Current Biology

Report

Unicellular Origin of the Animal MicroRNA Machinery

Highlights

- The animal-specific miRNA Microprocessor is discovered in unicellular Ichthyosporea
- The origin of the animal miRNA machinery was independent of animal multicellularity
- The Microprocessor is lost in ctenophores and is not an ancestral animal trait
- Several ichthyosporeans harboring the Microprocessor express bona fide miRNAs

Authors

Jon Bråte, Ralf S. Neumann,
Bastian Fromm, ..., Iñaki Ruiz-Trillo,
Paul E. Grini, Kamran Shalchian-Tabrizi

Correspondence

kamran@ibv.uio.no

In Brief

In animals, microRNAs and the miRNA biogenesis machinery are essential for correct organismal development. Bråte et al. demonstrate that the core of this machinery, the Microprocessor, is not an animal innovation but originated among their unicellular relatives. Several unicellular species harboring the Microprocessor also express bona fide miRNAs.

Bråte et al., 2018, Current Biology *28*, 1–8
October 22, 2018 © 2018 The Authors. Published by Elsevier Ltd.
https://doi.org/10.1016/j.cub.2018.08.018

Current Biology
Report

CellPress

Unicellular Origin of the Animal MicroRNA Machinery

Jon Bråte,[1] Ralf S. Neumann,[1] Bastian Fromm,[2,3] Arthur A.B. Haraldsen,[1] James E. Tarver,[4] Hiroshi Suga,[5] Philip C.J. Donoghue,[4] Kevin J. Peterson,[6] Iñaki Ruiz-Trillo,[7,8] Paul E. Grini,[1] and Kamran Shalchian-Tabrizi[1,9,*]

[1]Centre for Epigenetics, Development and Evolution (CEDE) and Centre for Integrative Microbial Evolution (CIME), Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, Oslo, Norway
[2]Department of Tumor Biology, Institute for Cancer Research, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
[3]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, 10691 Stockholm, Sweden
[4]School of Earth Sciences, University of Bristol, Bristol BS8 1TQ, UK
[5]Faculty of Life and Environmental Sciences, Prefectural University of Hiroshima, Nanatsuka 562, Shobara, Hiroshima 727-0023, Japan
[6]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA
[7]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[8]ICREA, 08010 Barcelona, Spain
[9]Lead Contact
*Correspondence: kamran@ibv.uio.no

https://doi.org/10.1016/j.cub.2018.08.018

SUMMARY

The emergence of multicellular animals was associated with an increase in phenotypic complexity and with the acquisition of spatial cell differentiation and embryonic development. Paradoxically, this phenotypic transition was not paralleled by major changes in the underlying developmental toolkit and regulatory networks. In fact, most of these systems are ancient, established already in the unicellular ancestors of animals [1–5]. In contrast, the Microprocessor protein machinery, which is essential for microRNA (miRNA) biogenesis in animals, as well as the miRNA genes themselves produced by this Microprocessor, have not been identified outside of the animal kingdom [6]. Hence, the Microprocessor, with the key proteins Pasha and Drosha, is regarded as an animal innovation [7–9]. Here, we challenge this evolutionary scenario by investigating unicellular sister lineages of animals through genomic and transcriptomic analyses. We identify in Ichthyosporea both *Drosha* and *Pasha* (*DGCR8* in vertebrates), indicating that the Microprocessor complex evolved long before the last common ancestor of animals, consistent with a pre-metazoan origin of most of the animal developmental gene elements. Through small RNA sequencing, we also discovered expressed bona fide miRNA genes in several species of the ichthyosporeans harboring the Microprocessor. A deep, pre-metazoan origin of the Microprocessor and miRNAs comply with a view that the origin of multicellular animals was not directly linked to the innovation of these key regulatory components.

RESULTS AND DISCUSSION

Recent genomic and molecular data have revealed that the unicellular ancestors of animals already had most of the complex genetic repertoire essential for multicellular development and cellular differentiation [2, 10, 11]. One striking exception is the animal microRNA (miRNA) pathway. This pathway is required for correct development of most animal lineages but has not been discovered outside of the animal kingdom [6] (among animals, only Ctenophora lack the miRNA pathway [12–14]). It consists of the Microprocessor protein machinery, which is essential for miRNA biogenesis, and the resulting miRNAs that post-transcriptionally regulate mRNAs (Figure 1A) [15]. The view that the animal miRNA pathway is specific to animals is supported by the fact that the closest unicellular relatives to animals, the choanoflagellates (Figure 1B), lack the *Drosha* and *Pasha* (*DGCR8* in vertebrates) genes that make up the Microprocessor, as well as other key components of the miRNA processing machinery [6]. This evolutionary scenario is compelling and could give insight into the genetic mechanisms underlying the origin of animals. However, as only a single unicellular holozoan (the clade that comprises Metazoa and their closest unicellular relatives) has been sampled thus far, the absence of the Microprocessor in choanoflagellates could reflect the loss of an ancient pathway invented prior to the animal-choanoflagellate divergence. Indeed, gene losses, especially within the choanoflagellates, are much more frequent in eukaryotic evolution than previously thought [16]. Thus, robust inferences of the timing and sequence of innovations of the animal miRNA processing machinery, and the origin of animal miRNAs, require analysis of other unicellular sister lineages to the animals. Filasterea and Ichthyosporea are particularly interesting because, with respect to animals, they are the deepest lineages within Holozoa (Figure 1B) and have proven especially influential in correctly resolving the origin of transcription factors and cell-signaling molecules [4, 17].

