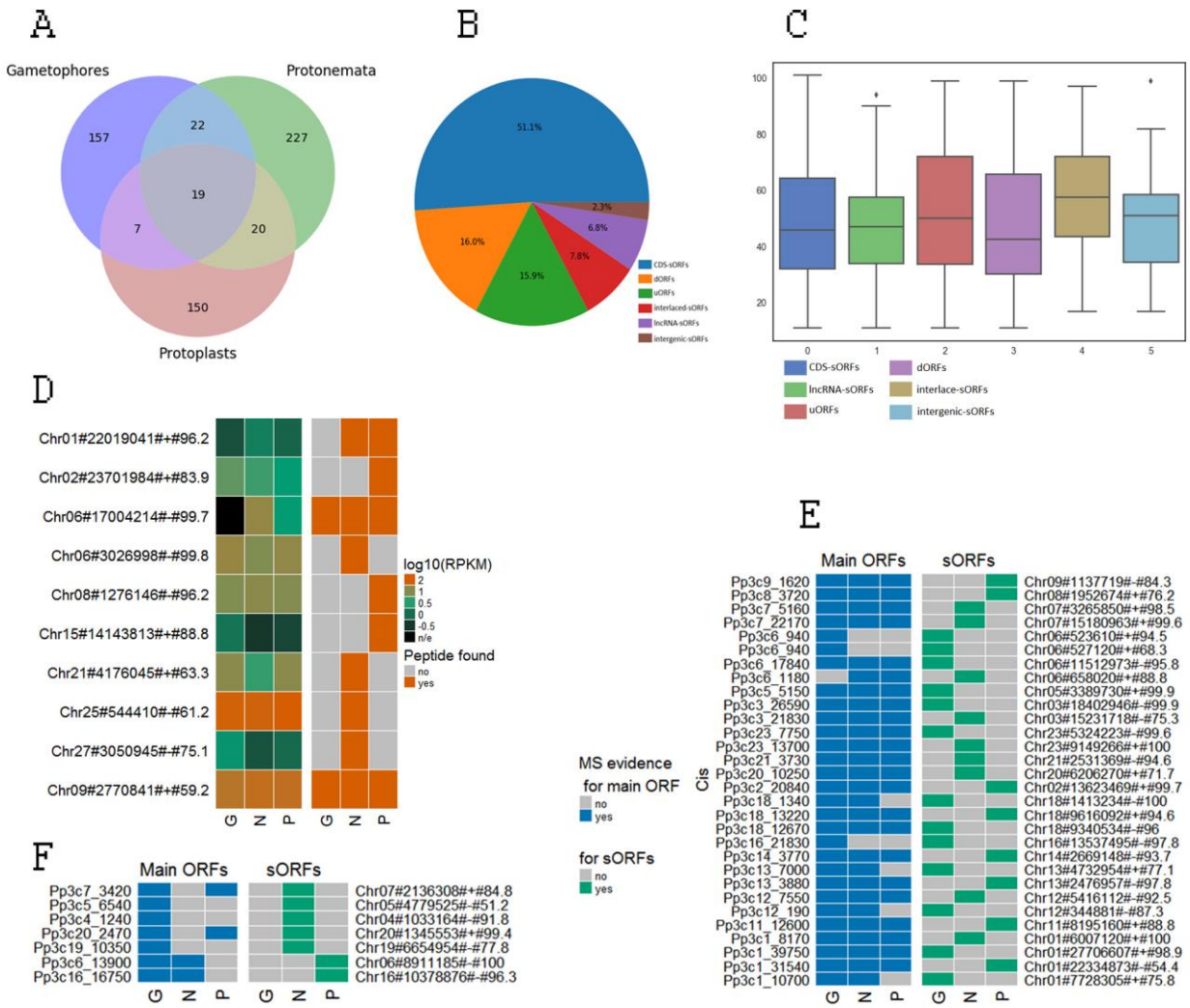



**Figure 3. Moss contains hundreds of translatable sORFs.** A – Venn diagram showing the distribution of the identified sORFs among three types of moss cells; B - Distribution of translatable sORFs based on the suggested classification; C - Length distribution of various groups of translatable sORFs; D - Heatmap of RPKM expression levels of transcripts carrying lnsRNA-sORFs and evidence of translation for three moss tissues; G, N and P correspond to gametophores, protonemata and protoplasts, respectively; E – Heatmap showing evidence of translation for sORFs and proteins in polycistronic genes in three moss tissues. G, N and P correspond to gametophores, protonemata and protoplasts, respectively. F - Examples of contrasting translational patterns of the main ORF and CDS-sORF. Only proteins confirmed by more than three unique tryptic peptides in the MS data are shown.

255 Interestingly, the quality of spectra and the values of PSMs supporting the expression of SEPs were
256 better in the "peptidomic" dataset (Supplemental Figure S4C). Also, translatable sORFs were longer
257 for those identified in the peptidomic dataset (Supplemental Figure S4D).

258 There were no significant dependencies between the level of expression of a transcript and
259 the chance of finding peptides from sORFs located on this transcript (logistic regression, P-value >>
260 0.05). However, among the 19 sORFs with evidence of translation in all types of moss cells, lncRNA-
261 sORFs were significantly overrepresented (Fisher's exact test, P-value = 0.001). Moreover, lncRNA
262 transcripts were highly expressed and produced peptides that were also detected in gametophores,
263 protonemata and protoplasts (Figure 3D). These data may point to biological significance for the
264 peptides translated from these sORFs rather than the sORFs having regulatory functions in the
265 translation of the main ORF. To investigate this notion, we explored the activity of one such SEP
266 encoded by an lncRNA (see below).

267 Standard proteomic validation requires the presence of two non-overlapping tryptic peptides
268 to confirm the translation of a protein. However, it is unlikely that more than one tryptic fragment
269 will be detected in the case of an SEP. Nevertheless, we identified more than one unique peptide for
270 six sORFs in the peptidomic dataset. Moreover, two of these SEPs, Chr09#2770841#+#59.21 (41aa)
271 and Chr25#544410#-#61.2 (61aa), were common to all three cell types and were confirmed by 15
272 and 17 unique endogenous peptides, respectively (Figure 3C). In the proteomic dataset, we observed
273 two non-overlapping tryptic peptides for only one 89-aa sORF (Chr20#14861199#+#85.5).
274 Presumably, the analysis of endogenous peptide pools may be more suitable for detecting smaller
275 SEPs.

276

277 ### sORFs can be translated together with proteins

278 Several reports provide evidence that eukaryotic mRNA can have more than one coding ORF (bi- and
279 polycistronic genes) in both plants and animals (Blumenthal 1998; Rohrig et al. 2002; Pi et al. 2009;
280 Tautz 2009). We analyzed our MS data to detect the translation of main ORFs and sORFs from the
281 same gene (putatively polycistronic). We identified 144 genes for which at least two ORFs (one main
282 ORF and one sORF) were translated, according to our MS data, including 82 connected to the

283 translation of CDS-sORFs (Supplemental Table S3). Some of these were translated simultaneously
284 with protein-coding ORFs in the same type of moss cell (Figure 3E), while others showed tissue-
285 specific expression patterns (Figure 3F). This observation suggests that specific regulatory
286 mechanisms may exist to fine-tune the translation of both sORFs and proteins situated in the same
287 gene locus. We next sought to determine whether these sORFs encode some known protein domains.
288 Interestingly, all ten sORFs that were identified in this analysis harbor intrinsically disordered
289 regions (IDRs).

290 Taken together, our findings indicate that at least 27% of translatable CDS-sORFs are
291 expressed simultaneously with main ORFs and that the moss genome has more than 100 putative
292 bicistronic and three polycistronic genes with detectable translation products. Moreover, the
293 translation of sORFs and proteins located together in the same locus might be regulated in a tissue-
294 specific manner.

295

296 **Most translatable sORFs are not evolutionarily conserved**

297 Analysis of the evolutionary conservation of sORFs is often a key step in revealing biologically active
298 sORFs (Andrews and Rothnagel 2014). To determine whether the translatable sORFs were more
299 highly conserved than the other sORFs, we analyzed the intactness of these sORFs in the
300 reconstructed genomes of three *P. patens* ecotypes, 'Villersexel', 'Reute' and 'Kaskasia', as well as the
301 ten abovementioned species. We found that 19 (3.2%) of 602 translatable sORFs in the ecotypes
302 either lost the start/stop codon or had a frameshift or premature termination codon (PTC). This
303 number was not significantly different from the number (2.4%) occurring by chance suggesting that
304 sORF translation does not disrupt trends of sORF elimination in these ecotypes.

305 To investigate whether the trend in translatable sORF evolution differs from that of the other sORFs,
306 we estimated the age (number of species in which homologs can be found) and the selection
307 pressure (Da/Ds) on translatable sORFs on an evolutionary timescale using the transcriptomes of the
308 ten abovementioned species. Overall, we found 74 sORFs had evidence of translation and
309 conservation in at least one species while only 11 were under negative selection (Ka/Ks << 1)
310 (Supplemental Figure S5).

311 Sixty-four of these were CDS-sORFs or interlaced-sORFs. These results point to a high level of
312 conservation of these sORFs, which is apparently connected to the conservation of overlapping
313 protein-coding genes. Although conservative sORFs were significantly enriched in a set of
314 translatable sORFs (Fisher's exact test, P = 2.716567e-05), we found that most translatable sORFs
315 (525, 87.6%) were not conserved.

316 We next examined whether the translatable sORFs detected in this study share similarity
317 with a recently defined set of 13,748 putative SEPs in the *A. thaliana* (Hazarika et al. 2017). We
318 identified two sORFs (Chr20#13303500#-#88.2 (uORF), Chr11#14549091#+#97.9 (CDS-sORF))
319 with evidence of translation according to our MS analysis that shared similarity with ARA-PEP
320 peptides (e-value < 0.01), implying that these sORFs are evolutionarily conserved and may produce
321 peptides in *A. thaliana* cells.

