Abundantly expressed class of non-coding RNAs conserved through the multicellular evolution of dictyostelid social amoebae

Jonas Kjellin[1], Lotta Avesson[2,3], Johan Reimegård[4], Zhen Liao[1,5], Ludwig Eichinger[6], Angelika Noegel[6], Gernot Glöckner[6], Pauline Schaap[7], and Fredrik Söderbom[1]

[1]Department of Cell and Molecular Biology, Uppsala University, Box 596 Uppsala, S-75124 Sweden, [2]Department of Molecular Biology, Biomedical Center, Swedish University of Agricultural Sciences, Box 590, S-75124 Uppsala, Sweden, [4]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Box 596 Uppsala, S-75124 Sweden, [6]Centre for Biochemistry, Institute of Biochemistry I, Medical Faculty, University of Cologne, Cologne, Germany and [7]College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom.

[3]Present address: Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen

[5]Present address: Department of Plant Biology, Swedish University of Agricultural Sciences, Box 7080 Uppsala, S-750 07 Sweden

Corresponding author. fredrik.soderbom@icm.uu.se. Phone: 46 184714901

Abstract

Background: Aggregative multicellularity has evolved multiple times in diverse groups of eukaryotes. One of the most well-studied examples is the development of dictyostelid social amoebae, e.g. *Dictyostelium discoideum*. However, it is still poorly understood why multicellularity emerged in these amoebae while the great majority of other members of Amoebozoa are unicellular. Previously a novel type of non-coding RNA, Class I RNAs, was identified in *D. discoideum* and demonstrated to be important for normal multicellular development. In this study we investigated Class I RNA evolution and its connection to multicellular development.

Results: New Class I RNA genes were identified by constructing a co-variance model combined with a scoring system based on conserved up-stream sequences. Multiple genes were predicted in representatives of each major group of Dictyostelia and expression analysis validated that our search approach can identify expressed Class I RNA genes with high accuracy and sensitivity. Further studies showed that Class I RNAs are ubiquitous in Dictyostelia and share several highly conserved structure and sequence motifs. Class I RNA genes appear to be unique to dictyostelid social amoebae since they could not be identified in searches in outgroup genomes, including the closest known relatives to Dictyostelia.

Conclusion: Our results show that Class I RNA is an ancient abundant class of ncRNAs, likely to have been present in the last common ancestor of Dictyostelia dating back at least 600 million years. Taken together, our current knowledge of Class I RNAs suggests that they may have been involved in evolution of multicellularity in Dictyostelia.

Keywords (3-10): social amoeba, *Dictyostelium discoideum*, non-coding RNA, evolution, Class I RNA, dictyostelids, multicellularity, Dictyostelia

Background

The role of RNA goes far beyond it being an intermediate transmitter of information between DNA and protein, in the form of messenger (m)RNAs. This has been appreciated for a long time for some non-coding RNAs (ncRNAs), such as transfer (t)RNAs, ribosomal (r)RNAs, small nuclear (sn)RNAs, and small nucleolar (sno)RNAs. Today, we know that ncRNAs are involved in regulating most cellular processes and the advent of high-throughput sequencing technologies have facilitated the identification of numerous different classes of ncRNAs [1]. These regulatory RNAs vary greatly in size from 21-24 nucleotides (nt), e.g. micro (mi)RNAs and small interfering (si)RNAs, to several thousands of nucleotides, such as long non-coding (lnc)RNAs. Several classes of ncRNAs are ubiquitously present in all domains of life while others are specific to certain evolutionary linages, contributing to their specific characteristics. This can be exemplified by ncRNAs in Metazoa, where an increase in number of ncRNAs, e.g. miRNAs, is associated with increased organismal complexity and is believed to be essential for the evolution of metazoan multicellularity [2]. Multicellularity in plants and animals is achieved by clonal division and development originating from a single cell. This is in contrast to aggregative multicellularity, were cells stream together to form multicellular structures upon specific environmental changes. Aggregative multicellularity has evolved independently multiple times and is found both among eukaryotes and prokaryotes [3–10]. The complexity of the aggregative multicellular life stages varies for different organisms, but they all share the transition from unicellularity to coordinated development upon environmental stress, e.g. starvation, which eventually leads to formation of fruiting bodies containing cysts or spores [3]. Probably the most well-studied aggregative multicellularity is the development of the social amoeba *Dictyostelium discoideum* belonging to the group Dictyostelia within the supergroup Amoebozoa. Dictyostelia is a monophyletic group estimated to date back at least 600 million years [11], which is similar to the age of Metazoa [12]. Dictyostelia is currently divided into four major groups (Group 1-4) where all members share the ability of transition from uni- to multicellularity upon starvation [13]. However, the complexity

71 of the development and the morphology of the fruiting bodies varies between different dictyostelids,
72 where the highest level of multicellular complexity is found among group 4 species, which includes *D.*
73 *discoideum* [14, 15]. Recently a new taxonomy was proposed for many dictyostelids [16]. As this new
74 taxonomy has not yet been fully adopted by the research community, we choose to use the previous
75 designations throughout this study (old and new names, including NCBI accession numbers, are
76 summarized in Additional file 1).
77
78 Well-annotated genome sequences are available for representative species of all four major groups of
79 Dictyostelia [11, 17–19] and multiple draft genome sequences are available for additional dictyostelids.
80 This has allowed for comparative genomics, which has provided information about the protein coding
81 genes that are important for the diversification of Dictyostelia from other amoebozoans. Comparison
82 between genomes has also given insight into the genes required for the evolution of the distinct
83 morphological characteristics, which define each group [11, 17, 18, 20, 21]. However, evolution of
84 complex traits such as multicellularity in Dictyostelia as well as other eukaryotic groups, cannot solely be
85 explained by the appearance of novel genes but also relies on an increased ability to regulate pre-existing
86 genes and their products so that they can function in novel genetic networks [20, 22]. This is also
87 supported by the major transcriptional reprogramming during multicellular development in *D. discoideum*
88 [23].
89
90 *D. discoideum* harbors several classes of developmentally regulated ncRNAs with regulatory potential, e.g.
91 microRNAs [24–27], long non-coding RNAs [28] and long antisense RNAs [28, 29]. In addition, a large part
92 of the ncRNA repertoire of *D. discoideum* is constituted by Class I RNAs, originally identified in full length
93 cDNA libraries [30]. So far, Class I RNAs have only been validated in *D. discoideum* [30, 31], but they have
94 also been computationally predicted in *Dictyostelium purpureum* [18]. Both species belong to the same

95 evolutionary group of Dictyostelia, i.e. group 4 [13, 16]. In *D. discoideum,* Class I RNAs are 42-65 nt long
96 and are expressed at high levels from a large number of genes. Members of Class I RNAs are characterized
97 by a short stem structure, connecting the 5' and 3' ends, and a conserved 11 nt sequence motif adjacent
98 to the 5´part of the stem. The remainder of the RNA is variable both in sequence and structure. Class I
99 RNAs mainly localize to the cytoplasm [30] where one of the RNAs has been shown to associate with four
100 different proteins of which at least one, the RNA recognition motifs (RRM) containing protein CIBP,
101 directly binds to the Class I RNA [31]. Furthermore, the Class I RNAs appear to be involved in regulating
102 multicellular development. This is based on the observations that Class I RNAs are developmentally
103 regulated and that cells where a single Class I RNA gene has been knocked out show aberrant early
104 development [30, 31].
105
106 In this study we investigated the prevalence of Class I RNAs within Dictyostelia as well as in other
107 organisms with the aim to understand if Class I RNAs are restricted to dictyostelids and perhaps required
108 for their aggregative multicellularity. Based on the known Class I RNAs from *D. discoideum*, we know that
109 the major part of the RNA is variable and hence sequence-based searches alone, such as BLAST, would
110 not reliably identify new genes. This was solved by constructing a Class I classifier based on a co-variance
111 model, which includes both sequence and structure information, combined with a scoring system for
112 conserved up-stream sequence motifs, e.g. promoter motifs. Using this search approach, we identified
113 approximately 300 Class I RNA genes predicted to be expressed in the genomes of 16 different
114 dictyostelids. In addition, we validate the expression of ~100 Class I RNAs from six different species, using
115 both northern blot and RNA-seq, which supports the high accuracy and sensitivity of our search approach
116 to identify expressed Class I RNAs. Comparative studies of identified Class I RNA loci show several well
117 conserved features, like stem forming properties connecting the 5' and 3' ends, as well as preserved
118 sequence motifs. Importantly, Class I RNAs appear to be specific to Dictyostelia as no Class I RNA genes

119 were identified in genomes of organisms outside this group of social amoebae, including their closest
120 known relatives or organisms exhibiting different kinds of multicellularity. Taken together, this unique
121 class of ncRNAs constitute a very large number of conserved highly expressed genes involved in the
122 evolution of Dictyostelia multicellular development.
123

124 **Results**

125 **Co-variance model identifies Class I RNA genes in evolutionary distinct groups of Dictyostelia social**
126 **amoebae.**

127 We have previously identified and characterized Class I RNAs from the social amoeba *D. discoideum*.
128 Members of Class I RNAs are characterized by a short stem-structure connecting the 5' and 3' ends and a
129 conserved 11 nt sequence motif adjacent to the 5´part of the stem. The remainder of the Class I RNAs are
130 highly variable both in sequence and structure (Fig. 1a). This class of ncRNA have so far only been validated
131 in *D. discoideum* where it is associated with development [30, 31], but has also been predicted in *D.*
132 *purpureum* [18].