We searched for the presence of the enzymes responsible for miRNA processing and function in ten unicellular holozoan

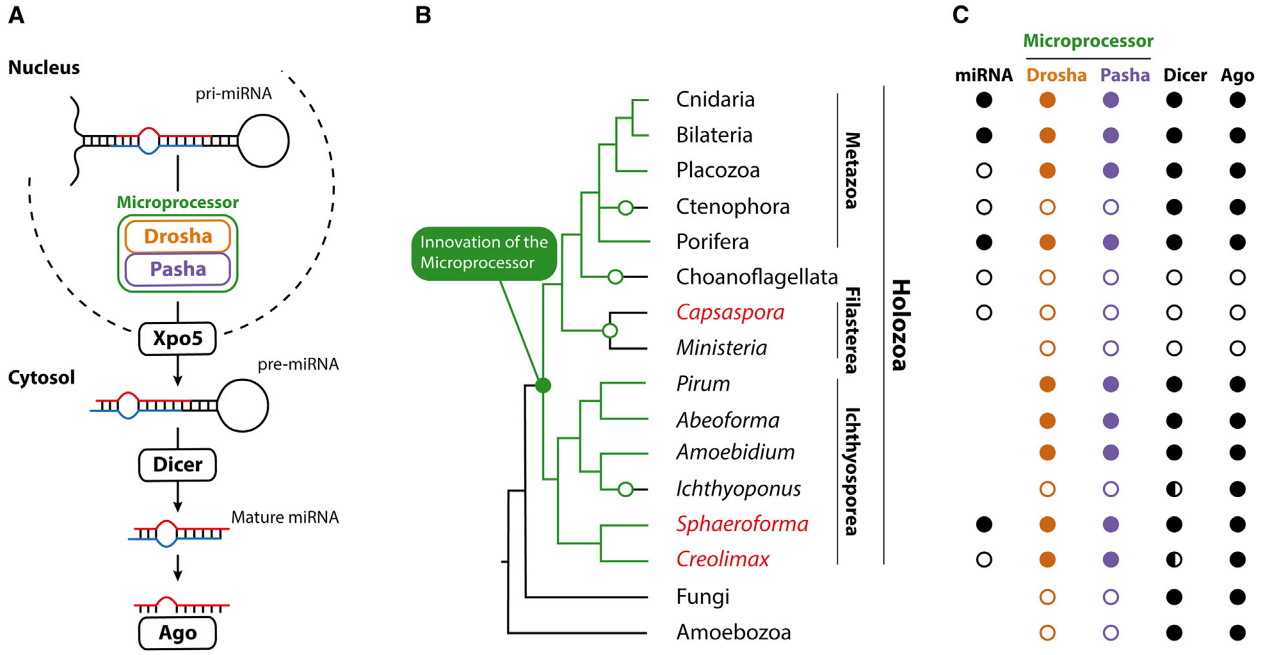


Figure 1. The Evolution of the Animal miRNA Biogenesis Pathway across Holozoa

(A) Schematic drawing of the canonical miRNA pathway in animals. Key proteins are indicated inside rectangles.

(B) Phylogenetic tree of Holozoa with Fungi and Amoebozoa as outgroups. Green branches on the tree indicate the hypothesized origin and evolutionary trajectory of the Microprocessor components (*Drosha* and *Pasha*), and black branches indicate the absence of Microprocessor components. Open circles indicate loss of both Microprocessor components. Taxa highlighted in red have been sequenced for small RNAs in this study.

(C) Presence of miRNAs and genes involved in miRNA biogenesis and function are indicated by filled circles, and absence is indicated by empty circles. For *Dicer*, filled circles means that two or more *Dicers* were discovered, and half-filled circles means a single *Dicer* was identified. Taxa with no circles for miRNAs indicate that small RNAs have not been sequenced.

See also Tables S2 and S3.

species; two filastereans (*Capsaspora owczarzaki* and *Ministeria vibrans*) and eight ichthyosporeans (*Abeoforma whisleri*, *Amoebidium parasiticum*, *Creolimax fragrantissima*, *Ichthyophonus hoferi*, *Pirum gemmata*, *Sphaeroforma arctica*, *S. sirkka*, and *S. napiecek*). In addition, we searched for expressed miRNAs in *C. owczarzaki*, *C. fragrantissima*, *S. arctica*, *S. sirkka*, and *S. napiecek* by small RNA sequencing.

The proteins Drosha (class 3 RNase III protein) and Pasha, which cleave newly transcribed RNA hairpins inside the nucleus (Figure 1A) [18–20], are unique to animal miRNA biogenesis. Export of these miRNAs from the nucleus to the cytoplasm is mediated by the protein Exportin 5 (Xpo5) [18], followed by a second cleavage of the miRNA hairpin by the Dicer protein, another RNase III protein (class 4) [18]. After processing by RNases, miRNAs interface with the proteins of the Argonaute (Ago) family to affect mRNA translation and stability [21]. In plants, which lack both Drosha and Pasha, the entire processing of the RNA hairpins is performed by Dicer before the mature miRNA interacts with Ago [22].

We searched for these genes in transcriptomes of deeply branching holozoan taxa using reciprocal BLAST against animal genomes, BLAST against public databases, and domain annotation (including protein structure analysis). With these approaches, we were able to identify genes similar to *Ago*, *Xpo5*, *Pasha*, and several different RNases, including orthologs of both *Drosha* and *Dicer* in several ichthyosporean species across different genera (Figures 1C and 2; Table S2). The *Dicer* and *Drosha* genes contained two consecutive RNase III domains (i.e., RNase III-A and RNase III-B), which is the defining criterion for these two gene families [25]. Another diagnostic character we identified in the ichthyosporean *Drosha* genes was a unique insert in the RNase III-A, which forms the so-called "bump helix" [25]. Modeling the tertiary structure of these *Drosha* and *Dicer* gene sequences based on homologs with a known 3D structure consistently placed the insert and the bump helix of the ichthyosporean Drosha as in the folded human protein homolog (Figures 3A and S1), while these features were not present in the *Dicer* genes. Congruent with the structural data, all the double-RNase III-containing genes with the insertion and bump helix formed a clade in the phylogenetic analyses, excluding the genes annotated as *Dicer* (Figure 3B; the topology was also recovered independent of the inclusion of the bump helix insertion in the phylogenetic analysis). Hence, all data inferences, covering reciprocal BLAST, domain annotation, and phylogenetic analyses, strongly suggest two types of double-RNase III-containing genes in ichthyosporeans, where one is an ortholog of the Drosha component of the animal Microprocessor complex [20, 25].