322

323 **Alternative splicing regulates the number of sORFs in protein-coding transcripts**

324 Alternative splicing (AS) events may lead to the specific gain, loss or truncation of different groups of
325 sORFs located on the transcripts of the same gene. For example, AS can generate sORFs that are
326 truncated version of proteins (see below). We found 6092 alternatively spliced sORFs (AS-sORFs)
327 belonging to transcripts from 4389 genes. CDS-sORFs were significantly overrepresented (Figure
328 4A), while interlaced-sORFs, uORFs and dORFs were significantly underrepresented among AS-
329 sORFs compared to the control dataset (AS-REF). The number of translatable sORFs in a set of AS-
330 sORFs did not significantly differ from that expected by chance (Fisher's exact test p-value=0.9423),
331 suggesting that AS does not preferentially occur in peptide-encoding sORFs.

332 We randomly selected ten different translatable AS-sORFs and searched for the
333 corresponding isoforms with/without sORFs in the transcriptomes of three types of moss cells. RT-
334 PCR analysis revealed the transcription of these isoforms, confirming that they could indeed be
335 translated (Supplemental Figure S6). Moreover, four sORFs contained isoforms showing tissue-
336 specific transcription. These observations led to the hypothesis that the translation of sORFs is
337 extensively regulated by AS.

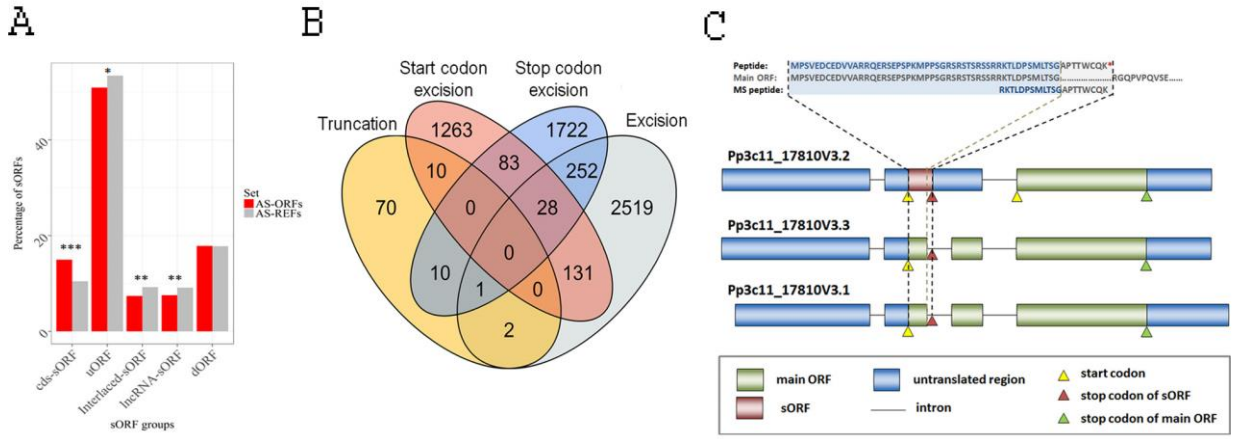



**Figure 4. Alternative splicing regulates the expression of sORFs.** A – Enrichment analysis of different sORF groups in a set of AS-sORFs and AS-REFs. P-value was calculated by Fisher's exact test. ***P < 1e-10; **P < 0.001; *P < 0.05; B – Venn diagram showing the number of AS-sORFs influenced by different AS events. C – Example of a translatable CDS-sORF, which was generated by an AS event and partially overlaps with the main ORF of Pp3c11_17810. Intron retention caused the formation of the isoform with the sORF, while splicing of this intron led to the excision of the sORF stop codon and its disruption. MS detection of the peptide located at the exon-AS-intron junction allowed the translation of the sORF to be unambiguously distinguished from the translation of the main ORF. Upper panel shows the amino acid sequence of the sORF-encoded peptide, MS detected peptide and (partial) protein translated from the main ORF. Black and gray dotted lines mark the borders of the sORF and the canonical intron start site, respectively. The intron-exon structure of three transcript isoforms of the gene was retrieved from Phytozome (v12).

We then performed GO enrichment analysis of the genes carrying AS-sORFs and found that they were significantly enriched (Fisher's exact test, p-value < 0.01) for 13 GO terms. Ten GO terms linked with nucleic acid binding (GO:0001071, GO:0003700), signal transducer activity (GO:0004871), aminopeptidase activity (GO:0004177), transferase activity (GO:0003950, GO:0016772, GO:0016775) and kinase activity (GO:0004672, GO:0004673, GO:0000155) were specifically enriched in a set of AS-sORF-carrying genes. These overrepresented GO terms demonstrate that alternative splicing does influence the sORF landscape for regulatory genes and suggest that sORFs play an important role in the regulation of their translation.

360 We then classified the events leading to changes in sORF sequences into four groups: 1)
361 truncation, if the middle part of the sORF was excised; 2) stop codon excision; 3) start codon excision
362 and 4) excision, if the complete sORF was removed from an isoform. We found that half of the sORFs
363 (48%, 2933) had undergone complete excision from transcripts, whereas only 93 sORFs were
364 truncated (Figure 4B). Moreover, the complete excision of sORFs occurred significantly more
365 frequently in uORFs than in the other sORF groups (57% vs. 20–44%, Fisher's exact test P-value <
366 1e-05). In addition, in the set of AS-sORFs with complete excision, evolutionarily conserved sORFs
367 (conserved in >1 species) were significantly underrepresented (6.76e-42) compared to the other
368 sets of AS-sORFs ("Truncation", "Stop codon excision", "Start codon excision"). Thus, our analysis
369 demonstrated that AS and evolutionary forces lead to the excision of sORFs from the transcriptome
370 and genome of *P. patens*.
371

372 ### The role of sORFs in modulating protein–protein interactions

373 Protein–protein interactions (PPI) are critical for the formation of higher order protein complexes.
374 Competitive inhibitors of PPI are referred to as MicroProteins (miPs) or small interfering peptides
375 (siPEPs) (Seo et al. 2011; Eguen et al. 2015). These proteins, which are usually small, can be
376 generated by alternative splicing or evolutionarily generated by domain loss (Staudt and Wenkel
377 2011; Eguen et al. 2015). We hypothesized that sORFs with similarity to known proteins could
378 impair the functions of these proteins by mimicry or by regulating the activity of proteins translated
379 from the main ORF. To identify such sORFs, we performed BLASTP (E-value < e-5) similarity
380 searches between the encoded amino acid sequences of sORFs and the annotated proteins of *P.*
381 *patens.* We identified 363 sORFs resulting from AS events that partially overlapped with the main
382 ORF, thereby generating truncated versions of the proteins (cis-sORFs, see in Supplemental Table
383 S4). First, we analyzed how many cis-ORFs contained known complete or incomplete protein
384 domains, finding that 60 sORFs harbored IDRs, while 30 cis-sORFs contained parts of 28 different
385 domains (Supplemental Table S4). Among these, we observed the protein kinase domain (PS50011,
386 Chr13#1404821#+#28.4), protein tyrosine kinase (PF07714, Chr11#4429996#+#64.5) and MYB-
387 like DNA-binding domain (TIGR01557, Chr19#1622814#+#55.3). Potential SEPs from these sORFs

388 can be considered potential candidate microPeptides (Straub and Wenkel 2017). Interestingly, the
389 genes containing cis-sORFs were enriched in kinase and kinase-like domains. GO enrichment analysis
390 also revealed significant overrepresentation of terms associated with protein modifications, such as
391 GO:0006468 (protein phosphorylation) and GO:0036211 (protein modification process).
392 Another group of peptides carrying similarity with DNA binding and protein-protein interaction
393 domains of transcription factors is small interfering peptides. Because similarity with transcription
394 factors (TFs) domains they can act as dominant-negative repressors of TFs. Among genes containing
395 cis-sORFs, we found some moss small interfering peptides that had similarity to putative
396 transcription factor genes such as genes encoding GROWTH-REGULATING FACTOR (e.g.,
397 Pp3c20_10590), C2H2 zinc finger domain containing (e.g., Pp3c1_16920)), BTB/POZ domain
398 containing (e.g., Pp3c16_9230), B3 DNA binding domain containing (e.g., Pp3c7_7990) and MYB-CC
399 type transcription factor (e.g., Pp3c21_2850).
400 To obtain evidence for the translation of these sORFs, we analyzed MS data and found at least
401 two examples (Figure 4C). A few detected translatable cis-sORFs could be explained by a significant
402 overlap with the protein sequences, whereas we filtered out the 'ambiguous' PSMs. Moreover, the
403 formation of a premature termination codon (PTC), for example, as a result of intron retention
404 events, leads to mRNA decay (Ge and Porse 2014; Karousis et al. 2016). Thus, we suggest that
405 peptides from the cis-sORFs are produced by moss cells and accordingly, they might be involved in
406 cis regulation of main ORF protein activity or have distinct functions.
407 We identified 272 sORFs that shared similarity with annotated proteins but were located on
408 other transcripts (trans-sORFs, see in Supplemental Table S4). The translation of six trans-sORFs was
409 confirmed by our MS data. We found 36 potential trans-SEPs with similarity to known protein
410 domains (Supplemental Table S4). Trans-sORFs may have originated through the divergence of
411 ancient paralogous genes, which occurred after the paleo duplication of the moss genome (Rensing et
412 al. 2007; Rensing et al. 2008). In fact, 159 (58.5%) trans-sORFs shared similarity to genes from at
413 least one species (age 1-10). In addition, all of these trans-sORFs are under strong purifying selection
414 (dN/dS << 1). The low rate of evolution of trans-sORFs suggests that the encoded peptides have
415 important biological functions and might be involved in proteomic networks based on their

416 similarity to functional proteins. Interestingly, we observed significantly fewer AS-sORFs in the set of
417 trans-sORFs, suggesting that mechanisms other than AS are responsible for their formation. We then
418 investigated which trans-sORFs share similarity to large gene families. Several distinct clusters with
419 sORF-encoded peptides sharing similarity with more than four proteins from distinct genes were
420 detected (Supplemental Figure S7). Each cluster encompasses genes from different protein families,
421 including one containing leucine-rich repeat and zinc-finger domains involved in protein–protein
422 and protein–nucleic acid interactions, respectively. Potential SEPs from these clusters share
423 similarity with the respective domains and are therefore considered to be potential candidate
424 microPeptides (Straub and Wenkel 2017).