133

134 The presence of Class I RNA genes in two different dictyostelids and the fact that at least one Class I RNA
135 member is involved in controlling early multicellular development, led us to hypothesize that this class of
136 ncRNA may be a general effector for early development in all members of dictyostelid social amoebae.
137 Hence, it may have been present in the last common ancestor of Dictyostelia, dating back approximately
138 600 million years in evolution. In order to investigate this, we used the complete and well-annotated
139 genome sequences for representatives of each major group of Dictyostelia, i.e. *D. discoideum* [17], *D.*
140 *purpureum* [18], *Dictyostelium lacteum* [20], *Polysphondylium pallidum* [11], *Acytostelium subglobosum*
141 [19] and *Dictyostelium fasciculatum* [11] (Fig. 1b). Class I RNAs cannot be reliably detected by sequence
142 searches alone due to the high sequence variability. Therefore, we constructed a co-variance model (CM)
143 with Infernal [32] where both sequence and secondary structure information of 34 *D. discoideum* Class I
144 RNAs were taken into account (see Materials and methods for details). The initial CM search of the six
145 Dictyostelia genomes, followed by manual inspection of the results, indicated the presence of Class I RNAs
146 in all major groups of Dictyostelia (Additional file 2: Fig. S1). These candidates scored ≥ 25 in the CM search
147 and had the potential to form a short stem similar to the *D. discoideum* Class I RNAs. In order to improve

148 the CM, these candidates (CM score ≥ 25 and potential to form stem) were added to the CM followed
149 by new genome searches. This process was repeated until no new candidates fulfilling the criteria were
150 identified after which a final search with increased sensitivity was performed (see Materials and methods).
151 In total, 126 loci distributed over all major groups of Dictyostelia were identified including 36 of the 40
152 published *D. discoideum* Class I RNAs [30, 31] and all the 26 previously predicted *D. purpureum* Class I loci
153 [18] (Additional file 2: Fig. S1).
154
155 **Refining the search for Class I RNA genes using conserved promoter elements**
156 Many *D. discoideum* ncRNA genes have an upstream putative promoter element, DUSE (*Dictyostelium*
157 upstream sequence element), which in most cases is situated ~60 nt from the transcriptional start site
158 (TSS) [33]. However, for *D. discoideum* Class I RNAs, DUSE is often found further upstream. In these cases,
159 a TGTG-box (AAATGTG) is located ~60 nt downstream of DUSE while the distance from the start of the
160 mature RNA varies. Whether the TGTG-box is an additional promotor element or the TSS of a precursor
161 transcript is currently not known. DUSE appears to be conserved within group 4 of Dictyostelia, as it was
162 also identified ~60 nt in front of the predicted *D. purpureum* Class I RNAs [18]. In order to investigate the
163 presence of conserved upstream motifs in the rest of Dictyostelia, we searched for enriched motifs in the
164 150 nt upstream sequence of all the 126 Class I RNA gene candidates identified in the CM search.
165 Intriguingly, DUSE like motifs could be identified ~60 nt upstream of the predicted start for the majority
166 of the Class I RNA gene candidates (73 of 126) in all organisms. In contrast, the TGTG-box was only found
167 in a subset (21 of 126) of the upstream sequences of which all belonged to *D. discoideum* Class I RNA
168 genes (Fig. 1c). As both sequence and distance of DUSE appeared to be conserved in all major groups of
169 Dictyostelia, we used this information to create a scoring system anticipating accurate prediction of
170 expressed Class I RNA genes. (Fig. 1d). First the score produced by the CM search (Infernal) was used and
171 all candidates with a score ≥ 15 were included in order to capture more divergent Class I RNA genes. Next

172 we scored the presence and location of DUSE and TGTG-box 150 nt upstream of the candidates identified
173 in the CM search based on the motif identification program FIMO [34]. Lack of DUSE and/or non-canonical
174 distance from predicted TSS or TGTG-box was penalized with negative scores. Taken together, a total
175 score of 32 could be achieved if a high-scoring DUSE was identified at the predicted distance upstream of
176 a candidate Class I RNA gene with the lowest allowed Infernal score (≥ 15). Based on this, all candidates
177 scoring 32 or higher were classified as Class I RNA loci. Using this approach, we predicted 18-39 Class I
178 RNAs (146 in total) for each of the six dictyostelids investigated (see below).
179
180 **Class I RNAs of predicted sizes are expressed at high levels in all four groups of Dictyostelia**
181 Based on the Class I classifier, we predicted Class I RNA genes in all dictyostelids included in the CM build.
182 But are all these genes really expressed and how accurate are the size predictions? From our previous
183 studies, we know that Class I RNAs in *D. discoideum* are expressed at high levels at vegetative growth and
184 are readily detected by northern blot [30, 31]. Hence, we used the same approach to validate a subset of
185 randomly chosen candidates that made the total score threshold in *D. purpureum* (Group 4)*, D. lacteum*
186 (Group 3)*, P. pallidum* (Group 2A)*, A. subglobosum* (Group 2B) and *D. fasciculatum* (Group 1). RNA was
187 prepared from vegetative growing amoebae and specific Class I RNAs were analyzed by northern blot (Fig.
188 2a). For *D. purpureum*, we probed for DpuR-7, predicted to be 54 nt long, resulting in a strong signal. In
189 addition, we designed two probes predicted to recognize six different 85 nt long RNAs (DpuR-X) and the
190 majority (24/30) of Class I RNAs (DpuR-Y), respectively. As expected, probing for DpuR-X resulted in one
191 band on the northern blot. For DpuR-Y, we expected signals from several Class I RNAs to overlap due to
192 similar/identical sizes but we could still detect several distinct bands within the expected size range. In *D.*
193 *fasciculatum*, we probed for one Class I RNA predicted to be 62 nt long (DfaR-4) while at least two
194 candidates were probed for in the other organisms, i.e. *D: lacteum:* DlaR-1 (61 nt) and DlaR-5 (54 nt), *P.*
195 *pallidum*: PpaR-1/2 (59/62 nt) and PpaR-9 (58 nt), and *A. subglobosum*: AsuR-13/14 (61 nt) and AsuR-10

196 (82 nt). Distinct bands were detected for each Class I RNA candidate and the sizes matched the predictions
197 well although the northern results often indicated that the RNAs were a few nt longer than predicted (Fig.
198 2a and see below). The larger (but much weaker) signals observed for AsuR-13/14 and PpaR-9 are likely
199 cross hybridizations to longer Class I RNAs. Taken together, the results confirm that Class I RNAs are
200 conserved and expressed in all four groups of Dictyostelia social amoebae.
201
202 **Classifier accurately predicts expressed Class I RNAs in all major groups of Dictyostelia**
203 Next, we performed RNA-seq on *D. discoideum, D. lacteum, P. pallidum* and *D. fasciculatum* representing
204 each major group of Dictyostelia. RNA was prepared from growing cells as well as two multicellular life-
205 stages, i.e. mound and slug/finger stages, to increase our chances to also detect Class I RNAs that are only
206 expressed at specific life stages. Expression was evaluated based on the read count and coverage over all
207 loci identified in the CM search (Infernal score ≥ 15) as exemplified in Additional file 2: Fig. S2. Strikingly,
208 expression both during vegetative growth and development could be confirmed for almost all Class I RNA
209 candidates identified by the classifier (total score ≥ 32) (Fig. 2b, Additional file 3). Next, we calculated
210 receiver operator characteristics (ROC) curves in order to evaluate the classifier performance and
211 investigate if it improves Class I RNA identification compared to CM search alone. ROC curves were
212 generated for *D. discoideum, D. lacteum, P. pallidum* and *D. fasciculatum* individually (Additional file 2:
213 Fig. S3) as well as for the pooled data (Fig. 2c) based on the RNA-seq validation and either CM search score
214 (≥ 25) or classifier score (≥ 32). Evaluation of the two search approaches show an increase in both
215 sensitivity and accuracy of prediction for the classifier, i.e. when the promoter (DUSE) presence and
216 distance were included in the classification of Class I RNA gene candidates. To summarize, the classifier
217 reliably detects expressed Class I RNAs in all the tested dictyostelids with almost no false positives.
218
219

220 **Conserved features of Dictyostelia Class I RNA**

221 The wealth of newly identified Class I RNA genes in six different and evolutionary separated dictyostelids
222 allowed us to construct a general/unifying picture of Class I RNAs. This will also be of importance when
223 searching for Class I RNA genes in other species in order to track down the birth of this class of ncRNAs in
224 evolution (see below).

225

226 *Class I transcription is dependent on DUSE*

227 Both the sequence of DUSE and its upstream location is highly conserved in all of the analyzed amoebae
228 (Fig. 3a). In the group 1 and 4 dictyostelids, DUSE contains three consecutive C residues, while two
229 consecutive C's are found in the majority of the DUSE of *A. subglobosum* (Group 2B) and *D. lacteum* (Group
230 3) and in all *P. pallidum* (Group 2A). The RNA-seq data show that the promoter element is essential for
231 transcription as it is found in front of all expressed Class I RNA genes. Further strengthening its importance
232 is the observation that high scoring Class I RNAs (both considering score produced by classifier or CM
233 alone) lacking DUSE at the correct up-stream location are not expressed (Additional file 3). The TGTG-box,
234 situated 60 nt down-stream of the DUSE and in front of the predicted TSS, is only found in *D. discoideum*
235 suggesting that this is a rather late addition in the evolution of Class I RNA genes. The genes are most
236 likely transcribed by RNA polymerase III (Pol III) as the majority of the Class I RNAs from the different
237 dictyostelids exhibit a stretch of at least four consecutive T-residues downstream of the predicted end of
238 transcription, which is a common Pol III termination signal [36].

239

240 *5' and 3' ends of Class I RNAs are conserved*

241 The RNA-seq analyses showed that the *D. discoideum* Class I RNAs started at the predicted (and previously
242 defined [30]) G residue while the majority of group 1-3 Class I RNAs started one to two nucleotides
243 upstream of the conserved G (Additional file 2: Fig. S2). When we compared all loci, we noticed that the

244 two nucleotides preceding the completely conserved G residue, are highly conserved A-residues in all six
245 species investigated. The difference in 5' ends of the mature Class I RNAs indicates that either
246 transcription initiation or 5´ processing of Class I RNAs differ between group 4 species and species
247 belonging to the other evolutionary groups of Dictyostelia. Also, the very 3' end of Class I RNAs is highly
248 similar with an almost perfectly conserved CTGT sequence in the genomic loci. The coverage from the
249 RNA-seq data indicates that these four nucleotides are transcribed so that the CUGU sequence is included
250 in the mature Class I RNA (Additional file 2: Fig. S2) where the C residue always have base pairing potential
251 with the conserved 5' G residue (Fig. 3a-b). The RNA-seq coverage agrees with the slightly longer than
252 predicted lengths of Class I RNAs observed by northern blot (see above).
253
254 *Class I RNA GC-content, size, sequence motif, and stem-structure are conserved throughout Dictyostelia*
255 Comparison of all identified Class I RNAs showed that both Class I RNA length (median of ~60 nt) and GC
256 content (32-41%) are highly conserved (Additional file 2: Fig. S4a-b). The stable GC content of Class I RNAs
257 is remarkable considering the variation in overall genome GC content for these six dictyostelids (Additional
258 file 2: Fig. S4b). In spite of these conserved features and specific sequence motifs discussed above and
259 later, the overall sequence variability of Class I RNAs is extensive both within and between species. Only
260 a few examples of loci with identical sequences are found within *D. discoideum, D. purpureum* and *P.*
261 *pallidum,* respectively (Additional file 3). No Class I genes with identical sequences were found between
262 these six analyzed species.
263
264 In contrast to the overall variable sequences, the 11 nt sequence motif identified among the *D. discoideum*
265 Class I RNAs, is highly conserved both within and between all six dictyostelids (marked in grey in Fig. 3a).
266 The motif is nearly perfectly conserved within group 4, while some positions of the motif are variable in
267 group 1-3. However, T, C, C, A, and A at position 3, 6, 9, 10 and 11 (counting from the 5´most nt of the