The other Microprocessor gene, *Pasha,* was also identified in Ichthyosporea with largely the same domain composition as that of the human homolog, including two consecutive double-stranded RNA-binding domains (dsRBDs; Figures 2 and 3C). For *P. gemmata, A. whisleri*, and *A. parasiticum,* we also

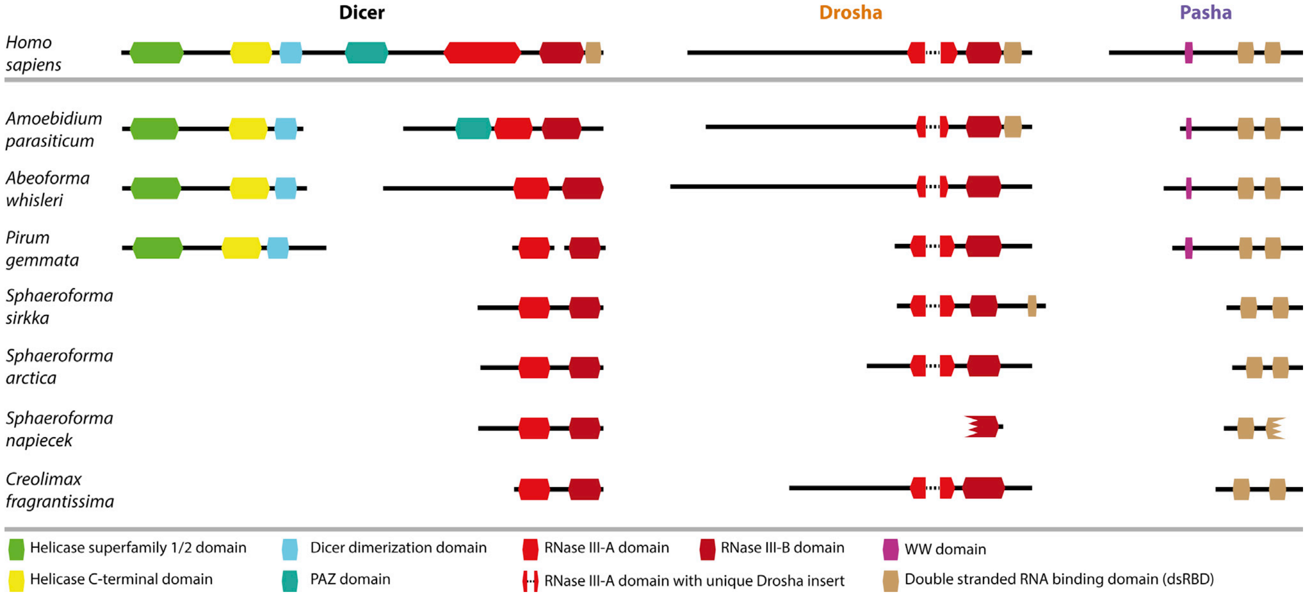


Figure 2. Comparison between the Domain Composition of the Human and Ichthyosporean miRNA Biogenesis Machinery

The domain composition of the ichthyosporean *Dicer*, *Drosha*, and *Pasha* sequences discovered in the reciprocal BLAST searches was compared against their human counterparts (*Dicer* [DICER1; Q9UPY3], *Drosha* [Q9NRR4], and *Pasha* [DGCR8; Q8WYQ5]), as annotated in InterPro [23]. The sequences identified in the reciprocal BLAST searches were annotated using InterProScan5 [23] and CD-Search [24] and by comparing sequence alignments and secondary structures (see STAR Methods). All domains were identified by both InterProScan and CD-Search annotation programs except the following: Dicer dimerization domain of *A. parasiticum*; WW domains of *A. parasiticum*, *A. whisleri*, and *P. gemmata*; dsRBD domains of *S. sirkka* Drosha and of Pasha in *P. gemmata*, *S. arctica*, and *C. fragrantissima* (C-terminal domain only), which were identified by CD-Search only; and the N-terminal dsRBD domains of Pasha in *P. gemmata* and *C. fragrantissima,* which were identified by InterProScan only. The RNase III-A domains of *S. sirkka*, *S. arctica*, and *S. napiecek* Dicer were identified using an alignment and structural modeling approach as described in STAR Methods. In addition, the single RNaseIII domain of *S. napiecek* Drosha was only identified by InterProScan. Incomplete domains are indicated by a jagged border. All boxes and lines are drawn to scale according to their InterProScan annotation (in cases where InterProScan did not identify a domain, the size was chosen based on the homologous domain from a closely related sequence). Except for *C. fragrantissima*, all genes are from *de novo* assembled transcriptome data; hence, the many short contigs and aberrant domains are likely due to incomplete assemblies.