425 We then addressed two questions: 1) Are proteins with higher numbers of interactions
426 overrepresented among those with BLASTP hits; and 2) Is there a connection between the
427 expression of sORFs and similar proteins? To answer the first question, we identified orthologous
428 genes in the *A. thaliana* genome and used data from the interactome database (AtPID) (Lv et al.
429 2017). However, we did not detect a significant difference from the set of randomly selected genes.
430 To determine whether sORF-protein pairs more frequently coexist in a cell, we examined the
431 coexpression data and compared the distribution of correlation coefficient values with those from
432 randomly selected pairs (10 iterations) of genes. On average, these sORF-protein pairs had higher
433 correlation coefficients than randomly selected gene pairs (Wilcoxon Rank Sum and Kolmogorov-
434 Smirnov Tests P-value < 0.05), implying that sORF-bearing and target genes are frequently
435 coexpressed.

436 **SEPs regulate moss growth**

437 Despite the recent finding that 10% of overexpressed intergenic sORFs have clear phenotypes in
438 Arabidopsis (Hanada et al. 2013), the functions of most sORFs and SEPs in plants are generally
439 unknown. Known bioactive SEPs in plants are encoded by sORFs located on short non-protein-coding
440 transcripts, which can be referred to as lncRNAs (Rohrig et al. 2002; Chilley et al. 2006). In this
441 context, it would be intriguing to determine how many plant lncRNAs encode peptides, as well as the
442 biological functions of these SEPs. Our pipeline allowed us to identify hundreds of translated sORFs,
443 including those encoded by lncRNAs. Some of these lncRNA-sORFs showed tissue-specific

transcription and translation patterns, while others were expressed in all types of moss cells (Figure 3C). We reasoned that stably expressed lncRNA-sORFs can produce peptides that play fundamental roles in various cellular processes. To explore this hypothesis, we examined the impact of overexpression and knockout of these lncRNA-sORF sequences on moss morphology. Here, we present an example of such analysis using a 41-aa peptide (SEP1) encoded by the stably expressed lncRNA-sORF Chr09#2770841#+#59.21 (Figure 3C).

We obtained multiple independent mutant lines and confirmed the expression of the transgenes or the knockout of sORF sequences (Supplemenatl Figure S8). The overexpression as well as knockout of sORF sequences led to clear morphological changes, implying that these peptides play a role in regulating filamentous architecture in *P. patens* (Figure 5). The overexpression of *SEP1* induced the formation of long caulonema cells compared with the wild type and knockout lines (Figure 5B and 5C). Moreover, there was a significant difference between the growth rates of the wild type and *SEP1* mutant lines (Figure 5D).

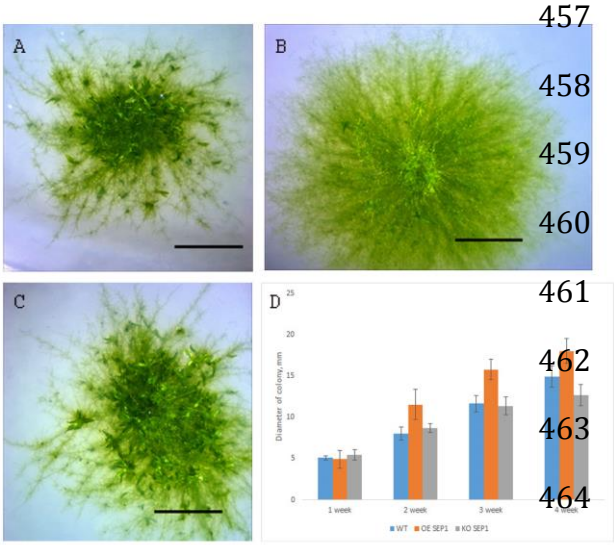
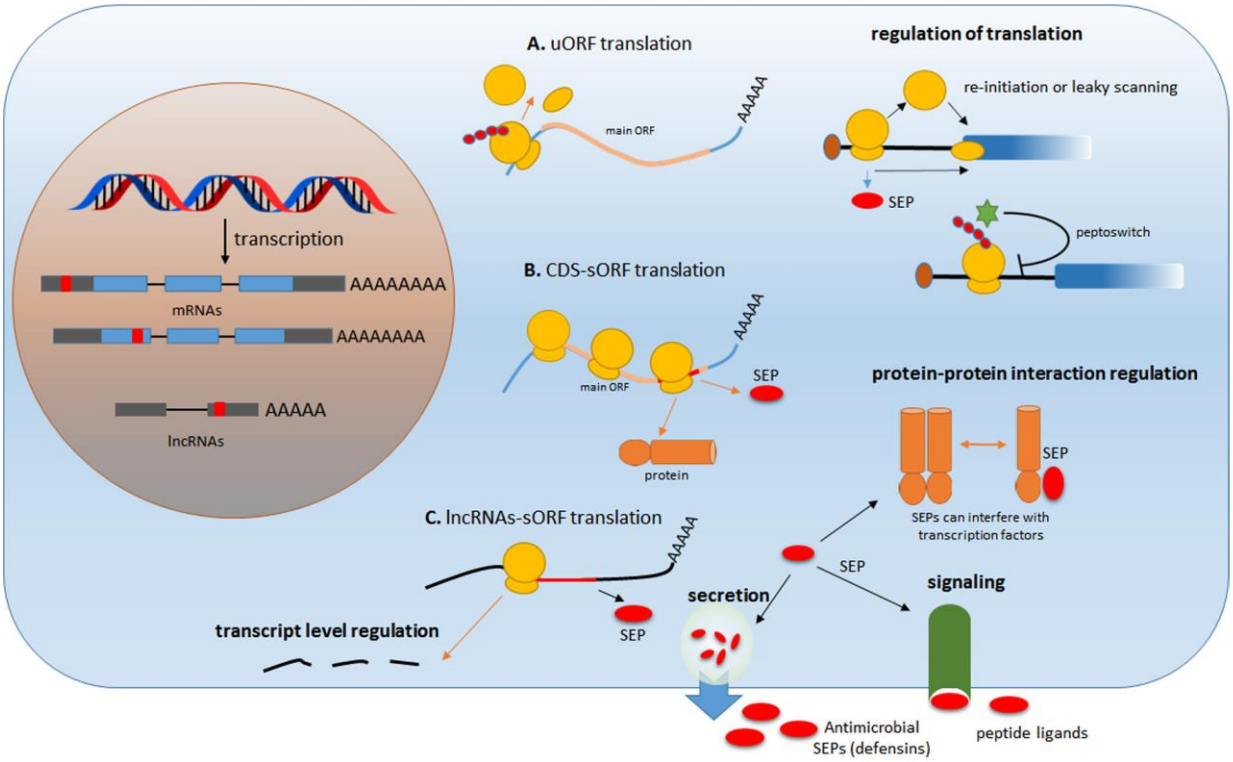



**Figure 5. Morphology of wild type and mutants lines grown on BCD-medium.** A – WT; B –overexpression mutant line; C – knockout mutant line; Scale bar: 50 mm; D – The diameter of WT and mutant colonies on BCD medium.

The knockout of the *SEP1* sequence led to moss protonemata with a reduced growth rate and a significant delay in senescence. Conversely, the *SEP1* overexpression lines were characterized by rapid growth and senescence compared to the wild type and knockout lines. Taken together, our findings suggest that lncRNA-sORFs can influence growth and development in moss and that our pipeline allows biologically active peptides to be identified.

472 **DISCUSSION**

473 Although functionally characterized SEPs have been shown to play fundamental roles in key
474 physiological processes, sORFs are arbitrarily excluded during proteome annotation. Given the
475 difficulty in identifying translatable, functional sORFs, we know little about their origin, evolution
476 and regulation in the genome. In the present study, we investigated the abundance, evolutionary
477 history and possible functions of sORFs in the genome of the model moss *Physcomitrella patens*. The
478 use of an integrated pipeline that includes transcriptomics, proteomics and peptidomics data allowed
479 us to identify hundreds of translatable sORFs in three types of moss cells. We propose that several
480 distinct classes of sORFs that differ in terms of their position on transcripts, the level of evolutionary
481 conservation and possible functions are present in the moss genome (Figure 6).




482

483 **Figure 6. Proposed functions of sORF-encoded peptides.** A – uORFs can function in the regulation
484 of translation of the downstream main ORF. The functions of peptides encoded by uORFs are
485 unknown, and most are likely to represent "noise" from protein translation; B – CDS-sORF-encoded
486 peptides can help regulate protein-protein interactions, and some interfere with the translation of
487 the main ORF; C – long non-coding RNAs or intergenic-sORFs can produce biologically active
488 peptides that perform signaling, defense or regulatory functions. In addition, the translation of sORFs

489 can activate the nonsense-mediated decay (NMD) mechanism, which leads to the degradation of the
490 corresponding transcripts.
491
492 According to our MS data, the translation patterns of most sORFs tend to be tissue specific. Our
493 results suggest that the evolutionary rates of various types of sORFs differ, with weak overall
494 conservation and fast elimination from a genome. We also showed that alternative splicing is an
495 additional mechanism to control sORF expression in plant cells, changing their sequences at the
496 transcriptome level. Finally, our results suggest that stably expressed sORFs located on lncRNAs can
497 play important roles in plant growth.