268 motif) are almost identical between all the Class I RNAs regardless of species. Other nucleotides are well
269 conserved in most of the evolutionary groups but not all. The sequence motif does not seem to extensively
270 engage in base-pairing since computational prediction indicates that the conserved motif is less structured
271 compared to the full-length RNA in most of the organisms (Additional file 2: Fig. S4c-d). However, the first
272 5´-nucleotide of the motif is often part of the stem-structure (see below), while the base pairing potential
273 for the remainder of the sequence drops in a pattern similar for all six dictyostelids (Additional file 2: Fig.
274 S4e).
275
276 Another distinct feature common to all *D. discoideum* Class I RNAs is the short (six bp) stem structure,
277 connecting the 5' and 3' ends of the RNA (Fig. 3a). This stem is predicted to be present in all Class I RNAs
278 in all six species. However, in contrast to the conserved sequence motif, the nucleotide sequence of the
279 stem-structure has changed substantially during Dictyostelia evolution. Nevertheless, the base pairing
280 potential is retained, indicating that it is the structure rather than sequence that is crucial for function
281 (Fig. 3b). This is further supported by the high number of compensatory mutations found in the predicted
282 stem of Class I RNAs within each species (Fig. 3b). Notably, in spite of the sequence variation in the stems,
283 the 5' most G is completely conserved within all Class I RNAs from all six dictyostelids representing each
284 evolutionary group of Dictyostelia. The predicted base-paired structure of the stem and the unstructured
285 feature of the conserved sequence motif correspond well with previous *in vitro* probing results of one
286 Class I RNA, DdR21, from *D. discoideum* [31]. Taken together, Class I GC-content, length, stem structure
287 and 11 nt motif are highly conserved in all evolutionary groups of Dictyostelia, indicating that these parts
288 are essential for Class I RNA function.
289
290
291

292 **Class I RNAs are developmentally regulated and highly conserved throughout Dictyostelia**

293 We knew from our previous work that *D. discoideum* Class I RNAs are developmentally regulated and that
294 Class I knock-out cells, *DdR-21* k.o., are disturbed in early development, leading to more and smaller
295 fruiting bodies compared to wild-type cells [31]. In order to investigate if the developmental regulation is
296 conserved also in other dictyostelids, we performed principal component analysis (PCA) of Class I
297 expression in *D. discoideum, P. pallidum* and *D. fasciculatum* based on the RNA-seq data. *D. lacteum* was
298 not included in this analysis since only one replicate per timepoint was available. The PCA plots show
299 developmental regulation of Class I RNAs in all three amoebae as the different life stages are clearly
300 separated (Additional file 2: Fig. S5). Taken together, this suggests a role for Class I RNAs in regulating
301 multicellular development in Dictyostelia.

302

303 If the prediction that Class I RNAs are involved in and important for multicellular development holds true,
304 these ncRNAs should be present in all dictyostelids. To analyze this, we used the Class I classifier to
305 investigate the presence of Class I RNA genes in ten additional social amoebae genome sequences. Class
306 I RNA genes were detected in all species, 9-31 genes in each genome, of which the great majority passed
307 manual curation based on the ability to form a short stem connecting the 5' and 3' end (Fig. 4). It should
308 be noted that these are draft genome sequences of varying degree of completeness (Additional file 1).
309 Comparison of all curated loci reinforced the previously identified conserved Class I features i.e. the high
310 sequence conservation of the terminal residues and the 11 nt motif as well as the short stem where
311 structure but not sequence is preserved. In addition, presence of DUSE ~60 nt up-stream of the majority
312 of the identified Class I RNA genes strongly suggests that expressed Class I RNAs exist in all members of
313 Dictyostelia. The TGTG-box, previously only found upstream of *D. discoideum* Class I RNA loci, was
314 identified in four additional genomes all belonging to group 4 or the *P. violaceum* complex strengthening
315 the hypothesis that this motif emerged rather late in Class I evolution. In addition, both Class I RNA lengths

316 and GC content are conserved also when considering all species (Additional file 2: Fig. S6). Taken together,
317 this proves the existence and emphasize the importance of Class I RNA genes throughout the evolution of
318 Dictyostelia social amoebae. We have named all curated Class I RNA genes according to the naming
319 convention previously defined for *D. discoideum* Class I RNA genes [30] (Additional file 4).
320
321 **Genomic distribution of Class I RNA genes**
322 In *D. discoideum*, all Class I RNA genes are located in intergenic regions and frequently found in clusters
323 of two or more genes [33]. Clusters of at least two (different) Class I RNA genes are present in all analyzed
324 Dictyostelia genomes, except for *D. citrinum* (Additional file 5). The absence of Class I RNA gene clusters
325 in *D. citrinum* is likely a consequence of the quality of the genome assembly (Additional file 1), which is
326 also reflected in the low number of identified Class I RNA genes. Clusters with a higher number of Class I
327 RNA genes (three or more) are only found in five genomes, where the distribution in *P. pallidum* resembles
328 that in *D. discoideum,* i.e. many of the genes are collected in two larger clusters on the same chromosome
329 (Additional file 5). Even though many Class I RNA genes cluster together, they rarely have identical
330 sequences. Only a few species-specific identical Class I RNAs were found in *D. polycephalum, P. violaceum,*
331 *D. purpureum*, *P. pallidum* and *D. discoideum*, where in the latter two the identical genes are located in
332 clusters (Additional file 5). Are there Class I RNAs that are identical between two different species? The
333 only examples found were two Class I RNA loci shared between the group 4 species *D. discoideum* and *D.*
334 *firmibasis* (Additional file 4).
335
336 To explore the origin of Class I RNAs further, we used the most well-annotated genomes to search for
337 shared synteny by identifying orthologous genes in the 10 kb region flanking each Class I RNA locus
338 (Materials and methods). Using this approach, we did not identify any strong evidence for shared synteny
339 for Class I RNA genes between the different groups of Dictyostelia. Next we investigated if shared synteny

340 could be detected within group 4 only by performing the same search using the genomes of *D. discoideum,*
341 *D. purpureum* and *D. firmibasis*. Almost half of the Class I RNAs in *D. firmibasis* appear to share synteny
342 with *D. discoideum* Class I RNAs, supported by several protein gene orthologues (Additional file 6) while
343 no well-supported examples were found for *D. purpureum*. For the identical Class I RNA genes in *D.*
344 *discoideum* and *D. firmibasis*, shared synteny was detected for DfiR-4 and DdR-47 (Additional file 6).
345 Shared synteny between the other two identical Class I RNAs, DfiR-12 and DdR-50, could not be properly
346 assessed due to the lack of available genome sequence surrounding DfiR-12.
347

348 **Class I RNAs are unique to dictyostelid social amoebae**

349 The omnipresence of Class I RNAs within Dictyostelia, their developmental regulation as well as the
350 aberrant development of *D. discoideum* cells lacking *DdR-21* [31] led us to hypothesize that this class of
351 ncRNAs might be involved in the evolution of Dictyostelia aggregative multicellularity. In order to
352 investigate this further, we searched for Class I RNA genes in genomes of unicellular amoebae and
353 amoebae able to form unicellular fruiting bodies. Furthermore, we explored representative genomes of
354 other major eukaryotic groups, i.e. archeaplastida and ophistokonta. We also chose to include the
355 proteobacteria *Myxococcus xhantus* as these bacteria exhibit aggregative multicellularity [37], which in
356 many aspects are analogous to Dictyostelia multicellularity (Fig. 4). We searched these genomes using the
357 same successful approaches as for Dictyostelia, i.e. using the Class I classifier, based on promoter
358 characteristics combined with RNA structure and sequence, as well as CM search alone. Only a few
359 candidates were identified with the Class I classifier. This was anticipated since we did not expect the
360 DUSE sequence or its distance to the TSS to be conserved outside Dictyostelia. The Infernal search resulted
361 in a slightly higher number of Class I RNA gene candidates. However, manual inspection revealed that the
362 candidates are unlikely to represent true Class I RNA genes as they were few in numbers and did not share
363 characteristics, such as conserved 5' and 3' ends and presence of conserved sequence motif (Fig. 4,

364 Additional file 7). Taken together, no Class I RNA genes were identified outside Dictyostelia, suggesting
365 that this class of ncRNAs is unique to dictyostelid social amoebae and important for their aggregative
366 multicellularity.
367
368 **Conserved Class I RNA interacting proteins**
369 Class I RNAs are conserved throughout the evolution of Dictyostelia but does this also apply to proteins
370 associated with this class of ncRNA? We previously identified four Class I RNA interacting proteins in *D.*
371 *discoideum* by using one specific Class I RNA, DdR-21, as bait in pull-down experiments. Two of these
372 proteins, GuaB and NdkC, are involved in nucleotide metabolism while the function of DDB_G0281243 is
373 unknown. The fourth identified Class I RNA interacting protein, CIBP (also known as Rnp1A [38]), harbors
374 two RNA binding motifs (RRMs) and was demonstrated to bind directly to the Class I RNA [31]. We
375 searched for orthologues of the four proteins in the best annotated genomes of Dictyostelia (Fig. 1b,
376 Additional file 1) and could identify orthologues to GuaB, DDB_G0281243 and CIBP in all Dictyostelia
377 genomes investigated. Next, we used mRNA-seq data from *D. discoideum* (Group 4) [39] as well as *D.*
378 *lacteum* (Group 3), *P. pallidum* (Group 2) and *D. fasciculatum* (Group 1) [20] to investigate expression and
379 developmental regulation of the genes for each of the orthologues. According to the RNA-seq data, only
380 *CIBP* were expressed in all four species. Interestingly, during early development the expression pattern of
381 the gene is similar to that of Class I RNAs [31], i.e. *CIBP* appears to be down-regulated during early
382 development in all species. However, as *D. discoideum* Class I RNAs continue to decrease in expression
383 throughout development, *CIBP* expression seems to increase in the final stages of development
384 (Additional file 2: Fig. S7a). Next, we performed gene predictions in the ten additional Dictyostelia
385 genomes (Fig 4) and found *CIBP* orthologues in all species except for *D. citrinum* (Additional file 2: Fig.
386 S7b). The absence of *CIBP* in *D. citrinum* is likely due to the quality of the genome assembly (Additional
387 file 1). The majority of the orthologues are predicted to encode an approximately 300 amino acids (aa)