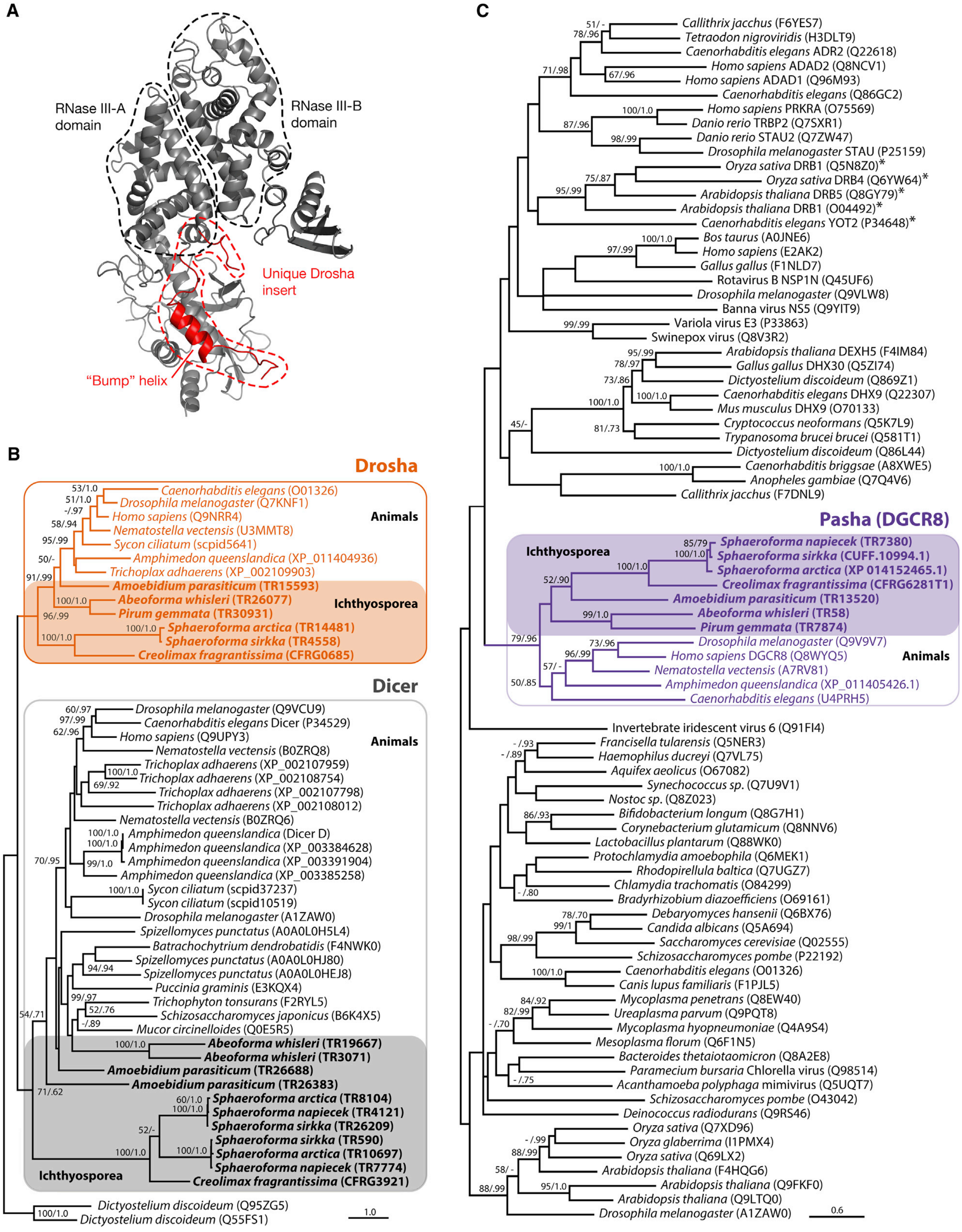
See also Figure S1.

identified a WW domain upstream of the dsRBDs, thereby displaying the full complement of human Pasha domains. Phylogenetic analysis confirmed the annotation of *Pasha* by placing the ichthyosporean genes as sister to animal *Pasha* within a tree composed of all dsRBD-containing sequences in the Pfam database [27] (Figure 3C). This annotation was further strengthened by giving animal *Pasha* as the most significant hit against the NCBI RefSeq, nr, and UniProt databases. The template-based modeling approach also identified *Pasha* as the most similar tertiary model to these sequences. The ichthyosporean *Pasha* did not cluster together with HYL1, which is a partner of Dicer in plants and has been identified in ctenophores, sponges, and cnidarians, but not bilaterians [28]. This suggests that HYL1 has been lost both in Bilateria and in Ichthyosporea.

In contrast, searches for these animal miRNA processing genes in the other holozoan lineages, Filasterea and Choanoflagellata, as well as in all available data from fungi and unicellular relatives (i.e., Holomycota), did not recover any strong candidates for Microprosessor genes (Figure 1C; Table S2).

Altogether, these data contradict earlier hypotheses that *Drosha* and *Pasha* are animal innovations [12, 25]. Rather, our results show that the entire Microprocessor complex originated long before animals, preceding even the last ancestor shared with their nearest unicellular holozoan relatives (Figure 1B). Furthermore, the phylogenies of *Drosha* and *Pasha* resolve animal and Ichthyosporea orthologs in monophyletic groups, suggesting that each of these genes originated once from a common precursor. Lack of *Drosha* and *Pasha* among Holomycota (fungi and their unicellular relatives) suggests that invention of *Drosha* from a *Dicer* precursor [12, 25] occurred early in holozoan evolution. An even earlier origin pre-dating Opisthokonta (i.e., Holozoa plus Holomycota) is possible but requires subsequent losses of *Drosha* and *Pasha* among Holomycota. Such a pre-holozoan origin would require the presence of the Microprocessor proteins among other eukaryote lineages, but so far, only the distantly related green alga *Chlamydomonas reinhardtii* has been reported to have an RNase III gene with possible *Drosha*-like functions (but no *Pasha*) [29].

In any case, the presence of homologous Microprocessor components in Ichthyosporea and animals suggests independent losses of *Drosha* and *Pasha* in choanoflagellates [6] and filastereans (Figures 1B and 1C; Table S2), as well as the only animal lineage that lacks these genes, the ctenophores [12–14] (Placozoa has long been thought to lack the Microprocessor because of the absence of *Pasha* in *Trichoplax adhaerens* [6], but this gene was recently discovered in the strain *Trichoplax* sp. H2 [30]). Absence of the Microprocessor complex in