498 **sORFs with high coding potential are not conserved among genomes**

499 Even though the analysis of sequence conservation is somewhat biased against the detection of short
500 sequences (Ladoukakis et al. 2011), this technique is widely used to select candidate functional
501 sORFs. Although analyzing the conservation of short amino acid sequences is not trivial (Moyers and
502 Zhang 2016), hundreds of conserved sORFs have recently been identified in plants, yeast and
503 animals (Ladoukakis et al. 2011; Hanada et al. 2013; Mackowiak et al. 2015). The number of sORFs
504 conserved in the plant kingdom is undoubtedly underestimated due to the low sensitivity of tools
505 used for conservation analysis and the limited number of available sequenced genomes from closely
506 related species. Our pipeline allowed us to identify 5034 conserved sORFs among the transcriptomes
507 of ten different plant species, 74 of which showed evidence of translation according to our MS data.
508 However, we suggest that the possibly functional sORFs might significantly outnumber the
509 conserved ones.

510 Despite the observation that approximately 1% of uORFs and dORFs had evidence of
511 translation, a significant "loss" of these sORF types was observed among the closest related species.
512 We even detected rapid inactivation of uORFs and dORFs (756 sORFs) in the reconstructed genomes
513 of three *P. patens* ecotypes due to the disruption of start or stop codons. As the occurrence of sORFs
514 downstream or upstream of the main ORF can be deleterious to its translation, we cannot rule out
515 the possibility that this may cause strong selection pressure and the rapid elimination of uORFs and
516 dORFs (Iacono et al. 2005; Neafsey and Galagan 2007; Johnstone et al. 2016). Moreover, we observed

517 significant depletion (Fisher's exact test P-value = 5.25e-13) of uORFs and dORFs in a set of
518 translatable conservative sORFs. Taken together, these findings suggest that sORFs located in
519 untranslated regions are evolving rapidly and may play a regulatory roles rather than encoding
520 bioactive peptides.

521 In a recent study, more than 1000 alternative proteins were experimentally detected by
522 mass-spectrometry in human cell lines (Vanderperre et al. 2013). In *P. patens*, we found tens of
523 thousands of sORFs (CDS-sORFs) that overlapped with the CDS of protein-coding genes, 306 of which
524 were translatable. The evolutionary origins, functions and mechanisms of translation of such
525 alternative sORFs are still unclear. According to our results, a significant number of CDS-sORFs are
526 under positive selection and are dramatically eliminated from the genomes of distant species. This
527 process must not be stochastic, as opposite results were obtained for randomly selected CDS. The
528 speed and direction of the evolution of protein-coding genes depends on their level of transcription,
529 biological functions, protein-protein connectivity and functional redundancy (Hirsh and Fraser 2001;
530 Zhang and Li 2004; Julenius and Pedersen 2006; Enard et al. 2014). Moreover, only specific domains
531 or regions of protein-coding sequences can evolve under positive selection (Montoya-Burgos 2011).

532 The evolution of CDS-sORFs is undoubtedly an expensive process for the cell, as these
533 elements may be located in regions encoding protein domains and influence the structure and
534 function of the protein encoded by the main ORF (Cherry 2010). In addition to being located in
535 regions encoding functional domains, CDS-sORFs can be generated in fast-evolving regions of genes
536 (e.g., those encoding protein disordered regions). We found both CDS-sORFs originated from regions
537 associated with known protein domains and CDS-sORFs from disordered regions. Indeed, a
538 comparison among the conserved (homologs found in >1 species) sORFs of the number of the two
539 types of CDS-sORFs in similar species revealed significant differences, with higher conservation for
540 CDS-sORFs originated from protein domain-encoding regions. These results indicate that the
541 evolution of CDS-sORFs depends on their locations insight main CDS sequence.

542 In the current study, we found that both the transcription and translation of CDS-sORFs
543 occurred in a tissue-specific manner. Protein-coding genes with tissue-specific transcription patterns
544 and functional redundancy of the gene product are often under positive selection (Zhang and Li

545 2004; Montoya-Burgos 2011). This finding, together with other properties of CDS-sORFs, such as
546 their overlap with particular parts of protein-coding sequences, might explain the high turnover rate
547 of CDS-sORFs. However, whether sORFs are preferentially generated in fast-evolving regions of
548 proteins or whether the selective pressure on sORFs leads to changes in protein-coding sequences is
549 still unknown.

550 **Analysis of sORF translation: approaches that makes sense**

551 Clear evidence of transcription and translation points to the biological significance of sORFs. Thus,
552 identifying translatable sORFs is an important step that could lead to the discovery of new biological
553 functions. Ribosome profiling provides a direct readout of the ribosome occupancy of different
554 transcripts, thereby providing a measure of the level of translation. According to ribosome profiling
555 data from a wide variety of species, translation appears to occur in a pervasive manner (Ingolia et al.
556 2011; Guttman et al. 2013; Bazzini et al. 2014; Couso and Patraquim 2017). These observations have
557 led some researchers to conclude that the majority of non-coding RNAs (e.g., lncRNAs) in cells can be
558 translated and produce peptides (Crappe et al. 2013; Bazzini et al. 2014; Housman and Ulitsky 2016).
559 In addition to lncRNAs, the ribosome occupancy of short frames in the UTRs of mRNA has also been
560 investigated (Weatheritt et al. 2016). However, ribosome-profiling data alone are not sufficient to
561 classify transcripts as coding or noncoding (Guttman et al. 2013). Thus, alternative approaches such
562 as proteomics and peptidomics should be used to investigate the translation of sORFs (Slavoff et al.
563 2013; Ma et al. 2016). Comparisons of ribosome profiling and mass spectrometry results have led to
564 the conclusion that MS detects peptides arising from the most highly translated sORFs (Aspden et al.
565 2014; Bazzini et al. 2014). However, a recent study showed that there are no technical obstacles to
566 the detection of lncRNA-encoded peptides by mass spectrometry (Verheggen et al. 2017).
567 Nonetheless, peptide detection by MS analysis can be difficult, for example, due to the location of
568 SEPs in cellular membranes (Andrews and Rothnagel 2014; Couso and Patraquim 2017). Mass-
569 spectrometry studies have thus far confirmed the presence of a few dozen SEPs in the peptidomes of
570 animal cells (Slavoff et al. 2013; Prabakaran et al. 2014; Mackowiak et al. 2015; Ma et al. 2016). In the
571 present study, we found evidence for the translation of only a small portion of the predicted sORFs.

572 Moreover, we confirmed the presence of only a few sORFs in all three cell types. Therefore, it is
573 difficult to estimate the full extent of the presence of sORFs in a cell.

574 In previous studies, only standard proteomics analysis was used to identify SEPs. We
575 reasoned that analyzing endogenous peptide pools instead of tryptic peptides has several
576 disadvantages in terms of SEP identification: 1) standard proteomic approaches are not suitable for
577 the isolation and analysis of small and low-abundance peptide molecules; and 2) SEPs are shorter
578 than standard proteins and it is unlikely that more than one tryptic fragment will be detected in a
579 single proteomic experiment. Moreover, peptidomic approaches can theoretically be used to identify
580 full-length SEPs in a cell. We firstly used endogenous peptides pools to detect SEPs and according to
581 our data the values of PSMs, supporting expression of SEPs, were better in "peptidomic" dataset.
582 Moreover, some SEPs were confirmed by several endogenous peptides (up to 17), that an increase
583 the reliability of their detection. Notably, we did not observe any significant overlap between the
584 sORFs detected using proteomic and peptidomic approaches. Thus, our study demonstrates the
585 advantage of using complementary approaches for building a complete list of SEPs.

586 **Functionality of SEPs**

587 It was recently suggested that sORFs are randomly generated in a genome (Couso and Patraquim
588 2017). We detected more than 600,000 sORFs with high coding potential in the moss genome.
589 Assuming the average length of an sORF is approximately 60 bp and distinct sORFs are not
590 overlapping, these elements occupy a substantial portion of the moss genome. This raises the
591 following question: to what extent are sORFs present in the transcriptome and (even more
592 interesting) the proteome of a cell? Some of these sORFs are translated into peptides, suggesting they
593 might contribute to physiological processes, but the extent of selective pressure on these elements
594 has been unclear.

595 We identified hundreds of translatable sORFs of different types and suggested various
596 functions for these types of sORFs (Figure 6). The average conservation of sORFs within 5′ leader
597 sequences (uORFs) is low (Aspden et al. 2014); uORFs are thought to regulate the translation of the
598 downstream ORF (Johnstone et al. 2016). Based on our conservation analysis and MS data, we
599 suggest that the majority of uORFs and dORFs play regulatory roles instead of encoding peptides

600 (Figure 6A). By contrast, CDS-, interlaced- and lncRNA-sORFs have greater potential to encode
601 bioactive peptides, as they are more highly conserved, frequently contain protein domains and,
602 according to the MS data, often produce peptides. However, the functions of the peptides are unclear
603 and require more detailed investigations. One possible function of sORF-encoded peptides that are
604 similar to known proteins is mimicry of the main protein function.