388 long protein with two RRM's. However, shorter orthologues were identified in *A. leptosomum* (114 aa),
389 *A. ellipticum* (61 aa) and *D. purpureum* (203 aa). Also, only one RRM could be identified in the *CIBP*
390 orthologues in *D. purpureum* and *A. ellipticum* (Additional file 2: Fig. S7b). Collectively, the majority of the
391 disctyostelids are predicted to encode/express CIBP of similar length where the N- and C- terminal
392 sequences contains RNA binding motifs, RRMs, while the central part of the protein is less conserved.
393 Despite several efforts by us and others, all attempts to generate a CIBP knock out strain in *D. discoideum*
394 have been unsuccessful [31, 37]. In addition, no CIBP mutants were generated in the recent Genome Wide
395 *Dictyostelium* Insertion (GWDI) project (www.remi-seq.org). Taken together, this indicates that the CIBP
396 protein is essential.
397
398 **Discussion**
399 The development of high-throughput sequencing techniques has led to the discovery of numerous
400 ncRNAs. In particular, it has facilitated the identification of small ncRNAs such as mi- and si- and piwi-
401 interacting (pi)RNAs sized ~21-31 nt and long ncRNAs that can consist of up to several thousand nt.
402 However, "mid-sized" ncRNA have largely been overlooked partly due to the size selection commonly
403 carried out before sequencing to enrich for small RNAs and to avoid abundant RNAs such as rRNAs and
404 tRNAs or fragment thereof. Here we used genome analyses in combination with expression validation to
405 prove the existence of Class I RNAs in 16 dictyostelid social amoeba. This indicates that Class I RNAs were
406 present in the last common ancestor of Dictyostelia, dating back at least 600 million years, and were
407 involved in the transition from unicellular to multicellular life.
408
409 Class I RNAs play an important role in *D. discoideum*, as suggested by e.g. the large number of highly
410 expressed genes and requirement for normal multicellular development [30, 31]. This class of ncRNAs was
411 initially discovered by sequencing cDNA libraries of full-length RNA sized 50-150 nt [30]. In the same study,

412 a putative promoter element, DUSE, was identified to be associated with Class I RNA genes. Later, two
413 approaches to predict Class I RNAs were published. Fragrep, a tool which predict ncRNAs based on
414 sequence motifs separated by a variable region, identified 45 Class I RNA candidate genes in *D. discoideum*
415 of which 34 had been previously experimentally validated [40]. In *D. purpureum,* another group 4
416 dictyostelid, 26 Class I genes were predicted by searching for enriched 8-mers downstream of DUSE motifs
417 in the genome sequence [18]. Hence, previous to the present study, Class I RNAs had only been identified
418 in organisms belonging to one group of Dictyostelia. We now asked if Class I RNA genes also are present
419 in other organisms, both within Dictyostelia characterized by aggregative multicellularity and outside this
420 group of social amoebae.
421
422 In order to detect Class I RNAs, we first built a covariance model (CM) which predicts candidates based on
423 both sequence and structure information of Class I RNAs. Based on the CM, we created a classifier which
424 evaluates the identified candidates based on the presence of DUSE at the correct distance from TSS or
425 from the TGTG-box (found upstream of many *D. discoideum* Class I RNA genes). Next, we confirmed
426 expression of approximately 100 Class I RNAs, predicted by the classifier, by northern blot and/or RNA-
427 seq. We were now confident that we could use the Class I classifier to accurately identify expressed Class
428 I RNA genes also in other organisms. Based on this, we show that Class I RNAs are present in all
429 dictyostelids with available genome sequences (total of 16). Having established the presence of Class I
430 RNAs in all tested disctyostelids, we turned our focus on organisms outside of Dictyostelia. Outgroups
431 were chosen to both represent organisms with a unicellular life style and those that go through different
432 kinds of multicellular development. We used both the classifier and the CM search alone to search for
433 candidates, the latter to avoid the constraint of having a DUSE element in front of the Class I RNA genes.
434 Regardless of search approach, we did not identify Class I RNAs in any organism outside Dictyostelia,
435 suggesting that these RNAs are restricted to dictyostelid social amoebae.

436 The ubiquitous presence of Class I RNA genes in Dictyostelia and absence in the closest related unicellular
437 amoebae suggest that these RNAs were present in the last common ancestor of Dictyostelia and played a
438 role in the transition from unicellular life to aggregative multicellularity. This is further supported by the
439 developmental regulation of Class I RNAs and that at least one member, DdR-21, is required for normal
440 development [30, 31]. In *D. discoideum*, transition from unicellular growth to multicellular development
441 is associated with large transcriptional reprogramming of protein coding genes [23] and different cell
442 types, i.e. prespore and prestalk cells, can be separated based on the transcriptional signatures of
443 individual cells [41]. Interestingly, the majority of the protein coding genes that are essential for
444 multicellular development in *D. discoideum* is also present in strictly unicellular amoebae [20]. Hence, it
445 is likely that the ability to regulate how genes are expressed have been key in the evolution of multicellular
446 development. Maybe Class I RNAs can rewire gene expression of genes present in unicellular organisms
447 to create new networks adapted for development. We have no evidence that Class I RNA directly interacts
448 with mRNA to regulate gene expression even though we are not excluding this mode of action. Another
449 possibility is that Class I RNA regulate development by binding to proteins that directly or indirectly control
450 development, maybe by acting as a molecular sponge, where specific proteins are sequestered by Class I
451 RNAs. This could buffer the action of these proteins. Perhaps the observed down regulation of Class I RNAs
452 during development lead to an increase in free active proteins important for multicellular development.
453 CIBP would be a candidate for this kind of regulation since CIBP directly interacts with Class I RNAs, at
454 least with the tested DdR21, in *D. discoideum* [31]. Support for CIBP as an important protein for
455 development is its presence in all dictyostelids. We are currently attempting to decipher the function of
456 Class I RNAs in *D. discoideum* by using a coupled RNA-seq and proteomics approach to investigate the
457 involvement of Class I RNA in early multicellular development. The tendency to differentially regulate
458 genes in order to create new functions is seen also in metazoan evolution, where an increase in regulatory
459 miRNAs is correlated with increased organismal complexity [2]. Interestingly, *D. discoideum* is one of the

460 few organisms outside animals and land plants were miRNAs have been identified [24–27]. If miRNAs,
461 similar to Class I RNAs, are present in other dictyostelids is currently being investigated.
462
463 In Dictyostelia, the most complex multicellularity is found within group 4, exemplified by regulated
464 proportions of specialized cell types and a migrating slug stage [3]. Interestingly, in analogy to miRNA
465 expansion in complex animals, the number of expressed Class I RNA genes is correlated with Dictyostelia
466 complexity where group four has the largest number of Class I RNA genes (Fig. 2b). In addition, RNA-seq
467 data indicate that Class I RNAs are expressed at higher levels in group 4 (*D. discoideum*), further
468 strengthening that Class I RNAs are involved in increased organismal complexity. This increased expression
469 appears to be connected to the TGTG-box between DUSE and the TSS. So far, the TGTG-box have only
470 been identified in Class I RNA loci in species belonging group 4 (except *D. purpureum*) and *P. violaceum*
471 complex (Figs 3a and 4 and Additional file 3). Thus, this motif is likely a rather late addition in the evolution
472 of Class I RNAs. It should be noted that it is currently not known if the second motif is actually a promoter
473 element or the TSS of a longer precursor that is processed down to the mature RNA. In either case, the
474 TGTG-box is associated with Class I RNA genes in social amoebae with higher levels of complexity as
475 compared to other dictyostelids and appears to add another layer of regulation to Class I RNA expression.
476 The emergence of the TGTG-box somewhere after the split of group 3 and 4 and its connection to
477 increased phenotypic complexity seen in group 4 dictyostelids is somewhat analogous to changes in cis-
478 regulatory elements, such as enhancers, and morphological evolution in animals [42].
479
480 The high number of novel Class I RNA loci identified in Dictyostelia enabled comparative studies which
481 provides information on their key features. The short stem connecting the 5’ and 3’ ends of mature Class
482 I RNA is conserved. Furthermore, the sequence variability of the stem between organisms but also the
483 high number of compensatory mutations within each species strongly suggest that it is the structure

484 rather than sequence that is important for function. Flanking the stem structure at both ends are highly
485 conserved nucleotides present in almost all Class I RNAs: AAG and CTGT at the 5' and 3'ends respectively,
486 where the 5'G and 3'C forms the first base-pair of the stem. Curiously, in spite of the high conservation of
487 the three 5' nucleotides, the mature transcript of Class I RNAs in *D. discoideum* almost always starts with
488 the G residue, while it is more common in the group 1-3 representatives that also the two A's are included.
489 In either case, the start of each mature Class I RNA is well defined with the great majority of all RNA-seq
490 reads sharing the same 5' end. The 3'end is also remarkably conserved. Whether these nucleotides are
491 part of the mature transcript is hard to assess due to the sequencing approach used, which only allowed
492 us to capture a small fraction of the 3' ends. However, we do find reads covering the 3' conserved CTG in
493 the majority of the different Class I RNAs, suggesting that these constitute the 3' end of the mature Class
494 I RNAs. This is also supported by the full-length cDNA libraries of small RNAs in *D. discoideum* where this
495 class of RNA was first discovered [30].
496
497 The comparison of Class I RNA loci also confirms the importance of the ~11 nt motif present immediately
498 after the 5' part of the stem structure. Although some of the nucleotides vary within and in between
499 organisms, several residues are nearly perfectly conserved. Yet another conserved feature is the putative
500 promoter motif, DUSE. Based on studies of spliceosomal RNAs in *D. discoideum*, we previously
501 demonstrated that DUSE is associated with genes transcribed by both RNA Pol II and RNA Pol III [44].
502 However, we believe that Class I RNAs are transcribed by RNA Pol III since the canonical RNA Pol III
503 termination signal is present downstream of most Class I RNA loci and Class I RNA reads are lacking in
504 poly(A) enriched RNA-seq libraries.
505
506 Based on the findings in this study, we conclude that Class I RNAs were present in the last common
507 ancestor of Dictyostelia. In addition, our data also strongly suggest that the putative promoter element

508 DUSE was present 60 nt upstream of the ancestral Class I RNA gene and that the element was required
509 for expression of the gene. We also conclude that the ancient Class I RNA was characterized by a short
510 stem structure and a 11 nt sequence motif where at least six of the positions were identical to the
511 corresponding nucleotides in extant Class I RNAs. Due to the high sequence and structure variability of
512 the region between the 11 nt motif and the start of the 3' stem in identified Class I RNAs, we cannot
513 resolve this part of the ancestral sequence. However, the total length of the mature RNA was probably
514 approximately 60 nt long (Fig. 5).
515
516 Even though some motifs are highly conserved within all Class I RNAs, conservation of complete Class I
517 RNA genes are rare. Only a few examples of identical loci within the same genome are found in a handful
518 of dictyostelids and only two identical Class I RNAs shared by two different species were identified, i.e.
519 the group 4 dictyostelids *D. discoideum* and *D. firmibasis* (Additional file 4). In *D. discoideum* and *P.*
520 *pallidum* many Class I RNA genes are situated in larger clusters and it is within these clusters where the
521 species-specific identical Class I RNA genes are found, perhaps indicating expansion of Class I RNA genes
522 by duplication. Interestingly, the snRNA genes in *D. discoideum* are organized in a similar way where
523 closely related genes often are found in pairs situated very close together [33]. However, in general the
524 Class I RNA genes are spread out in the genomes and this is also true for the majority of the identical Class
525 I RNA genes in the different genomes. Hence, the low occurrence of shared synteny, low overall sequence
526 conservation and different number of loci in different organisms suggest that the expansion of Class I
527 RNAs mainly occurred after speciation.
528

529 **Conclusions**

530 We have identified Class I RNAs in 16 different dictyostelids and validated their expression in
531 representatives of each major group of Dictyostelia dating back at least ~600 million years. Despite the
532 large evolutionary distances, Class I RNA genes share promotor motifs and the mature RNAs have several
533 characteristics in common, i.e. short stem, conserved sequence motif and highly conserved 5' and 3' ends.
534 In addition, the *D. discoideum* Class I RNA interacting protein CIBP is conserved throughout Dictyostelia.
535 Furthermore, the gene shares its expression profile with Class I RNAs during early development, indicating
536 that function and mode of action is also conserved. Although Class I RNAs are present in all dictyostelids
537 investigated, no evidence was found for this class of RNAs in any other organism. Taken together, our
538 results suggest that Class I RNAs are important for the evolution of multicellularity in Dictyostelia.