A
RNase III-A domain
RNase III-B domain
Unique Drosha insert
"Bump" helix
B
Drosha
53/1.0
Caenorhabditis elegans (O01326)
51/1.0
Drosophila melanogaster (Q7KNF1)
-/.97
Homo sapiens (Q9NRR4)
58/.94
Nematostella vectensis (U3MMT8)
95/.99
Sycon ciliatum (scpid5641)
50/-
Amphimedon queenslandica (XP_011404936)
91/.99
Trichoplax adhaerens (XP_002109903)
Animals
Amoebidium parasiticum (TR15593)
100/1.0
Abeoforma whisleri (TR26077)
96/.99
Pirum gemmata (TR30931)
Ichthyosporea
100/1.0
Sphaeroforma arctica (TR14481)
100/1.0
Sphaeroforma sirkka (TR4558)
Creolimax fragrantissima (CFRG0685)
Dicer
60/.97
Drosophila melanogaster (Q9VCU9)
97/.99
Caenorhabditis elegans Dicer (P34529)
62/.96
Homo sapiens (Q9UPY3)
Animals
Nematostella vectensis (B0ZRQ8)
Trichoplax adhaerens (XP_002107959)
100/1.0
Trichoplax adhaerens (XP_002108754)
69/.92
Trichoplax adhaerens (XP_002107798)
Trichoplax adhaerens (XP_002108012)
Nematostella vectensis (B0ZRQ6)
100/1.0
Amphimedon queenslandica (Dicer D)
100/1.0
Amphimedon queenslandica (XP_003384628)
70/.95
99/1.0
Amphimedon queenslandica (XP_003391904)
Amphimedon queenslandica (XP_003385258)
100/1.0
Sycon ciliatum (scpid37237)
Sycon ciliatum (scpid10519)
Drosophila melanogaster (A1ZAW0)
Spizellomyces punctatus (A0A0L0H5L4)
Batrachochytrium dendrobatidis (F4NWK0)
94/.94
Spizellomyces punctatus (A0A0L0HJ80)
Spizellomyces punctatus (A0A0L0HEJ8)
Puccinia graminis (E3KQX4)
99/.97
Trichophyton tonsurans (F2RYL5)
52/.76
Schizosaccharomyces japonicus (B6K4X5)
54/.71
-/.89
Mucor circinelloides (Q0E5R5)
100/1.0
Abeoforma whisleri (TR19667)
Abeoforma whisleri (TR3071)
Amoebidium parasiticum (TR26688)
71/.62
Amoebidium parasiticum (TR26383)
60/1.0
Sphaeroforma arctica (TR8104)
100/1.0
Sphaeroforma napiecek (TR4121)
52/-
Sphaeroforma sirkka (TR26209)
Sphaeroforma sirkka (TR590)
100/1.0
100/1.0
Sphaeroforma arctica (TR10697)
Sphaeroforma napiecek (TR7774)
Ichthyosporea
Creolimax fragrantissima (CFRG3921)
100/1.0
Dictyostelium discoideum (Q95ZG5)
Dictyostelium discoideum (Q55FS1)
1.0
C
51/-
Callithrix jacchus (F6YES7)
78/.96
Tetraodon nigroviridis (H3DLT9)
Caenorhabditis elegans ADR2 (Q22618)
71/.98
Homo sapiens ADAD2 (Q8NCV1)
67/.96
Homo sapiens ADAD1 (Q96M93)
Caenorhabditis elegans (Q86GC2)
100/1.0
Homo sapiens PRKRA (O75569)
87/.96
Danio rerio TRBP2 (Q7SXR1)
98/.99
Danio rerio STAU2 (Q7ZW47)
Drosophila melanogaster STAU (P25159)
Oryza sativa DRB1 (Q5N8Z0)*
75/.87
Oryza sativa DRB4 (Q6YW64)*
95/.99
Arabidopsis thaliana DRB5 (Q8GY79)*
Arabidopsis thaliana DRB1 (O04492)*
Caenorhabditis elegans YOT2 (P34648)*
100/1.0
Bos taurus (A0JNE6)
97/.99
Homo sapiens (E2AK2)
Gallus gallus (F1NLD7)
Rotavirus B NSP1N (Q45UF6)
Drosophila melanogaster (Q9VLW8)
Banna virus NS5 (Q9YIT9)
99/.99
Variola virus E3 (P33863)
Swinepox virus (Q8V3R2)
95/.99
Arabidopsis thaliana DEXH5 (F4IM84)
78/.97
Gallus gallus DHX30 (Q5ZI74)
73/.86
Dictyostelium discoideum (Q869Z1)
100/1.0
100/1.0
Caenorhabditis elegans DHX9 (Q22307)
Mus musculus DHX9 (O70133)
45/-
81/.73
Cryptococcus neoformans (Q5K7L9)
Trypanosoma brucei brucei (Q581T1)
Dictyostelium discoideum (Q86L44)
100/1.0
Caenorhabditis briggsae (A8XWE5)
Anopheles gambiae (Q7Q4V6)
Callithrix jacchus (F7DNL9)
Pasha (DGCR8)
Ichthyosporea
85/79
Sphaeroforma napiecek (TR7380)
100/1.0
Sphaeroforma sirkka (CUFF.10994.1)
100/1.0
Sphaeroforma arctica (XP 014152465.1)
52/.90
Creolimax fragrantissima (CFRG6281T1)
Amoebidium parasiticum (TR13520)
99/1.0
Abeoforma whisleri (TR58)
Pirum gemmata (TR7874)
79/.96
73/.96
Drosophila melanogaster (Q9V9V7)
96/.99
Homo sapiens DGCR8 (Q8WYQ5)
Animals
57/-
Nematostella vectensis (A7RV81)
50/.85
Amphimedon queenslandica (XP_011405426.1)
Caenorhabditis elegans (U4PRH5)
Invertebrate iridescent virus 6 (Q91FI4)
-/.93
Francisella tularensis (Q5NER3)
-/.89
Haemophilus ducreyi (Q7VL75)
Aquifex aeolicus (O67082)
Synechococcus sp. (Q7U9V1)
Nostoc sp. (Q8Z023)
86/.93
Bifidobacterium longum (Q8G7H1)
Corynebacterium glutamicum (Q8NNV6)
Lactobacillus plantarum (Q88WK0)
Protochlamydia amoebophila (Q6MEK1)
Rhodopirellula baltica (Q7UGZ7)
Chlamydia trachomatis (O84299)
-/.80
Bradyrhizobium diazoefficiens (O69161)
78/.70
Debaryomyces hansenii (Q6BX76)
99/1
Candida albicans (Q5A694)
98/.99
Saccharomyces cerevisiae (Q02555)
Schizosaccharomyces pombe (P22192)
100/1.0
Caenorhabditis elegans (O01326)
Canis lupus familiaris (F1PJL5)
84/.92
Mycoplasma penetrans (Q8EW40)
82/.99
Ureaplasma parvum (Q9PQT8)
-/.70
Mycoplasma hyopneumoniae (Q4A9S4)
Mesoplasma florum (Q6F1N5)
Bacteroides thetaiotaomicron (Q8A2E8)
Paramecium bursaria Chlorella virus (Q98514)
-/.75
Acanthamoeba polyphaga mimivirus (Q5UQT7)
Schizosaccharomyces pombe (O43042)
Deinococcus radiodurans (Q9RS46)
Oryza sativa (Q7XD96)
-/.99
Oryza glaberrima (I1PMX4)
88/.99
Oryza sativa (Q69LX2)
58/-
Arabidopsis thaliana (F4HQG6)
88/.99
95/1.0
Arabidopsis thaliana (Q9FKF0)
Arabidopsis thaliana (Q9LTQ0)
Drosophila melanogaster (A1ZAW0)
0.6