605 MiPs (or siPEPs) are small, usually single domain proteins (Seo et al. 2011; Eguen et al.
606 2015). Some miPs are important modulators of protein–protein and protein–DNA interactions that,
607 for example, prevent the formation of functional protein complexes (Seo et al. 2013; Graeff et al.
608 2016). Most known miPs/siPEPs are small proteins evolutionarily generated by domain loss (Eguen
609 et al. 2015). We suggest that the potential for sORFs that overlap with the CDS of protein-coding
610 genes to be a source of small interfering peptides is currently underestimated (Figure 6B). Based on
611 the analogy of cis-miPs generated by alternative splicing events (Eguen et al. 2015), we refer to these
612 SEPs as cis-SEPs. In the present study, 363 sORFs encompassing parts of the main ORFs and
613 originating from AS were identified. According to the MS data, some of these sORFs have evidence of
614 translation. Analyzing sequence similarity to known proteins or the presence of particular domains
615 can be useful for predicting peptide function. We found that approximately 30% of cis-SEPs harbor
616 protein domains such as protein kinase domains and MYB-like DNA-binding domain or IDRs. This
617 finding points to the high regulatory potential of these SEPs resulting from their interference with
618 functional proteins (Eguen et al. 2015). At the same time, it should be noted that protein domains or
619 incomplete protein domains in isolation can have functions unrelated to those observed in multi-
620 domain proteins (Kelley and Sternberg 2015).

621 We also found SEPs sharing homology with proteins produced from other genes that likely
622 originated through the divergence of ancient paralogs. Perhaps SEPs with similarity to other
623 proteins, such as those translated from CDS-sORFs, can perform functions like those of miPs/siPEPs
624 and regulate the activity of protein or protein complexes. Indeed, we found that genes harboring
625 CDS-sORFs were enriched in GO terms connected to protein binding and transferase activity. Also,
626 some sORFs with disordered regions might mediate protein–protein or protein–nucleic acid

627 interactions, as suggested previously (Mackowiak et al. 2015). Taken together, these findings suggest
628 that sORFs may strongly interfere with protein interactions.

629 In this study, we explored several groups of sORFs, including those encoded by lncRNAs. The
630 translation of peptides from lncRNAs is intriguing, and there is some evidence that these peptides
631 play important biological roles in various processes (Kondo et al. 2010; Magny et al. 2013;
632 Matsumoto et al. 2017). Nevertheless, the biological functions of most lncRNA-sORF-encoded
633 peptides are currently unclear, especially those in the plant kingdom (Tavormina et al. 2015). For
634 example, it is unknown whether there is a connection between the functions of lncRNAs and the
635 peptides they encode. The transcription of the non-coding portions of the genome into lncRNAs is
636 thought to give rise to the translation of sORFs located within them. In this case, some of these
637 peptides would not be vital but may be important for survival under certain conditions by serving as
638 a raw material for evolution (Figure 6C). According to our data, the number of sORFs located on
639 single lncRNAs varied from 1 to 28. Thus, the translation of lncRNAs can potentially lead to the
640 production of many peptides in a cell. However, the opportunity for the translation of lncRNAs and
641 the subsequent stability of such peptides in the cell are under debate (Saric et al. 2004; Loose et al.
642 2007; Housman and Ulitsky 2016). There are only a few examples of lncRNA-encoded peptides
643 involved in signaling and plant growth. For example, POLARIS (PLS), encoding a 36-amino acid
644 peptide, is required for correct root growth and vascular development in Arabidopsis (Chilley et al.
645 2006). In the current study, we confirmed the translation of 42 SEPs encoded by lncRNAs. Plants
646 overexpressing an lncRNA-encoded peptide (41 aa) showed clear phenotypic differences from wild-
647 type plants, suggesting its possible role in regulating cell growth and development. Our results lay
648 the groundwork for systematic analysis of functional peptides encoded by sORFs.

649 The conservation rate of lncRNA- sORFs is not high. Moreover, most lncRNA-encoded
650 peptides are not conserved. In our view, they might serve as a reservoir of peptide activity (Bao et al.
651 2017) for adaptation processes and/or as a source for *de novo* generation of new genes, a topic that
652 is currently under intensive debate (Couso and Patraquim 2017).

653 Our analysis led to several conclusions concerning the extent, evolution and possible
654 biological functions of sORFs in plants. We discovered that most sORFs, including translatable ones,

655 are not widely conserved and that the most slowly evolving group overlaps with the CDS of protein-
656 coding genes. We demonstrated the role of alternative splicing in shaping the sORF landscape in
657 terms of transcripts, as well as isoform-specific transcription of sORFs. We identified a group of
658 sORFs homologous to known protein domains and suggested they function as small interfering
659 peptides. Finally, we demonstrated the functional potential of one peptide encoded by an lncRNA.
660 The high potential regulatory activity of peptides, high evolutionary rate and wide translation of
661 sORFs suggest that they may provide a reservoir of potentially active molecules and that some of
662 sORFs can give rise to new protein-coding genes. We provide a resource of putatively functional
663 peptides for further analysis.

664 
## METHODS

665 
### *Physcomitrella patens* growth conditions

666 *Physcomitrella patens* protonemata were grown on BCD medium supplemented with 5 mM
667 ammonium tartrate (BCDAT) under continuous white light at 25$^{0}$C in 9-cm Petri dishes (Nishiyama
668 et al. 2000). For all analyses, the protonemata were collected every 5 days. The gametophores were
669 grown on free-ammonium tartrate BCD medium under the same conditions, and 8-week-old
670 gametophores were used for analysis. Protoplast was prepared from protonema as described
671 previously (Fesenko et al. 2015).
672 For morphological analysis, samples of fresh protonemal tissue 2 mm in diameter were
673 inoculated on BCD and BCDAT Petri dishes. For colony growth rate measurements, photographs
674 were taken at 7 d intervals over 42 days. Colonies and protonema cells were photographed using a
675 Microscope Digital Eyepiece DCM-510 attached to a Stemi 305 stereomicroscope or Olympus CKX41.
676

677 
### Identification of coding sORFs in the *P. patens* genome

678 To identify sORFs with high coding potential, the sORFfinder (Hanada et al. 2010) tool was utilized.
679 Intron sequences and CDS were used as negative and positive sets, respectively. Intron sequences
680 and CDS were extracted from the *P. patens* v3.3 genome (Phytozome v12) by parsing a gff3 file of the

681 moss genome annotation using custom-made python scripts (available upon request), followed by
682 DNA sequence extraction using the subcommand getfasta from bedtools (Quinlan and Hall 2010).
683 These sequences were used to train sORFfinder as described in the user manual. Parameter -d was
684 set to "b" for searching the sORF sequences in both direct and reverse orientation. The search for
685 sORFs was performed using the whole genome sequence of *P. patens* (release 3.3,
686 https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ppatens). Of the 6,706,696 sORF
687 sequences found in the genome, 786,439 had high coding potential according to the sORF finder. To
688 eliminate sORFs that are transcribed, located in the exons of transcripts and have introns, a bed file
689 was generated using a custom-made python script and intersected with exon positions extracted
690 from a gff3 file of *P. patens* genome annotations. To identify intergenic-sORFs, the bed file was also
691 intersected with transcribed regions determined based on our RNAseq data (Fesenko et al. 2017).
692 Using an R script, sORFs fully overlapping with exons were selected; 75,685 sORFs remained after
693 this step. Identical sORFs were removed from the dataset. In addition, sORFs overlapping repetitive
694 regions identified by RepeatMasker, as well as sORFs comprising parts of annotated *P. patens*
695 proteins, were also removed from the dataset, resulting in a final dataset of sORFs comprising 70,095
696 sequences.
697

698 **sORF classification**

699 The step-by-step procedure performed for sORF classification is illustrated in Supplemental Figure
700 S9. In the first step, lncRNA-sORFs were identified by searching for identical sORFs in known lncRNA
701 databases, including CantataDB (Szczesniak et al. 2016), GreenC (Paytuvi Gallart et al. 2016) and our
702 previously published moss dataset (Fesenko et al. 2017). After this sORF bed file was intersected
703 with moss genome annotation, the locations of the sORFs on transcripts were determined, resulting
704 in the further classification of genic-sORFs into uORFs, dORFs, CDS-sORFs and interlaced-sORFs.
705 Because alternative splicing leads to inaccuracy in genome annotation, the locations of a
706 subset of genic-sORFs cannot be unambiguously classified, as they can be located in different regions
707 in different isoforms of the same gene. All sORFs located on transcripts that were not annotated in
708 the *P. patens* genome but were identified using our RNAseq data were classified as intergenic-sORFs.

709 To detect alternatively spliced sORFs (AS-sORFs), a bed file with sORF locations was
710 intersected with a bed file containing intron coordinates for all isoforms. Those sORFs that
711 overlapped for both exons (see above) and introns were classified as AS-sORFs.
712

713 ### Evolutionary conservation analysis

714 The transcriptomes of nine plant species were downloaded from Phytozome v12: *Sphagnum fallax*
715 (release 0.5), *Marchantia polymorpha* (release 3.1), *Selaginella moellendorffii* (release 1.0), *Spirodela*
716 *polyrhiza* (release 2), *Arabidopsis thaliana* (TAIR 10), *Zea mays* (Ensembl-18), *Oryza sativa* (release
717 7), *Volvox carteri* (release 2.1) and *Chlamydomonas reinhardtii* (release 5.5). The transcriptome of
718 *Ceratodon purpureus* was *de novo* assembled using Trinity (Haas et al. (2013)). To identify
719 transcribed homologous sequences, tBLASTn (word size = 3) was performed using sORF peptide
720 sequences as queries and the transcriptome sequences of the abovementioned species as subjects.
721 The following cutoffs parameters were used to distinguish reliable alignments: E-value < e-5 and
722 query coverage > 60%.