539

540 **Materials and methods**

541

542 **RNA isolation and northern blot**

543 The following strains were used for northern blot validation of Class I expression: *D. discoideum* AX2
544 (DBS0235521, www.dictybase.org), *D. purpureum* WS321, *P. pallidum* PN500, *D. lacteum* Konijn, *D.*
545 *fasciculatum* SH3, *A. subglobosum* LB1. All strains, except *D. discoideum,* were kindly provided by Dr Maria
546 Romeralo and Professor Sandra Baldauf. Total RNA was extracted with TRIzol (ThermoFisher Scientific)
547 from cells grown in association with *Klebsiella aerogenes* on non-nutrient agar. Northern blots were
548 performed as described in [30, 31]. Briefly, 10 µg total RNA was separated on 8 % PAGE/7 M Urea and
549 electroblotted to Hybond N+ nylon membranes (GE Healthcare). After UV crosslinking, immobilized RNA
550 was hybridized with ^{32}P-labeled oligonucleotides in Church buffer at 42 °C overnight. Signals were
551 analyzed with a Personal Molecular Imager (BIO-RAD) normally after a few hours' exposure. Membranes

552 analyzed more than once were stripped with 0.1×SSC/1 % SDS buffer at 95 °C for 1 h and controlled for
553 residual signal before reprobing. Oligonucleotide sequences are provided in Additional file 2: Table S1.
554
555 **RNA-seq validation**
556 Strain growth and RNA extraction for RNA-seq have been described previously [20]. For each strain, RNA
557 was prepared from growing cells and two multicellular developmental stages: aggregates and tipped
558 aggregates/fingers (biological duplicates except for *D. lacteum*). Truseq small RNA Sample Preparation kit
559 (product # RS- 200-0012, Illumina) was used to prepare sequencing libraries from 1 µg total RNA. The
560 library preparation was performed according to the manufacturer's protocol (#15004197 rev G) where
561 cDNA representing 18 nt – 70 nt RNA were isolated. Single read 50 bp sequencing was performed using
562 v4 sequencing chemistry on an Illumina HiSeq2500. To reduce influence of degradation products, only full
563 length 50 bp reads were mapped with bowtie allowing for 1 mismatch [45] and counted with feature
564 counts [46]. Read coverage over Class I candidate loci were calculated with BEDTools genomecov v. 2.26.0
565 [47]. Class I RNAs were considered to be validated by RNA-seq if the read coverage indicated a distinct 5'
566 end and reads specifically matched the predicted loci, i.e. did not appear to be part of a considerably
567 longer transcript. Principal component analyses were performed using DESeq2 [48]. ROC plot evaluation
568 of Class I prediction was performed using pROC package[49] in R.
569
570 **Strains and genomic resources**
571 Strain names and accession numbers for genomic sequences are listed in Additional file 1.
572
573 **Identification of Class I RNAs**

574 Of the 40 *D. discoideum* Class I RNAs annotated previous to this study, six were excluded from the model
575 build (r48, r53, r54, r55, r58 and r61) as they represented truncated fragments or lacked the canonical

576 features such as the stem or conserved sequence motif. The remaining 34 were aligned with MAFFT
577 v7.407 [50] using the ginsi setting. Consensus structure for the alignment was predicted with RNAalifold
578 2.3.3 [51] using the -T 22 option to account for the optimal growing temperature of the amoebae.
579 Alignment and consensus structure was combined to a Stockholm alignment file and a co-variance model
580 was created with Infernal 1.1.2 [32]. Infernal was then used to search the genomes of *D. discoideum, D.*
581 *purpureum, D. lacteum, P. pallidum, A. subglobosum* and *D. fasciculatum* using default settings and
582 candidates with a score ≥ 25 were added to the alignment using MAFFT (ginsi –add). The new alignment
583 was manually curated and used to predict consensus structure, create a new co-variance model, and
584 perform a new search in the same genomes. This procedure was iterated six time, i.e. until no new
585 candidates with an Infernal score ≥ 25 were identified. Enriched sequence motifs were identified up to
586 150 nt upstream of identified candidates with MEME v. 5.0.3 [52]. For final Class I identification, Infernal
587 searches were performed with increased sensitivity and all candidates scoring 15 or higher were kept
588 (cmsearch –nohmm –notrunc -T 15). The candidates were then evaluated based on the presence of DUSE
589 and TGTG-box in the 150 nt preceding the predicted start of transcription (see above) using FIMO v. 5.0.3
590 [34]. Infernal score, FIMO motif score, and a motif distance score (+5) were then added to a total score.
591 Missing DUSE or incorrect distance was penalized with -10 or -5 respectively. If a total score of 32 was
592 achieved, the candidate was considered likely to be an expressed true Class I RNA and kept for further
593 analyses. Representative sequence logos of manually curated sequence alignments (mafft --maxiterate
594 1000 –localpair) were created with WebLogo 3 [53].
595
596 **Orthologue identification and shared synteny search**
597 For Dictyostelia species lacking genome annotations, gene prediction was performed with Gene id v. 1.4
598 [54] using Dictyostelium parameter file. Orthologue identification was performed using OrthoFinder v.
599 2.3.3 [55]. Protein domain architectures for CIBP/Rnp1A orthologues were analyzed with hmmscan using

600 the HMMER web interface [56]. Orthologue information for *D. discoideum, D. firmibasis*, *D. lacteum*, *P.*
601 *pallidum, A. subglobosum* and *D. fasciculatum* was used to investigate shared synteny for Class I RNA loci.
602 For each Class I RNA locus, gene information within a 10 kb flanking region was retrieved. Next, we
603 searched for orthologues of these genes in the other organisms included in the search. If a Class I RNA
604 was found within 10 kb of an orthologous gene in another organism, it was manually inspected to
605 determine the level of shared synteny.

606 **Declarations**

607

608 **Ethics approval and consent to participate**

609 Not applicable

610

611 **Consent for publication**

612 Not applicable

613

614 **Availability of data and material**

615

616 **Competing interest**

617 The authors declare that they have no competing interests

618

619 **Authors' contribution**

620 JK, LA, JR, and FS participated in the design of the project; LA and ZL generated experimental data; JK
621 performed most of the data analysis and prepared figures; JR participated in the bioinformatic analysis;
622 L.E, A.N, G.G, and P.S supplied RNA for sequencing. JK and FS drafted the manuscript.

623

Funding

This work was supported by the Linneus Support from the Swedish Research Council to the Uppsala RNA Research Centre; Carl Tryggers Foundation under Grant number CST 18:381.

Acknowledgements

Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

633 **References**

634 1. Cech TR, Steitz JA. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. Cell.
635 2014;157:77–94.

636 2. Gaiti F, Calcino AD, Tanurdžić M, Degnan BM. Origin and evolution of the metazoan non-coding
637 regulatory genome. Developmental Biology. 2017;427:193–202.

638 3. Kawabe Y, Du Q, Schilde C, Schaap P. Evolution of multicellularity in Dictyostelia. Int J Dev Biol.
639 2019;63:359–69.

640 4. Brown MW, Spiegel FW, Silberman JD. Phylogeny of the "forgotten" cellular slime mold, Fonticula
641 alba, reveals a key evolutionary branch within Opisthokonta. Mol Biol Evol. 2009;26:2699–709.

642 5. Brown MW, Silberman JD, Spiegel FW. A contemporary evaluation of the acrasids (Acrasidae,
643 Heterolobosea, Excavata). Eur J Protistol. 2012;48:103–23.

644 6. He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. An alternative root for the eukaryote
645 tree of life. Curr Biol. 2014;24:465–70.

646 7. Tice AK, Silberman JD, Walthall AC, Le KND, Spiegel FW, Brown MW. Sorodiplophrys stercorea:
647 Another Novel Lineage of Sorocarpic Multicellularity. J Eukaryot Microbiol. 2016;63:623–8.

648 8. Lasek-Nesselquist E, Katz LA. Phylogenetic position of Sorogena stoianovitchae and relationships
649 within the class Colpodea (Ciliophora) based on SSU rDNA sequences. J Eukaryot Microbiol.
650 2001;48:604–7.

651 9. Brown MW, Silberman JD, Spiegel FW. "Slime molds" among the Tubulinea (Amoebozoa): molecular
652 systematics and taxonomy of Copromyxa. Protist. 2011;162:277–87.

653 10. Brown MW, Kolisko M, Silberman JD, Roger AJ. Aggregative multicellularity evolved
654 independently in the eukaryotic supergroup Rhizaria. Curr Biol. 2012;22:1123–7.

655 11. Heidel AJ, Lawal HM, Felder M, Schilde C, Helps NR, Tunggal B, et al. Phylogeny-wide analysis of
656 social amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res.
657 2011;21:1882–91.

658 12. dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. Uncertainty in the
659 Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. Curr Biol.
660 2015;25:2939–50.

661 13. Schilde C, Lawal HM, Kin K, Shibano-Hayakawa I, Inouye K, Schaap P. A well supported multi gene
662 phylogeny of 52 dictyostelia. Molecular Phylogenetics and Evolution. 2019;134:66–73.

663 14. Romeralo M, Skiba A, Gonzalez-Voyer A, Schilde C, Lawal H, Kedziora S, et al. Analysis of
664 phenotypic evolution in Dictyostelia highlights developmental plasticity as a likely consequence of
665 colonial multicellularity. Proceedings of the Royal Society B: Biological Sciences. 2013;280:20130976.

666 15. Schilde C, Skiba A, Schaap P. Evolutionary reconstruction of pattern formation in 98 Dictyostelium

667 species reveals that cell-type specialization by lateral inhibition is a derived trait. EvoDevo. 2014;5:34.

668 16. Sheikh S, Thulin M, Cavender JC, Escalante R, Kawakami S-I, Lado C, et al. A New Classification of
669 the Dictyostelids. Protist. 2018;169:1–28.

670 17. Eichinger L, Pachebat JA, Gloeckner G, Rajandream M-A, Sucgang R, Berriman M, et al. The
671 genome of the social amoeba Dictyostelium discoideum. Nature. 2005;435:43–57.