(legend on next page)

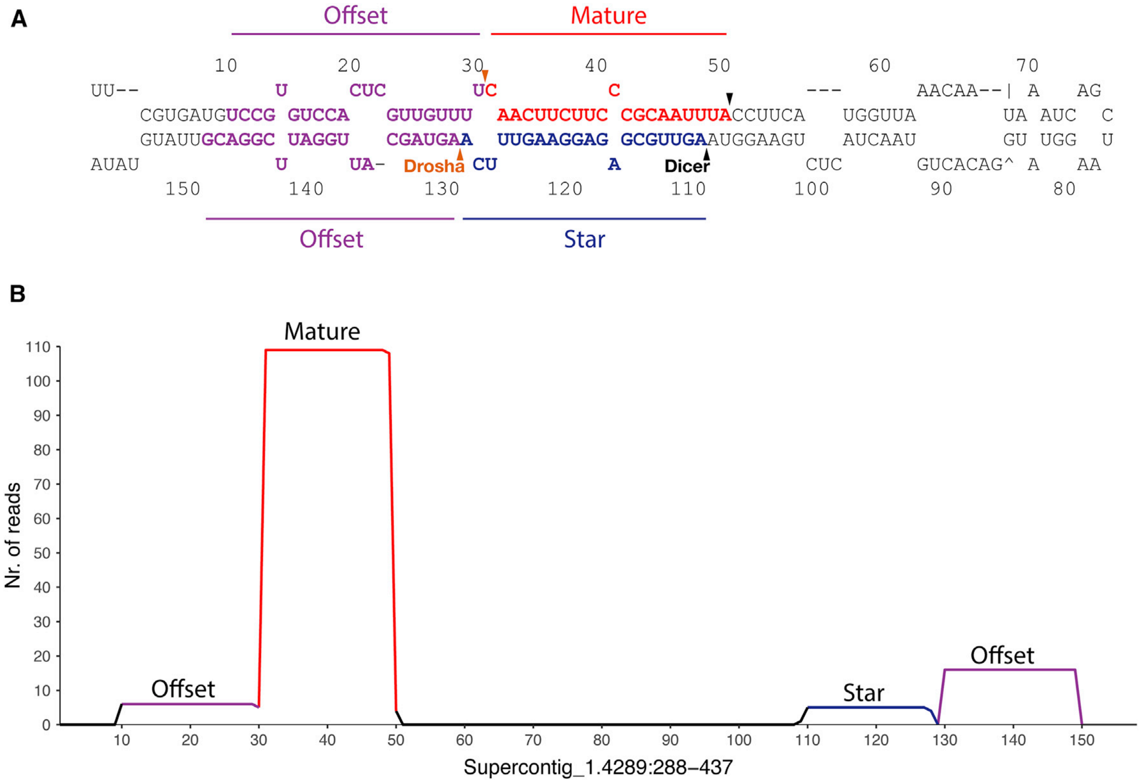


Figure 4. Secondary Structure and Small RNA Coverage of a Novel Ichthyosporean miRNA

(A) The secondary structure of the novel miRNA Sar-Mir-Nov-1 identified in *Sphaeroforma arctica* with the likely Drosha and Dicer cut sites indicated. Mature and star strands are indicated in red and blue, respectively, and magenta indicates the presence of offset reads resulting from the Drosha cuts.

(B) The mapping of small RNA reads on the genomic location of Sar-Mir-Nov-1. The numbers on the x axis correspond to the numbers in the secondary structure in (A). Note the presence of offset reads (external reads mapping outside the pre-miRNA) that are in accordance with Drosha processing. See Figure S2 for more miRNA structures.

See also Figures S2–S4 and Table S1.

ctenophores must, therefore, be derived and not a primitive state as previously suggested [12].

In animals, the main function of the Microprocessor is to process the primary miRNA transcripts, but miRNA genes have not been reported from deeply diverging Holozoa. It is, therefore, uncertain whether the ichthyosporean Microprocessor components identified here have the same function as in animals. Thus, we explored the presence of miRNAs using a combination of deep sequencing of small RNAs (Table S1) with computational searches of the genomes of our species. Eight miRNAs were identified in three species of the genus *Sphaeroforma* (Figures 4 and S2; Data S1). These fulfilled the criteria for the annotation of miRNA genes and were all expressed in two 20- to 26-nt cRNA strands from a hairpin precursor with a 2-nt offset, reflecting the sequential activity of two RNase III enzymes (Drosha and Dicer) [31, 32]. All eight of these miRNA genes were highly conserved across two of the three species of *Sphaeroforma*, with six of them conserved across all three (Data S1), supporting their

Figure 3. Identification of Ichthyosporean *Drosha* and *Pasha* Sequences

(A) The modeled protein structure of the *Drosha* homolog identified in the ichthyosporean *Abeoforma whisleri*. Indicated in red is the unique *Drosha* insertion, including the so-called ‘‘Bump’’ helix [25]. Modeled structures of other identified ichthyosporean *Drosha* genes are shown in Figure S1.