723 Pairwise Ka/Ks ratios (equivalent to dN/dS) were calculated using the codeml algorithm
724 with PAML software (Yang 2007). The calculation procedure, which was facilitated using a custom-
725 made python script (available under request), included alignment extraction from the tBLASTn
726 output, PAL2NAL (Suyama et al. 2006) correction of the nucleotide alignment using the
727 corresponding aligned protein sequences and calculation of Ka/Ks ratios using codeml. The script
728 implements packages from biopython (Cock et al. 2009).

729 To estimate homologous sORF lengths, a python script (available under request) was
730 designed. The script uses tBLASTn alignment output and estimates the presence of in-frame start
731 and stop codons within (as well as downstream and upstream of) alignment regions. If a stop codon
732 was found upstream of an alignment region, it was considered to be a premature termination codon
733 (PTC). Otherwise, start and stop codons closest to the alignment region were used for homologous
734 sORF length calculation.

735 ### GO enrichment analysis

736 GO enrichment analysis was performed using the topGO bioconductor R package using the Fisher's
737 exact test in conjunction with the 'classic' algorithm (false discovery rate [FDR] < 0.05). Gene
738 Ontology (GO) terms assigned to *P. patens* genes were downloaded from Phytozome. Only GO terms
739 containing >5 genes in a background dataset were considered in the enrichment analysis. Redundant
740 GO terms were removed using the web-based tool REVIGO (Supek et al. 2011).

741

742 ### Peptide and protein extraction

743 Endogenous peptide extraction was conducted as described previously (Fesenko et al. 2015).
744 Proteins were extracted as described previously (Fesenko et al. 2016). Four biological repeats for
745 gametophores, four for protonemal and four for protoplast samples were used.

746 ### Mass-spectrometry analysis and peptide identification

747 Mass-spectrometry analysis was performed using three biological and three technical repeats for the
748 proteomic (Fesenko et al. 2017) and peptidomic datasets. Analysis was performed on two different
749 mass spectrometers: a TripleTOF 5600+ mass spectrometer with a NanoSpray III ion source
750 (ABSciex,Canada) and a Q Exactive HF mass spectrometer (Q Exactive HF Hybrid Quadrupole-
751 Orbitrap mass spectrometer, Thermo Fisher Scientific, USA). For the Q Exactive HF mass
752 spectrometer (Thermo Fisher Scientific, USA), peptide samples were separated by high-performance
753 liquid chromatography (HPLC, Ultimate 3000 Nano LC System, Thermo Scientific, USA) in a 15-cm
754 long C18 column with a diameter of 75 µm (Acclaim® PepMap™ RSLC, Thermo Fisher Scientific, USA).
755 The peptides were eluted with a gradient from 5–35 % buffer B (80 % acetonitrile, 0.1 % formic
756 acid) for 45 min at a flow rate of 0.3 µl/min. The total run time, including 10 min to reach 99% buffer
757 B, flushing 5 min with 99% buffer B and 10 min re-equilibration to buffer A (0.1% formic acid)
758 amounted to 70 min. Mass spectra were acquired at a resolution of 60,000 (MS) and 15,000 (MS/MS)
759 in a range of 400–1500 m/z (MS) and 200–2000 m/z (MS/MS). An isolation threshold of 67,000 was
760 determined for precursor selection and (up to) the top 10 precursors were chosen for fragmentation
761 via high-energy collisional dissociation (HCD) at 25 NCE and 100 ms activation time. Precursors with

762 a charged state of +1 were rejected, and all measured precursors were excluded from measurement
763 for 20 s.

764 Mass-spectrometry analysis on a TripleTOF 5600+ mass spectrometer with a NanoSpray III
765 ion source (ABSciex, Framingham, MA 01701, USA) coupled with a NanoLC Ultra 2D+ nano-HPLC
766 system (Eksigent, Dublin, CA, USA) was performed as described (Fesenko et al, 2016).

767 All datasets were searched individually with MaxQuant v1.5.8.3 (Tyanova et al. 2016) against
768 a custom database containing the protein sequences from Phytozome v12.0 merged with chloroplast
769 and mitochondrial proteins, a database of common contaminant proteins and the set of predicted
770 sORFs. MaxQuant's protein FDR filter was disabled, while 1% FDR was used to select high-confidence
771 PSMs, and ambiguous peptides were filtered out. Moreover, any PSMs with Andromeda scores of less
772 than 30 were discarded (to exclude poor MS/MS spectra). For dataset of endogenous peptides, the
773 parameter "Digestion Mode" was set to "unspecific" and modifications were not allowed. All other
774 parameters were left at default values. Features of the PSMs (length, intensity, number of spectra,
775 Andromeda score, intensity coverage and peak coverage) were extracted from MaxQuant's msms.txt
776 files.

777 **RT-PCR analysis of AS-sORFs**

778 Total RNA from gametophores, protonema and protoplasts was isolated as previously described
779 (Cove et al. 2009). RNA quality and quantity were evaluated via electrophoresis in an agarose gel
780 with ethidium bromide staining. The precise concentration of total RNA in each sample was
781 measured using a Quant-iT™ RNA Assay Kit, 5–100 ng on a Qubit 3.0 (Invitrogen, US) fluorometer.
782 The cDNA for RT-PCR was synthesized using an MMLV RT Kit (Evrogen, Russia) according to the
783 manufacturer's recommendations employing oligo(dT)17 -primers from 2 μg total RNA after DNase
784 treatment. The primers were designed using Primer-BLAST (Ye et al. 2012) (Supplementary Table).
785 The minus reverse transcriptase control (-RT) contained RNA without reverse transcriptase
786 treatment to confirm the absence of DNA in the samples. The RT-PCR products were resolved on an
787 1.5% agarose gel and visualized using ethidium bromide staining.

788 ### Generation of overexpression and knockout lines

789 For the overexpression experiments, the plant LIC vector (De Rybel et al. 2011) was used. PCR was
790 carried out using genomic DNA isolated from *P. patens* as a template and PEP4f and PEP4r as primers
791 (Supplemental Table S5). Amplicons were cloned into the pPLV27 vector (GenBank JF909480) using
792 the Ligation-independent (LIC) procedure (Aslanidis and de Jong 1990). The resulting plasmid was
793 named pPLV-Hpa-4FR. The nucleotide sequence of the cloned fragment was verified by sequencing
794 using a BigDye Terminator Cycle Sequencing Kit (v. 3.1) and AbiPrism 3730xl (Applied Biosystems,
795 USA). For moss transformation, pPLV-Hpa-4FR purified using a Qiagen Plasmid Maxi Kit (Qiagen,
796 Germany) and linearized with SacI (pPLV-Hpa-4FR-SacI) was utilized.

797 Knockout lines were created using the CRISPR-Cas9 system (Collonnier et al. 2017). The
798 plasmid containing the sgRNA expression cassette was generated with internal restriction enzyme
799 sites that permit rapid, directional cloning of 20-mer guide sequences. This cassette consisted of the
800 U6 promoter from *P. patens* and the tracrRNA scaffold with two BbsI sites between them
801 (Supplemental Figure S10). The cassette was synthesized from oligonucleotides and cloned into the
802 TA vector pTZ57R/T (Thermo Fischer Scientific, USA). The resulting plasmid was named pBB.

803 Coding sequences of *SEP1* were used to search for CRISPR RNA (crRNA) preceded by a PAM
804 motif in *S. pyogenes* Cas9 (NGG) using the web tool CRISPR DESIGN (http://crispr.mit.edu/). The 3-1
805 crRNA closest to the translation start site (ATG) was selected for cloning.

806 The plasmid pBB with the guide RNA expression cassette was linearized by digestion with
807 BbsI. Oligonucleotides were designed to contain compatible overhangs and a 20-mer guide sequence
808 (targeting *SEP1*, Supplementary Table 1). The hybridized oligonucleotides (2G Top-2G Bottom) were
809 ligated into the digested plasmid, yielding the final complete sgRNA expression cassette. The
810 resulting plasmid was named pBB-3-1.

811 The plasmids pACT-CAS9 (for CAS9 expression) and pBNRF (resistance to G418) were kindly
812 provided by Dr. Fabien Nogué. For moss transformation, the plasmids were purified using a Qiagen
813 Plasmid Maxi Kit (Qiagen, Germany).

814 In the overexpression experiments, protoplasts were transformed with 20 µg pPLV-Hpa-4FR
815 and circular DNA (7.5 µg each) from pBB-3-1, pACT-CAS9 and pBNRF and spread onto regeneration

medium composed of BCDAT medium supplemented with 0.33 M mannitol, followed by 1 week of incubation before selection.

To select clones overexpressing of *SEP1*, the transformed protoplasts were planted on selective medium containing hygromycin, and knockout lines were selected on G418. Selection for hygromycin resistance was repeated twice, and after 1 week, the clones of survivors were placed in standard medium. The presence of the insert was determined by PCR with primer pairs p5-p4r, and HygF-HygR, and deletion was detected by sequencing the fragment with primers seqF and seqR (Supplemental Figure S8). Independent knockout and overexpression mutant lines showed similar phenotypes.

**DATA ACCESS**

All raw mass spectrometry data from this study have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino et al. 2016) partner repository with the dataset identifiers PXD005223, PXD007922, PXD007923, PXD007973.

**ACKNOWLEDGEMENTS**

This work was supported by the Russian Science Foundation (project No.17-14-01189). Some of mass spectrometric measurements were performed using the equipment of the “Human Proteome” Core Facility of the Orekhovich Institute of Biomedical Chemistry (Russia) which is supported by the Ministry of Education and Science of the Russian Federation.

**Authors’ contributions**

IF and IK conceived and designed experiments. AK performed moss transformation experiments. IF, RK, VL, DK, EG, VZ, IB and AM performed the proteomics analyses. IF, IK and GA performed the statistical and bioinformatics analyses. IF, IK, VI and VG wrote the manuscript with input from all authors. IF supervised the project. All authors read and approved the final manuscript.