672 18. Sucgang R, Kuo A, Tian X, Salerno W, Parikh A, Feasley CL, et al. Comparative genomics of the
673 social amoebae Dictyostelium discoideum and Dictyostelium purpureum. Genome Biol. 2011;12:R20.

674 19. Urushihara H, Kuwayama H, Fukuhara K, Itoh T, Kagoshima H, Shin-I T, et al. Comparative genome
675 and transcriptome analyses of the social amoeba Acytostelium subglobosum that accomplishes
676 multicellular development without germ-soma differentiation. BMC Genomics. 2015;16:80.

677 20. Glöckner G, Lawal HM, Felder M, Singh R, Singer G, Weijer CJ, et al. The multicellularity genes of
678 dictyostelid social amoebas. Nature Communications. 2016;7:1–11.

679 21. Hillmann F, Forbes G, Novohradská S, Ferling I, Riege K, Groth M, et al. Multiple Roots of Fruiting
680 Body Formation in Amoebozoa. Genome Biol Evol. 2018;10:591–606.

681 22. Deline B, Greenwood JM, Clark JW, Puttick MN, Peterson KJ, Donoghue PCJ. Evolution of
682 metazoan morphological disparity. Proc Natl Acad Sci U S A. 2018;115:E8909–18.

683 23. Rosengarten RD, Santhanam B, Fuller D, Katoh-Kurasawa M, Loomis WF, Zupan B, et al. Leaps and
684 lulls in the developmental transcriptome of Dictyostelium discoideum. BMC Genomics. 2015;16:294.

685 24. Avesson L, Reimegard J, Wagner EGH, Soderbom F. MicroRNAs in Amoebozoa: Deep sequencing
686 of the small RNA population in the social amoeba Dictyostelium discoideum reveals developmentally
687 regulated microRNAs. RNA. 2012;18:1771–82.

688 25. Liao Z, Kjellin J, Hoeppner MP, Grabherr M, Söderbom F. Global characterization of the Dicer-like
689 protein DrnB roles in miRNA biogenesis in the social amoeba Dictyostelium discoideum. RNA Biol.
690 2018;15:937–54.

691 26. Hinas A, Reimegård J, Wagner EGH, Nellen W, Ambros VR, Söderbom F. The small RNA repertoire
692 of Dictyostelium discoideum and its regulation by components of the RNAi pathway. Nucl Acids Res.
693 2007;35:6714–26.

694 27. Meier D, Kruse J, Buttlar J, Friedrich M, Zenk F, Boesler B, et al. Analysis of the Microprocessor in
695 Dictyostelium: The Role of RbdB, a dsRNA Binding Protein. PLoS Genet. 2016;12:e1006057.

696 28. Rosengarten RD, Santhanam B, Kokosar J, Shaulsky G. The Long Noncoding RNA Transcriptome of
697 Dictyostelium discoideum Development. G3 (Bethesda). 2016;7:387–98.

698 29. Hildebrandt M, Nellen W. Differential antisense transcription from the Dictyostelium EB4 gene locus:
699 Implications on antisense-mediated regulation of mRNA stability. Cell. 1992;69:197–204.

700 30. Aspegren A, Hinas A, Larsson P, Larsson A, Söderbom F. Novel non-coding RNAs in Dictyostelium

701 discoideum and their expression during development. Nucl Acids Res. 2004;32:4646–56.

702 31. Avesson L, Schumacher HT, Fechter P, Romby P, Hellman U, Söderbom F. Abundant class of non-
703 coding RNA regulates development in the social amoeba Dictyostelium discoideum. RNA Biol.
704 2011;8:1094–104.

705 32. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics.
706 2013;29:2933–5.

707 33. Hinas A, Soederbom F. Treasure hunt in an amoeba: non-coding RNAs in Dictyostelium discoideum.
708 Current Genetics. 2007;51:141–59.

709 34. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics.
710 2011;27:1017–8.

711 35. Richard P, Manley JL. Transcription termination by nuclear RNA polymerases. Genes Dev.
712 2009;23:1247–69.

713 36. Muñoz-Dorado J, Marcos-Torres FJ, García-Bravo E, Moraleda-Muñoz A, Pérez J. Myxobacteria:
714 Moving, Killing, Feeding, and Surviving Together. Front Microbiol. 2016;7.
715 doi:10.3389/fmicb.2016.00781.

716 37. Ngo T, Miao X, Robinson DN, Zhou Q. An RNA-binding protein, RNP-1, protects microtubules from
717 nocodazole and localizes to the leading edge during cytokinesis and cell migration in Dictyostelium cells.
718 Acta Pharmacologica Sinica. 2016;37:1449–57.

719 38. Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, et al. Conserved developmental
720 transcriptomes in evolutionarily divergent species. Genome Biology. 2010;11:R35.

721 39. Mosig A, Sameith K, Stadler P. Fragrep: An Efficient Search Tool for Fragmented Patterns in
722 Genomic Sequences. Genomics Proteomics Bioinformatics. 2006;4:56–60.

723 40. Antolović V, Lenn T, Miermont A, Chubb JR. Transition state dynamics during a stochastic fate
724 choice. Development. 2019;146. doi:10.1242/dev.173740.

725 41. Gaunt SJ, Paul Y-L. Changes in Cis-regulatory Elements during Morphological Evolution. Biology
726 (Basel). 2012;1:557–74.

727 42. Hinas A, Larsson P, Avesson L, Kirsebom LA, Virtanen A, Söderbom F. Identification of the major
728 spliceosomal RNAs in Dictyostelium discoideum reveals developmentally regulated U2 variants and
729 polyadenylated snRNAs. Eukaryotic Cell. 2006;5:924–34.

730 43. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short
731 DNA sequences to the human genome. Genome Biol. 2009;10:R25.

732 44. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning
733 sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

734 45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.

735 Bioinformatics. 2010;26:841–2.

736 46. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data
737 with DESeq2. Genome Biol. 2014;15:550.

738 47. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source
739 package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

740 48. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in
741 Performance and Usability. Mol Biol Evol. 2013;30:772–80.

742 49. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure
743 prediction for RNA alignments. BMC Bioinformatics. 2008;9:474.

744 50. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in
745 biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.

746 51. Crooks GE. WebLogo: A Sequence Logo Generator. Genome Research. 2004;14:1188–90.

747 52. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinformatics.
748 2007;Chapter 4:Unit 4.3.

749 53. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons
750 dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

751 54. HMMER web server: 2018 update | Nucleic Acids Research | Oxford Academic.
752 https://academic.oup.com/nar/article/46/W1/W200/5037715. Accessed 10 Mar 2020.

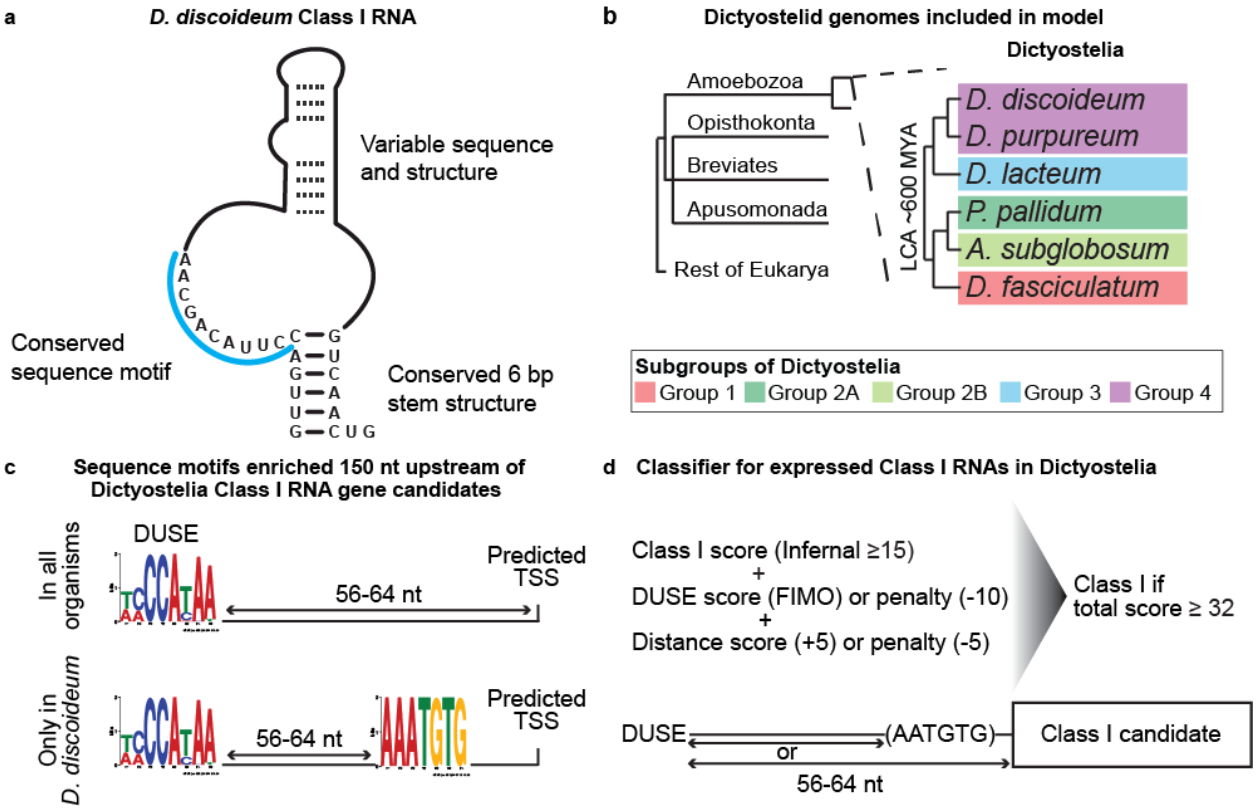
753 55. Burki F, Roger AJ, Brown MW, Simpson AGB. The New Tree of Eukaryotes. Trends in Ecology &
754 Evolution. 2020;35:43–55.

755

756

757 **Figures**

758



759

760 **Figure 1. Search strategy and classification of Class I RNAs** a) Schematic representation of previously

761 described *D. discoideum* Class I RNAs [31]. b) Schematic phylogeny showing the location of Amoebozoa,

762 a sister group to Obazoa (Opisthokonta, Breviates and Apusomonada) in the eukaryotic tree of life based

763 on [35]. Dictyostelia is represented by species belonging to each major group [13]. The genomes of

764 these dictyostelids were searched for Class I RNA genes and newly identified genes were used to refine

765 the co-variance model. c) Enriched sequence motifs identified upstream of Class I RNA gene candidates

766 in the different dictyostelids represented in panel b (Infernal score ≥ 25, n=126). The putative promoter

767 motif (DUSE) is found approximately 60 nt from the predicted start of transcription (TSS) in all organisms

768 (upper part). DUSE in combination with TGTG-box, only identified in *D. discoideum* (lower part). d)

769 Summary of scoring system used for the classifier of Dictyostelia Class I RNA based on Infernal score ≥

770 15, presence of DUSE, and distance between DUSE and predicted TSS or TGTG-box.