(B) Phylogeny of *Dicer* and *Drosha* sequences. *Drosha* sequences are indicated in the orange box; all other sequences are *Dicer*. The topology with the highest likelihood in a maximum-likelihood (ML) framework is shown, with ML bootstrap and Bayesian posterior probability (BP) nodal support values drawn onto the branching points (ML/BP). Only support values above 50% ML and/or 0.75 BP are shown. Accession numbers are given in parentheses. For all taxa, accession numbers refer to the UniProt database, except for *Trichoplax adhaerens* and *Amphimedon queenslandica*, which are from NCBI RefSeq (the *A. queenslandica* Dicer D sequence is taken from [26]), and *Sycon ciliatum*, which is from http://www.compagen.org. Ichthyosporean species are indicated in bold font. A *Drosha* ortholog was also detected in *S. napiecek*, but this sequence was incompletely assembled and did not cover the RNase III domains and was, therefore, not included in the analysis.

(C) Ichthyosporean sequences identified as *Pasha* in the reciprocal BLAST searches (bold font) analyzed together with double-stranded RNA binding motif (DSRM)-containing sequences from the Pfam database (see STAR Methods for details). All *Pasha* sequences are indicated in a purple box. HYL1 homologs are marked with an asterisk. UniProt accession numbers are given in parentheses (except for *Amphimedon queenslandica*, for which the NCBI RefSeq accession number is given). Tree topology and support values were created in the same way as for the phylogeny in (B).

See also Figure S1 and Table S3.

identification as functional miRNAs [31, 32]. In addition to conserved genomic sequences of these miRNAs, their expression and subsequent processing were also highly conserved between the different species. For species of *Sphaeroforma* with available genomic data, we were able to establish that the miRNAs are located either in intergenic regions or in the introns and UTRs of protein-coding genes. Two of the miRNAs were consistently located within *Ago* and *Dicer* (Figure S3; Data S1). Such genomic co-localization of miRNAs and miRNA processing genes is not found in animals and likely reflects additional instances of the exaptation of the primitive intronic sequence into miRNA genes [33]. None of the miRNA genes have homologs outside Ichthyosporea.

Altogether, the conserved sequence features and genome localization across species are suggestive of functional miRNA genes that are processed by an enzymatic machinery similar to that in animals. This functional link between the Microprocessor and miRNA genes is further strengthened by the co-occurrence of these two components in all holozoan lineages investigated so far. *C. fragrantissima* is the only species deviating from this pattern; it contains homologs of the Microprocessor but apparently no miRNA genes. Although, it could be possible that miRNAs were not detected in *C. fragrantissima* because their expression is restricted to certain developmental time points not present under our culture conditions. The existence of such stages have been suggested for closely related *Sphaeroforma* species [34] and could as well exist in *C. fragrantissima*. Drosha has also been found to cleave other types of secondary RNA stem-loop structures in mouse cell lines [35], which could represent an alternative function for the Drosha homolog in *C. fragrantissima*. In any case, the role of the Microprocessor and miRNAs in Ichthyosporea needs to be confirmed by functional studies, but this is currently not possible due to lack of developed protocols and an experimental system.

A deep holozoan origin of both miRNAs and the biogenesis machinery confirms that the genetic innovations that underpin miRNA biogenesis in animals are not linked phylogenetically with the origin of animal multicellularity itself [36, 37]. Rather, our findings complement the view that the unicellular ancestor of animals already had most of the genes, gene pathways, and regulatory mechanisms necessary, but evidently insufficient, for animal-grade multicellularity [11]. This repertoire includes genes involved in cell adhesion and communication, extra- and intra-cellular receptors, and transcription factors previously thought to be specific to animals; e.g., [1, 5, 38]. Beyond genes, this unicellular ancestor of animals also had other genomic regulatory mechanisms, including regulation of chromatin states, complex *cis*-regulation by enhancers, and cell-type-specific alternative splicing [4, 17]. We add post-transcriptional regulation of mRNA translation via miRNAs to this gene regulatory repertoire. It remains unclear whether the Microprocessor in Ichthyosporea functions as it does in animals, by targeting mRNAs and buffering noise in gene expression [39]. If this is not the case, the miRNA regulatory pathway was co-opted early in animal evolution for these purposes from an as-yet-unknown ancestral function. Nonetheless, our findings provide further support for the notion that many developmental features key to the emergence of animal multicellularity and phenotypic complexity evolved deep within the unicellular ancestry of animals before being co-opted and/or further expanded within multicellular Metazoa.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Identification of genes related to the miRNA processing machinery
  - ○ Phylogenetic annotation of miRNA processing proteins
  - ○ Culturing and RNA sequencing
  - ○ Mapping of RNA reads and miRNA detection
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Phylogenetic analyses
  - ○ Blast searches
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, three tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.cub.2018.08.018.

ACKNOWLEDGMENTS

We are grateful to Brandon Hassett for providing the *S. sirkka* and *S. napiecek* cultures, and we thank Notur (https://www.sigma2.no) and USIT at University of Oslo for providing computational resources and development of www.bioportal.no. B.F. is supported by South-Eastern Norway Regional Health Authority grant 2014041. H.S. was supported by JSPS KAKENHI 16K07468. P.C.J.D. is supported by the Natural Environment Research Council (NE/P013678/1). I.R.-T. acknowledges supported by an ERC Consolidator grant (ERC-2012-Co-616960), support from the Secretary's Office for Universities and Research of the Generalitat de Catalunya (project 2014 SGR 619), and a grant from the Spanish Ministry for Economy and Competitiveness (BFU2017-90114-P), the latter with European Regional Development Fund support. The postdoc grants (Nr. 213703 and 240284) to J.B. was funded by the Norwegian Research Council. Funding of the research project (including PhD fellowship for R.S.N.) and the www.bioportal.no infrastructure was granted to K.S.-T. by the Molecular Life Science board at University of Oslo.