**DISCLOSURE DECLARATION**

842 The authors declare that they have no significant competing financial, professional, or personal
843 interests that might have influenced the performance or presentation of the work described in this
844 manuscript.
845

846 **REFERENCES**


847 Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open
848 reading frames. *Nat Rev Genet* **15**(3): 193-204.
849 Aslanidis C, de Jong PJ. 1990. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic acids*
850 *research* **18**(20): 6069-6074.
851 Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MA, Brocard M, Couso JP. 2014. Extensive
852 translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**: e03528.
853 Bao Z, Clancy MA, Carvalho RF, Elliott K, Folta KM. 2017. Identification of Novel Growth Regulators in
854 Plant Populations Expressing Random Peptides. *Plant Physiol* **175**(2): 619-627.
855 Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT,
856 Rajewsky N, Walther TC et al. 2014. Identification of small ORFs in vertebrates using
857 ribosome footprinting and evolutionary conservation. *The EMBO journal* **33**(9): 981-993.
858 Blanvillain R, Young B, Cai YM, Hecht V, Varoquaux F, Delorme V, Lancelin JM, Delseny M, Gallois P.
859 2011. The Arabidopsis peptide kiss of death is an inducer of programmed cell death. *The*
860 *EMBO journal* **30**(6): 1173-1183.
861 Blumenthal T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *BioEssays : news*
862 *and reviews in molecular, cellular and developmental biology* **20**(6): 480-487.
863 Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo
864 CA, Barbette J, Santhanam B et al. 2012. Proto-genes and de novo gene birth. *Nature*
865 **487**(7407): 370-374.
866 Cherry JL. 2010. Expression level, evolutionary rate, and the cost of expression. *Genome biology and*
867 *evolution* **2**: 757-769.
868 Chilley PM, Casson SA, Tarkowski P, Hawkins N, Wang KL, Hussey PJ, Beale M, Ecker JR, Sandberg GK,
869 Lindsey K. 2006. The POLARIS peptide of Arabidopsis regulates auxin transport and root
870 growth via effects on ethylene signaling. *Plant Cell* **18**(11): 3058-3072.
871 Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
872 Wilczynski B et al. 2009. Biopython: freely available Python tools for computational
873 molecular biology and bioinformatics. *Bioinformatics* **25**(11): 1422-1423.
874 Collonnier C, Epert A, Mara K, Maclot F, Guyon-Debast A, Charlot F, White C, Schaefer DG, Nogue F.
875 2017. CRISPR-Cas9-mediated efficient directed mutagenesis and RAD51-dependent and
876 RAD51-independent gene targeting in the moss Physcomitrella patens. *Plant Biotechnol J*
877 **15**(1): 122-131.
878 Couso JP. 2015. Finding smORFs: getting closer. *Genome Biol* **16**.
879 Couso JP, Patraquim P. 2017. Classification and function of small open reading frames. *Nature reviews*
880 *Molecular cell biology*.
881 Cove DJ, Perroud PF, Charron AJ, McDaniel SF, Khandelwal A, Quatrano RS. 2009. Isolation of DNA,
882 RNA, and protein from the moss Physcomitrella patens gametophytes. *Cold Spring Harbor*
883 *protocols* **2009**(2): pdb prot5146.
884 Crappe J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, Menschaert G. 2013.
885 Combining in silico prediction and ribosome profiling in a genome-wide search for novel
886 putatively coding sORFs. *Bmc Genomics* **14**.
887 D'Lima NG, Ma J, Winkler L, Chu Q, Loh KH, Corpuz EO, Budnik BA, Lykke-Andersen J, Saghatelian A,
888 Slavoff SA. 2017. A human microprotein that interacts with the mRNA decapping complex.
889 *Nat Chem Biol* **13**(2): 174-180.
890 De Coninck B, Carron D, Tavormina P, Willem L, Craik DJ, Vos C, Thevissen K, Mathys J, Cammue BP.
891 2013. Mining the genome of Arabidopsis thaliana as a basis for the identification of novel
892 bioactive peptides involved in oxidative stress tolerance. *J Exp Bot* **64**(17): 5297-5307.

893 De Rybel B, van den Berg W, Lokerse A, Liao CY, van Mourik H, Moller B, Peris CL, Weijers D. 2011. A
894 versatile set of ligation-independent cloning vectors for functional studies in plants. *Plant*
895 *Physiol* **156**(3): 1292-1299.
896 Djordjevic MA, Mohd-Radzman NA, Imin N. 2015. Small-peptide signals that control root nodule
897 number, development, and symbiosis. *J Exp Bot* **66**(17): 5171-5181.
898 Eguen T, Straub D, Graeff M, Wenkel S. 2015. MicroProteins: small size-big impact. *Trends Plant Sci*
899 **20**(8): 477-482.
900 Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution.
901 *Genome Res* **24**(6): 885-895.
902 Fesenko I, Khazigaleeva R, Kirov I, Kniazev A, Glushenko O, Babalyan K, Arapidi G, Shashkova T,
903 Butenko I, Zgoda V et al. 2017. Alternative splicing shapes transcriptome but not proteome
904 diversity in Physcomitrella patens. *Scientific reports* **7**(1): 2698.
905 Fesenko I, Seredina A, Arapidi G, Ptushenko V, Urban A, Butenko I, Kovalchuk S, Babalyan K, Knyazev
906 A, Khazigaleeva R et al. 2016. The Physcomitrella patens Chloroplast Proteome Changes in
907 Response to Protoplastation. *Front Plant Sci* **7**: 1661.
908 Fesenko IA, Arapidi GP, Skripnikov AY, Alexeev DG, Kostryukova ES, Manolov AI, Altukhov IA,
909 Khazigaleeva RA, Seredina AV, Kovalchuk SI et al. 2015. Specific pools of endogenous
910 peptides are present in gametophore, protonema, and protoplast cells of the moss
911 Physcomitrella patens. *Bmc Plant Biol* **15**: 87.
912 Ge Y, Porse BT. 2014. The functional consequences of intron retention: alternative splicing coupled to
913 NMD as a regulator of gene expression. *BioEssays : news and reviews in molecular, cellular and*
914 *developmental biology* **36**(3): 236-243.
915 Giannakakis A, Zhang J, Jenjaroenpun P, Nama S, Zainolabidin N, Aau MY, Yarmishyn AA, Vaz C,
916 Ivshina AV, Grinchuk OV et al. 2015. Contrasting expression patterns of coding and
917 noncoding parts of the human genome upon oxidative stress. *Scientific reports* **5**: 9737.
918 Graeff M, Straub D, Eguen T, Dolde U, Rodrigues V, Brandt R, Wenkel S. 2016. MicroProtein-Mediated
919 Recruitment of CONSTANS into a TOPLESS Trimeric Complex Represses Flowering in
920 Arabidopsis. *PLoS genetics* **12**(3): e1005959.
921 Guillen G, Diaz-Camino C, Loyola-Torres CA, Aparicio-Fabre R, Hernandez-Lopez A, Diaz-Sanchez M,
922 Sanchez F. 2013. Detailed analysis of putative genes encoding small proteins in legume
923 genomes. *Front Plant Sci* **4**.
924 Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome Profiling Provides
925 Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154**(1): 240-251.
926 Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, Shiu SH. 2010. sORF finder: a program
927 package to identify small open reading frames with high coding potential. *Bioinformatics*
928 **26**(3): 399-400.
929 Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi
930 C, Iida K, Tanaka M et al. 2013. Small open reading frames associated with morphogenesis are
931 hidden in plant genomes. *P Natl Acad Sci USA* **110**(6): 2395-2400.
932 Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH. 2007. A large number of novel coding small open
933 reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed
934 and/or under purifying selection. *Genome Res* **17**(5): 632-640.
935 Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BP, van Noort V. 2017. ARA-PEPs: a
936 repository of putative sORF-encoded peptides in Arabidopsis thaliana. *Bmc Bioinformatics*
937 **18**(1): 37.
938 Hellens RP, Brown CM, Chisnal MAW, Waterhouse PM, Macknight RC. 2016. The Emerging World of
939 Small ORFs. *Trends Plant Sci* **21**(4): 317-328.
940 Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* **411**(6841): 1046-
941 1049.
942 Housman G, Ulitsky I. 2016. Methods for distinguishing between protein-coding and long noncoding
943 RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochimica et*
944 *biophysica acta* **1859**(1): 31-40.
945 Huang JZ, Chen M, Chen, Gao XC, Zhu S, Huang H, Hu M, Zhu H, Yan GR. 2017. A Peptide Encoded by a
946 Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol Cell* **68**(1): 171-184 e176.
947 Iacono M, Mignone F, Pesole G. 2005. uAUG and uORFs in human and rodent 5'untranslated mRNAs.
948 *Gene* **349**: 97-105.