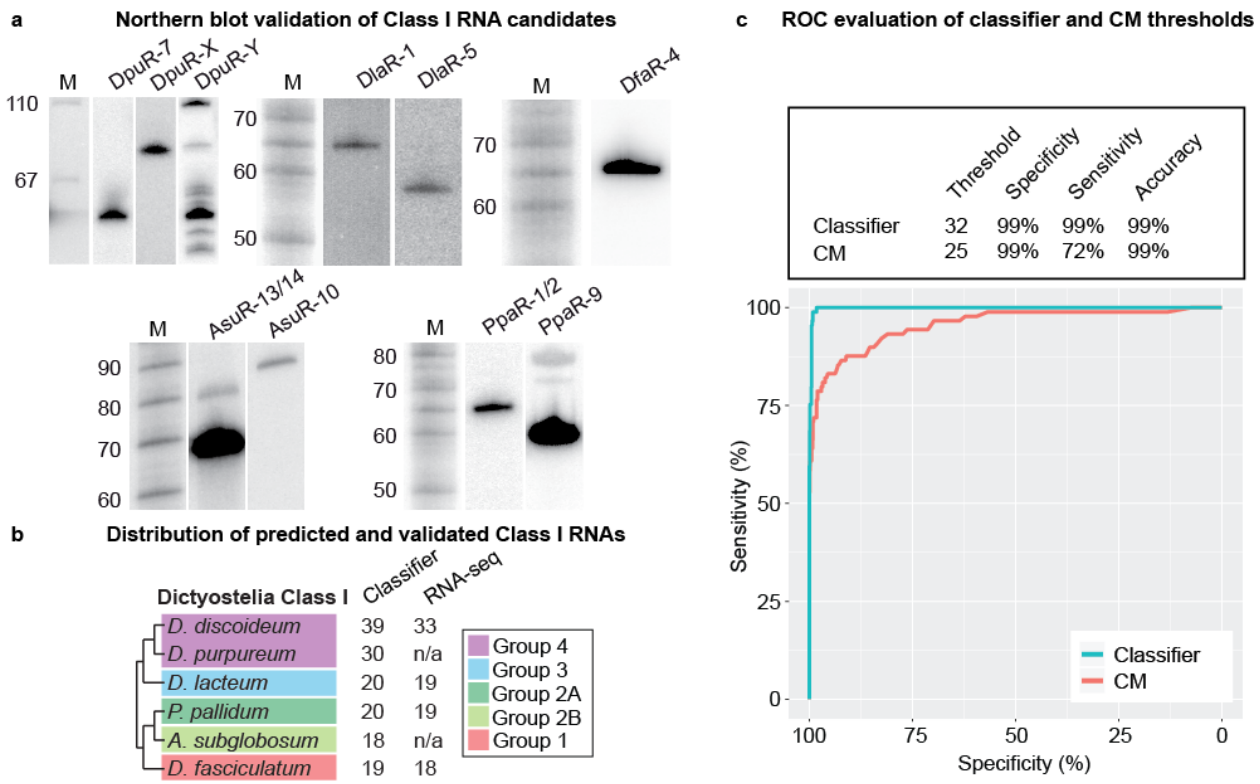


Figure 2. Expression of predicted Class I RNA genes. a) Northern blot validation of different Class I RNAs from *D. purpureum* (DpuR), *D. lacteum* (DlaR), *D. fasciculatum* (DfaR), *A. subglobosum* (AsuR), and *P. pallidum* (DpaR). The number after the species-specific designations indicates which Class I RNA the probe recognizes. When two numbers are given, the probe recognize two different Class I RNAs. DpuR-X indicates that the probe is expected to hybridize to six different Class I RNAs predicted to be 85 nt long. DpuR-Y indicates that the probe is expected to hybridize to 24 Class I RNAs. For each organism except for *D. fasciculatum* (DfaR), the same membrane was probed, stripped, and reprobed for the different Class I RNAs. Radioactively labeled size marker is indicated by M and numbers to the left indicate sizes in nucleotides. b) Number of Class I RNA genes in each species according to the classifier. RNA-seq designate the number of expressed Class I RNA genes verified by RNA-seq. c) ROC curves based on the RNA-seq validation and either classifier score or CM score for all Class I candidates in *D. discoideum, D. lacteum, P. pallidum,* and *D. fasciculatum* identified in the CM search. Input data available in Additional

784 file 3. Evaluation of the classifier and CM thresholds used throughout the study are shown above the

785 plot. Individual ROC curves for each organism are found in Additional file 2: Fig. S2.

786

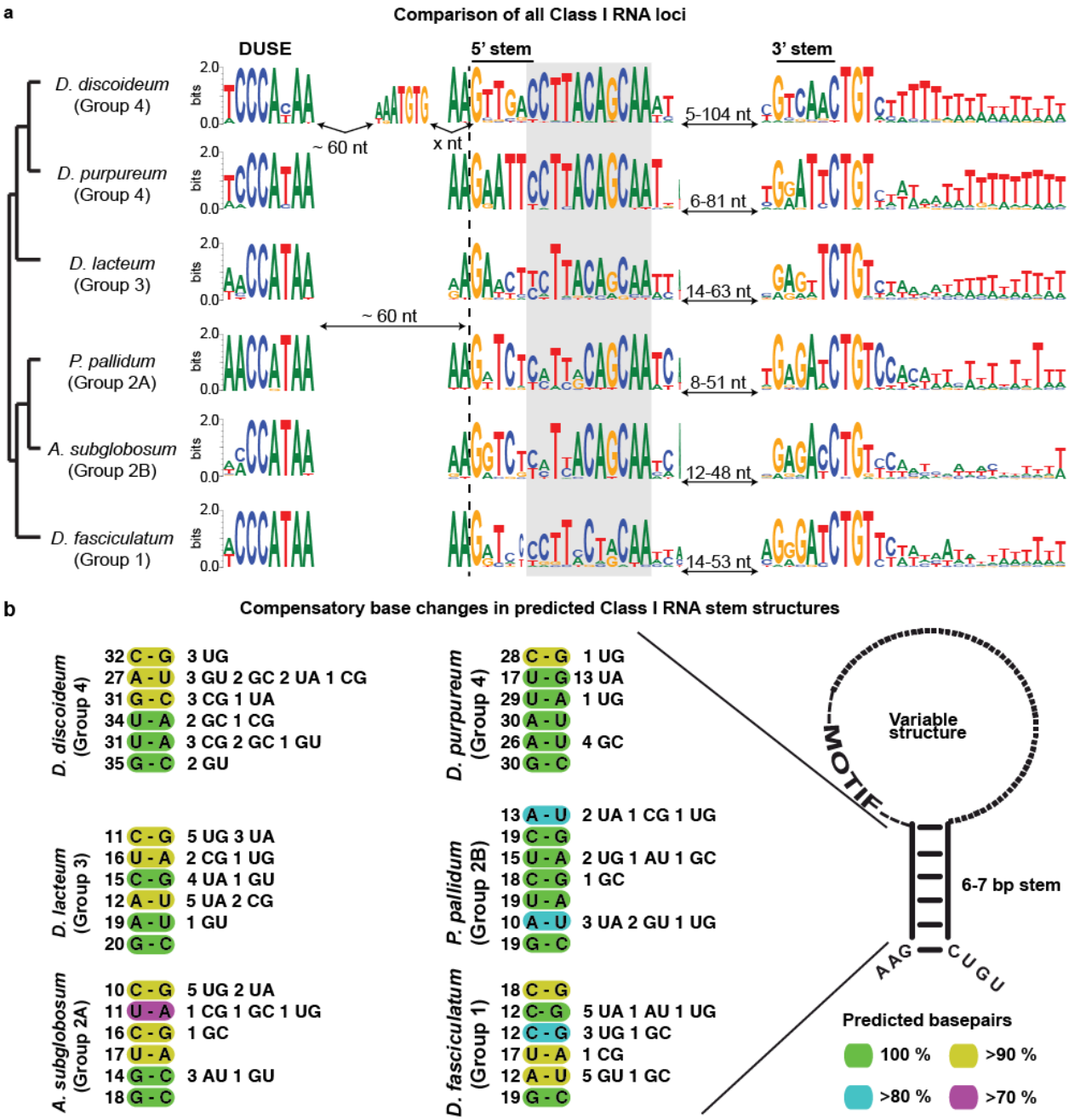


Figure 3. Conserved characteristics of Class I RNAs a) Sequence logo representing species specific alignments of conserved features of Class I RNA loci. DUSE indicate the putative promoter element and the 5' and 3´stem sequences of the conserved stem-structure are indicated. The conserved 11 nt sequence motif adjacent to the 5´part of the stem is boxed in gray. The dashed vertical line denotes the predicted start of the Class I RNAs based on the CM search. The sequence logo between DUSE and the 5' stem motif for *D. discoideum* represent the TGTG-box. Numbers of nucleotides (nt) correspond to the distances

794 between indicated motifs. b) Displayed are nucleotides representing the most common base pair for each
795 position (numbers to the left) of the conserved stem structure for each organism (color key down right).
796 Numbers to the right represent less common nucleotide combinations predicted to base pair. The few
797 combinations of nucleotides not predicted to base pair are not shown. Schematic structure of Class I RNA
798 is presented to the right.
799