949 Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome Profiling of Mouse Embryonic Stem Cells
950 Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147**(4): 789-802.
951 Johnstone TG, Bazzini AA, Giraldez AJ. 2016. Upstream ORFs are prevalent translational repressors in
952 vertebrates. *The EMBO journal* **35**(7): 706-723.
953 Julenius K, Pedersen AG. 2006. Protein evolution is faster outside the cell. *Molecular biology and*
954 *evolution* **23**(11): 2039-2048.
955 Karousis ED, Nasif S, Muhlemann O. 2016. Nonsense-mediated mRNA decay: novel mechanistic
956 insights and biological impact. *Wiley interdisciplinary reviews RNA* **7**(5): 661-682.
957 Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD et
958 al. 2006. Functional genomics of genes with small open reading frames (sORFs) in S-
959 cerevisiae. *Genome Res* **16**(3): 365-373.
960 Kelley LA, Sternberg MJ. 2015. Partial protein domains: evolutionary insights and bioinformatics
961 challenges. *Genome Biol* **16**: 100.
962 Kim TS, Liu CL, Yassour M, Holik J, Friedman N, Buratowski S, Rando OJ. 2010. RNA polymerase
963 mapping during stress responses reveals widespread nonproductive transcription in yeast.
964 *Genome Biol* **11**(7): R75.
965 Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y.
966 2010. Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila
967 Embryogenesis. *Science* **329**(5989): 336-339.
968 Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional
969 small open reading frames in Drosophila. *Genome Biol* **12**(11).
970 Laing WA, Martinez-Sanchez M, Wright MA, Bulley SM, Brewster D, Dare AP, Rassam M, Wang D,
971 Storey R, Macknight RC et al. 2015. An Upstream Open Reading Frame Is Essential for
972 Feedback Regulation of Ascorbate Biosynthesis in Arabidopsis. *Plant Cell* **27**(3): 772-786.
973 Lauressergues D, Couzigou JM, Clemente HS, Martinez Y, Dunand C, Becard G, Combier JP. 2015.
974 Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**(7545): 90-93.
975 Lease KA, Walker JC. 2006. The Arabidopsis unannotated secreted peptide database, a resource for
976 plant peptidomics. *Plant Physiol* **142**(3): 831-838.
977 Loose CR, Langer RS, Stephanopoulos GN. 2007. Optimization of protein fusion partner length for
978 maximizing in vitro translation of peptides. *Biotechnol Prog* **23**(2): 444-451.
979 Lv Q, Lan Y, Shi Y, Wang H, Pan X, Li P, Shi T. 2017. AtPID: a genome-scale resource for genotype-
980 phenotype associations in Arabidopsis. *Nucleic acids research* **45**(D1): D1060-D1063.
981 Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, Yates JR, 3rd, Saghatelian A. 2016.
982 Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides.
983 *Analytical chemistry* **88**(7): 3967-3975.
984 Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S,
985 Selbach M et al. 2015. Extensive identification and analysis of conserved small ORFs in
986 animals. *Genome Biol* **16**.
987 Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, Couso JP. 2013. Conserved
988 regulation of cardiac calcium uptake by peptides encoded in small open reading frames.
989 *Science* **341**(6150): 1116-1120.
990 Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama
991 KI, Clohessy JG, Pandolfi PP. 2017. mTORC1 and muscle regeneration are regulated by the
992 LINC00961-encoded SPAR polypeptide. *Nature* **541**(7636): 228-232.
993 Mazin PV, Fisunov GY, Gorbachev AY, Kapitskaya KY, Altukhov IA, Semashko TA, Alexeev DG,
994 Govorun VM. 2014. Transcriptome analysis reveals novel regulatory mechanisms in a
995 genome-reduced bacterium. *Nucleic acids research* **42**(21): 13254-13268.
996 McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-
997 coding genes in eukaryotic evolutionary innovation. *Philosophical transactions of the Royal*
998 *Society of London Series B, Biological sciences* **370**(1678): 20140332.
999 Montoya-Burgos JI. 2011. Patterns of positive selection and neutral evolution in the protein-coding
1000 genes of Tetraodon and Takifugu. *Plos One* **6**(9): e24800.
1001 Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene
1002 Birth in Genome Evolution. *Molecular biology and evolution* **33**(5): 1245-1256.

1003 Narita NN, Moore S, Horiguchi G, Kubo M, Demura T, Fukuda H, Goodrich J, Tsukaya H. 2004.
1004 Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and
1005 alters leaf shape in Arabidopsis thaliana. *Plant J* **38**(4): 699-713.
1006 Neafsey DE, Galagan JE. 2007. Dual modes of natural selection on upstream open reading frames.
1007 *Molecular biology and evolution* **24**(8): 1744-1751.
1008 Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR,
1009 Chen X, Kavalali ET et al. 2016. A peptide encoded by a transcript annotated as long
1010 noncoding RNA enhances SERCA activity in muscle. *Science* **351**(6270): 271-275.
1011 Nishiyama T, Hiwatashi Y, Sakakibara I, Kato M, Hasebe M. 2000. Tagged mutagenesis and gene-trap
1012 in the moss, Physcomitrella patens by shuttle mutagenesis. *DNA research : an international*
1013 *journal for rapid publication of reports on genes and genomes* **7**(1): 9-17.
1014 Paytuvi Gallart A, Hermoso Pulido A, Anzar Martinez de Lagran I, Sanseverino W, Aiese Cigliano R.
1015 2016. GREENC: a Wiki-based database of plant lncRNAs. *Nucleic acids research* **44**(D1):
1016 D1161-1166.
1017 Pi H, Lee LW, Lo SJ. 2009. New insights into polycistronic transcripts in eukaryotes. *Chang Gung*
1018 *medical journal* **32**(5): 494-498.
1019 Prabakaran S, Hemberg M, Chauhan R, Winter D, Tweedie-Cullen RY, Dittrich C, Hong E,
1020 Gunawardena J, Steen H, Kreiman G et al. 2014. Quantitative profiling of peptides from RNAs
1021 classified as noncoding. *Nature communications* **5**: 5429.
1022 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
1023 *Bioinformatics* **26**(6): 841-842.
1024 Rasheed S, Bashir K, Nakaminami K, Hanada K, Matsui A, Seki M. 2016. Drought stress differentially
1025 regulates the expression of small open reading frames (sORFs) in Arabidopsis roots and
1026 shoots. *Plant signaling & behavior* **11**(8): e1215792.
1027 Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R. 2007. An ancient genome
1028 duplication contributed to the abundance of metabolic genes in the moss Physcomitrella
1029 patens. *BMC evolutionary biology* **7**: 130.
1030 Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist
1031 EA, Kamisugi Y et al. 2008. The Physcomitrella genome reveals evolutionary insights into the
1032 conquest of land by plants. *Science* **319**(5859): 64-69.
1033 Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M. 2002. Soybean ENOD40 encodes two peptides
1034 that bind to sucrose synthase. *Proc Natl Acad Sci U S A* **99**(4): 1915-1920.
1035 Saric T, Graef CI, Goldberg AL. 2004. Pathway for degradation of peptides generated by proteasomes:
1036 a key role for thimet oligopeptidase and other metallopeptidases. *The Journal of biological*
1037 *chemistry* **279**(45): 46723-46732.
1038 Seo PJ, Hong SY, Kim SG, Park CM. 2011. Competitive inhibition of transcription factors by small
1039 interfering peptides. *Trends Plant Sci* **16**(10): 541-549.
1040 Seo PJ, Park MJ, Park CM. 2013. Alternative splicing of transcription factors in plant responses to low
1041 temperature stress: mechanisms and functions. *Planta* **237**(6): 1415-1424.
1042 Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL,
1043 Saghatelian A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in
1044 human cells. *Nat Chem Biol* **9**(1): 59-+.
1045 Staudt AC, Wenkel S. 2011. Regulation of protein function by 'microProteins'. *EMBO reports* **12**(1):
1046 35-42.
1047 Straub D, Wenkel S. 2017. Cross-Species Genome-Wide Identification of Evolutionary Conserved
1048 MicroProteins. *Genome biology and evolution* **9**(3): 777-789.
1049 Supek F, Bosnjak M, Skunca N, Smuc T. 2011. REVIGO summarizes and visualizes long lists of gene
1050 ontology terms. *Plos One* **6**(7): e21800.
1051 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments
1052 into the corresponding codon alignments. *Nucleic acids research* **34**(Web Server issue):
1053 W609-612.
1054 Szczesniak MW, Rosikiewicz W, Makalowska I. 2016. CANTATAdb: A Collection of Plant Long Non-
1055 Coding RNAs. *Plant Cell Physiol* **57**(1): e8.
1056 Tautz D. 2009. Polycistronic peptide coding genes in eukaryotes--how widespread are they? *Briefings*
1057 *in functional genomics & proteomics* **8**(1): 68-74.

1058 Tavormina P, De Coninck B, Nikonorova N, De Smet I, Cammue BP. 2015. The Plant Peptidome: An
1059 Expanding Repertoire of Structural Features and Biological Functions. *Plant Cell* **27**(8): 2095-
1060 2118.
1061 Tyanova S, Temu T, Cox J. 2016. The MaxQuant computational platform for mass spectrometry-based
1062 shotgun proteomics. *Nat Protoc* **11**(12): 2301-2319.
1063 Vanderperre B, Lucier JF, Bissonnette C, Motard J, Tremblay G, Vanderperre S, Wisztorski M, Salzet M,
1064 Boisvert FM, Roucou X. 2013. Direct Detection of Alternative Open Reading Frames
1065 Translation Products in Human Significantly Expands the Proteome. *Plos One* **8**(8).
1066 Verheggen K, Volders PJ, Mestdagh P, Menschaert G, Van Damme P, Gevaert K, Martens L,
1067 Vandesompele J. 2017. Noncoding after All: Biases in Proteomics Data Do Not Explain
1068 Observed Absence of lncRNA Translation Products. *J Proteome Res* **16**(7): 2508-2515.
1069 Vizcaino JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F,
1070 Ternent T et al. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic acids*
1071 *research* **44**(22): 11033.
1072 Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. 2016. The ribosome-engaged landscape of alternative
1073 splicing. *Nature structural & molecular biology* **23**(12): 1117-1123.
1074 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*
1075 **24**(8): 1586-1591.
1076 Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: a tool to
1077 design target-specific primers for polymerase chain reaction. *Bmc Bioinformatics* **13**: 134.
1078 Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific
1079 genes. *Molecular biology and evolution* **21**(2): 236-239.
1080