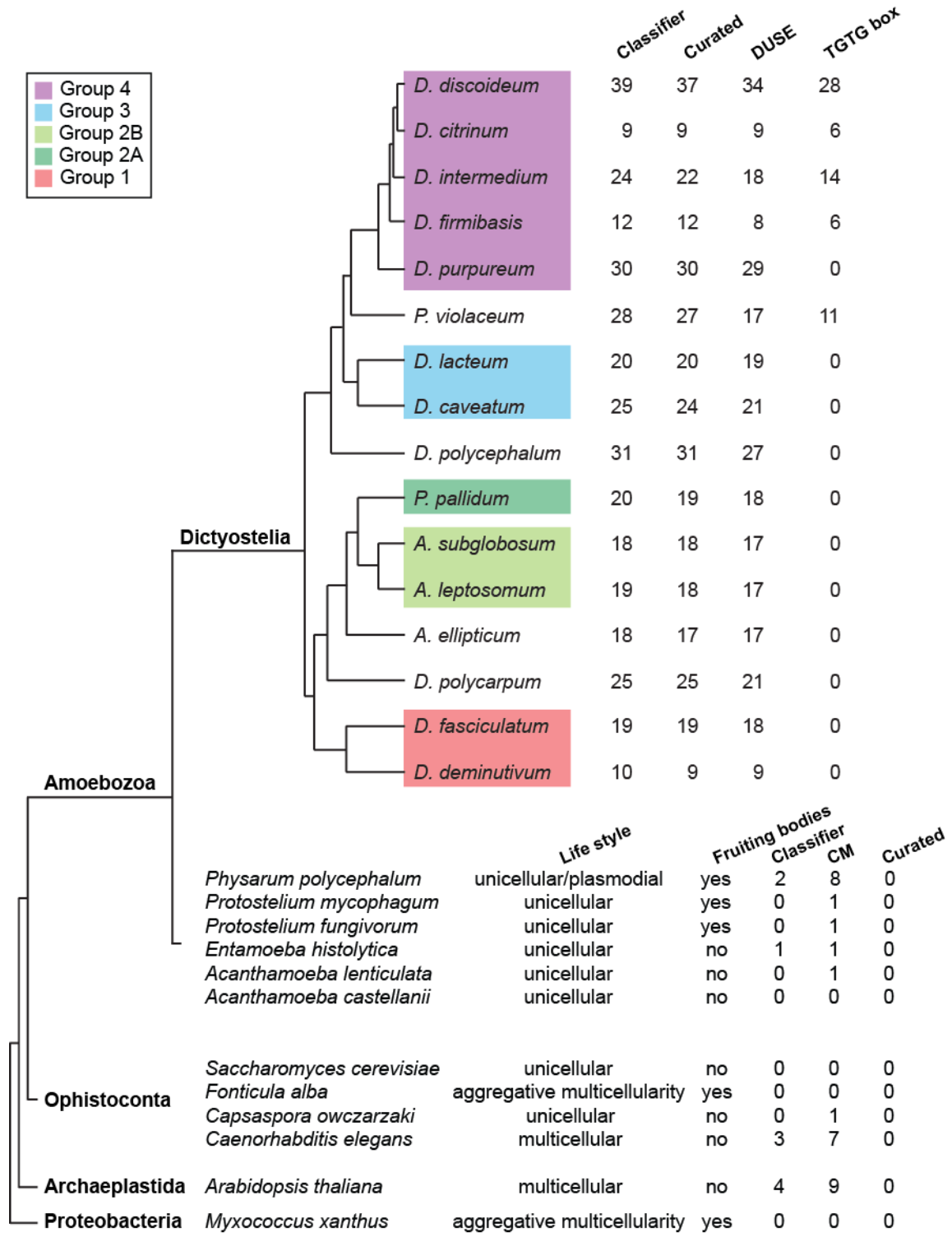


Figure 4. Class I RNAs are ubiquitous in and restricted to dictyostelid social amoebae.

Upper part: Class I RNA loci were searched for in the genomes of 16 different Dictyostelia (Additional file 1). The number of hits identified by the classifier is indicated as well as the number of these loci that

804 passed manual curation. The number of curated loci with DUSE and the TGTG-box at the correct
805 distance are shown. Bottom part: Result from searches for Class I RNA loci outside Dictyostelia. Life style
806 indicate unicellular or multicellular organisms. Fruiting bodies denotes if the organism life style involves
807 formation of fruiting bodies. CM indicates number of candidates identified with a CM score ≥ 25.
808 Headings Classifier and Curated as described above. Further information about outgroup Class I RNA
809 candidates is available in Additional file 7.

810

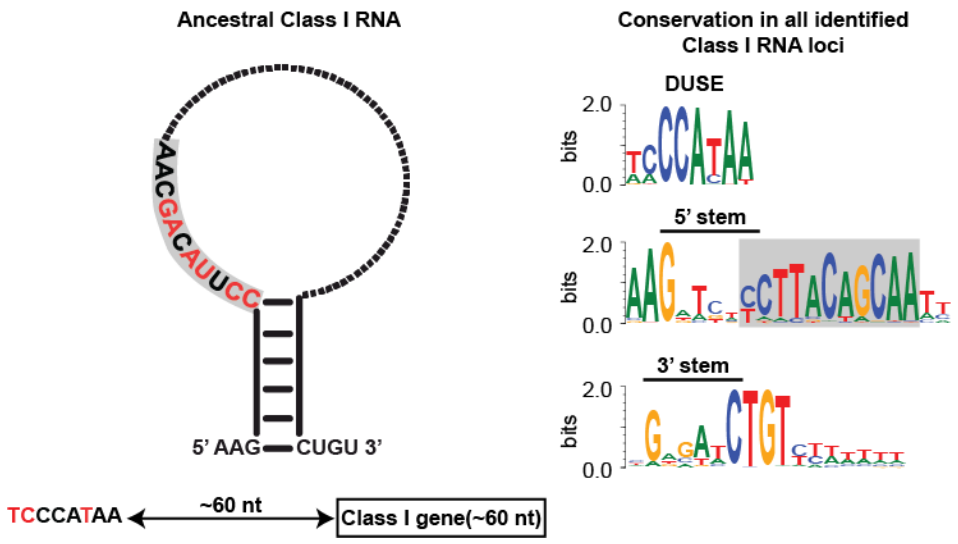


Figure 5. Ancestral Class I RNA and conserved key features

Left: Schematic representation of the ancestral Class I RNA transcript. The putative promoter element DUSE and its distance to the Class I RNA gene is indicated below the Class I RNA structure. Strongly conserved features, nucleotide and base paired stem, are colored black while red denotes more variable positions (based on data presented in figure 3). The dotted part of the loop indicates highly variable sequence. Right: Sequence logos of alignments of conserved sequence motifs from all identified and curated Class I RNAs. Only DUSE identified at the correct distance were included in the alignment. The 11 nt motif is boxed in gray in both the ancestral Class I RNA (left) and sequence logo (right).

821 **Description of additional files**

822

823 **File name:** Additional file 1

824 **File format:** excel (.xlsx)

825 **Title of data:** Genomic resources used in this study.

826 **Description:** References for dictyostelid genomic sequences used in the study together with alternative

827 names (Synonyms) for organisms (where applicable) and basic genome sequence statistics. Genome

828 sequences used in the construction of the co-variance model (CM) are indicated in bold. For outgroup

829 genome sequences used in the study, only the GenBank or RefSeq reference are given.

830

831 **File name:** Additional file 2

832 **File format:** PDF (.pdf)

833 **Title of data:** Supplementary figures and tables

834 **Description:** Supplementary figures 1-7 and table 1

835

836 **File name:** Additional file 3

837 **File format:** excel (.xlsx)

838 **Title of data:** Spreadsheet with search summary and expression validation for Class I RNAs from selected
839 dictyostelids.

840 **Description:** The spreadsheet show data for all predicted Class I RNAs, from respective organism,
841 passing the CM score ≥ 15. Candidates considered to be true Class I RNAs have been named according to
842 previous naming rules. Classifier summary includes classifier score and infernal (CM) score as well as
843 identified DUSE sequence (DUSE_seq), the duse score obtained from FIMO (duse_score), the duse
844 distance from the predicted Class I RNA start (duse_dist), the TGTG-box sequence (TGTG-box_seq) and
845 score obtained from FIMO (TGT-box_score), TGTG distance from predicted Class I start (TGT-box_dist)
846 and the distance between DUSE and TGTG-box (inter_dist). Read counts denotes raw RNA-seq read
847 counts for each Class I RNA candidate in each sequencing library for the organisms where we performed
848 RNA-seq. Read counts include time points, for growning cells (0h) and time after starvation, when RNA
849 was collected and A and B denotes biological replicates. Expression validation is based on read counts
850 and coverage over each respective locus and northern blot. For Read coverage, the values 1 and 0
851 denotes validated and not validated, respectively. The read coverage validation in combination with
852 either CM score or classifier score was used to generate ROC curves in figure 2 and Additional file 2: Fig.
853 S3. Northern blot indicates Class I RNAs analyzed by northern blot in this study.

854

855 **File name:** Additional file 4

856 **File format:** excel (.xlsx)

857 **Title of data:** Spreadsheet with name, genome location, validation and score of all curated Class I genes
858 for all investigated dictyostelids.

859 **Description:** For each Class I RNA gene Genomic location includes chromosome/contig/scaffold
860 (Chromosome), start and end of predicted RNA (Start and Stop, respectively), on which DNA strand the
861 RNA is located (Strand), and the predicted length (Class I RNA length). The Classifier summary includes
862 the classifier score, the CM score the DUSE sequence and score (DUSE_seq and duse_score,
863 respectively)) and TGTG-box sequence and score (TGTG-box_seq and TGTG-box_score, respecively).
864 Columns duse_dist and TGTG-box_dist denotes the distance between the motif and predicted TSS and
865 inter_dist denotes the distance between the two motifs. Under Upstream motifs, columns DUSE and
866 TGTG-box summarize the classifier output and indicate if respective motif is found at the correct
867 distance (1) or not (0). Where applicable, information about RNA-seq (1 denotes expressed Class I based
868 on read coverage and 0 if expression could not be validated) and/or northern blot validation is given.

869

870

871 **File name:** Additional file 5

872 **File format:** PDF (.pdf)

873 **Title of data:** Genomic distribution of Class I RNA loci.

874 **Description:** Chromosome/contig/scaffold names are indicated to the left and respective lengths are
875 normalized to the longest one (length in bp are presented to the very right of each schematic stretch of
876 DNA). Total number of Class I RNA genes for each organism is given at the top and for each
877 chromosome/contig/scaffold to the very right. Class I RNA genes are indicated by black arrows except
878 for genes with identical sequences, which are indicated with colored arrows. Short vertical lines specify
879 every 0.5 mbp. In *D. discoideum*, the duplication on chromosome 2 (DDB0232429) is indicated by grey
880 boxes with vertical lines.

881 **File name:** Additional file 6

882 **File format:** PDF (.pdf)

883 **Title of data:** Shared synteny between *D. discoideum* and *D. firmibasis* Class I RNA loci.

884 **Description:** Representation of *D. firmibasis* (dfi) Class I RNA loci +/- 10 kb that share synteny with *D.*
885 *discoideum* (ddi) Class I RNAs supported by at least two orthologous genes (connected with dashed
886 lines). Chromosome/contig names are given to the left of the genes. Genes situated on the forward and
887 the reverse strand are indicated above and below the chromosome/contig, respectively. Vertical grey
888 lines are given at every 1000 nt.

889

890 **File name:** Additional file 7

891 **File format:** excel (.xlsx)

892 **Title of data:** Identified Class I RNA candidates in outgroups.

893 **Description:** For each candidate (classifier score ≥ 32 and/or CM score ≥ 25) the following information is
894 given: classifier score, CM score as well as identified DUSE sequence (DUSE_seq), the duse_score
895 obtained from FIMO, the distance of DUSE (duse_dist) from the predicted Class I start, the TGTG-box
896 sequence (TGTG-box_seq) and score (TGTG-box_score) obtained from FIMO, TGTG_distance from
897 predicted Class I start and the distance between DUSE and TGTG-box (inter_dist). The last column (Seq),
898 contains the full genomic sequence of each candidate.

899