AUTHOR CONTRIBUTIONS

J.B. participated in the study design, took part in all the data analyses, designed the figures, and drafted the manuscript. R.S.N. participated in the study design, cultured and isolated RNA from *S. arctica*, analyzed the *S. arctica* small RNAs and the miRNA pathway genes, and wrote the initial manuscript draft. B.F. analyzed the small RNA data, identified and annotated miRNAs, provided critical evaluation of the miRNA structures, participated in figure design, and commented on the manuscript. A.A.B.H. maintained the cultures and isolated mRNA and total RNA, assembled novel transcriptomes, analyzed the small RNAs, developed the reciprocal BLAST pipeline, ran phylogenetic analyses, and commented on the manuscript. J.E.T. prepared small RNA libraries, analyzed the small RNA data, and commented on the manuscript. H.S. cultured *S. arctica*, *C. owczarzaki*, and *C. fragrantissima*; was involved in the analyses of the genetic machinery; and commented on the manuscript. P.C.J.D. prepared small RNA libraries, took part in the small RNA sequencing, and contributed to the manuscript. K.J.P. analyzed the small RNA data, identified and annotated miRNAs, provided critical evaluation of the miRNA structures, participated in figure design, and contributed to the manuscript. I.R.-T.

provided culture material, was involved in the analyses of the genetic machinery, and contributed to the manuscript. P.E.G. participated in the study design, provided critical discussion on miRNA function, and commented on the manuscript. K.S.-T. participated in the study design, evaluated all the data analyses and figures, and contributed on the initial and final manuscripts. All authors have read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 27, 2018
Revised: July 22, 2018
Accepted: August 7, 2018
Published: October 11, 2018

REFERENCES

1. Shalchian-Tabrizi, K., Minge, M.A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K.S., and Cavalier-Smith, T. (2008). Multigene phylogeny of choanozoa and the origin of animals. PLoS ONE *3*, e2098.
2. Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M.W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., et al. (2013). The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat. Commun. *4*, 2325.
3. King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., et al. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature *451*, 783–788.
4. de Mendoza, A., Suga, H., Permanyer, J., Irimia, M., and Ruiz-Trillo, I. (2015). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. eLife *4*, e08904.
5. Sebé-Pedrós, A., Roger, A.J., Lang, F.B., King, N., and Ruiz-Trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc. Natl. Acad. Sci. USA *107*, 10142–10147.
6. Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193–1197.
7. Peterson, K.J., Dietrich, M.R., and McPeek, M.A. (2009). MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. BioEssays *31*, 736–747.
8. Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., and Peterson, K.J. (2009). The deep evolution of metazoan microRNAs. Evol. Dev. *11*, 50–68.
9. Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. Nat. Rev. Genet. *12*, 846–860.
10. Sebé-Pedrós, A., and Ruiz-Trillo, I. (2017). Evolution and classification of the T-box transcription factor family. Curr. Top. Dev. Biol. *122*, 1–26.
11. Gaiti, F., Calcino, A.D., Tanurdžić, M., and Degnan, B.M. (2016). Origin and evolution of the metazoan non-coding regulatory genome. Dev. Biol. *35*, 76–83.
12. Maxwell, E.K., Ryan, J.F., Schnitzler, C.E., Browne, W.E., and Baxevanis, A.D. (2012). MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore *Mnemiopsis leidyi*. BMC Genomics *13*, 714.
13. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.
14. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.-D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Smith, S.A., et al.; NISC Comparative Sequencing Program (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science *342*, 1242592.
15. Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20–51.
16. O'Malley, M.A., Wideman, J.G., and Ruiz-Trillo, I. (2016). Losing complexity: the role of simplification in macroevolution. Trends Ecol. Evol. *31*, 608–621.
17. Sebé-Pedrós, A., Peña, M.I., Capella-Gutiérrez, S., Antó, M., Gabaldón, T., Ruiz-Trillo, I., and Sabidó, E. (2016). High-throughput proteomics reveals the unicellular roots of animal phosphosignaling and cell differentiation. Dev. Cell *39*, 186–197.
18. Kim, Y.-K., Kim, B., and Kim, V.N. (2016). Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. Proc. Natl. Acad. Sci. USA *113*, E1881–E1889.
19. Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat. Rev. Mol. Cell Biol. *10*, 126–139.
20. Nguyen, T.A., Jo, M.H., Choi, Y.G., Park, J., Kwon, S.C., Hohng, S., Kim, V.N., and Woo, J.S. (2015). Functional anatomy of the human Microprocessor. Cell *161*, 1374–1387.
21. Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608–613.
22. Moran, Y., Agron, M., Praher, D., and Technau, U. (2017). The evolutionary origin of plant and animal microRNAs. Nat. Ecol. Evol. *1*, 27.
23. Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. *45* (D1), D190–D199.
24. Marchler-Bauer, A., and Bryant, S.H. (2004). CD-Search: protein domain annotations on the fly. Nucleic Acids Res. *32*, W327–W331.
25. Kwon, S.C., Nguyen, T.A., Choi, Y.-G., Jo, M.H., Hohng, S., Kim, V.N., and Woo, J.-S. (2016). Structure of Human DROSHA. Cell *164*, 81–90.
26. Mukherjee, K., Campos, H., and Kolaczkowski, B. (2013). Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. Mol. Biol. Evol. *30*, 627–641.
27. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44* (D1), D279–D285.
28. Moran, Y., Praher, D., Fredman, D., and Technau, U. (2013). The evolution of microRNA pathway protein components in Cnidaria. Mol. Biol. Evol. *30*, 2541–2552.
29. Valli, A.A., Santos, B.A.C.M., Hnatova, S., Bassett, A.R., Molnar, A., Chung, B.Y., and Baulcombe, D.C. (2016). Most microRNAs in the single-cell alga *Chlamydomonas reinhardtii* are produced by Dicer-like 3-mediated cleavage of introns and untranslated regions of coding RNAs. Genome Res. *26*, 519–529.
30. Kamm, K., Osigus, H.-J., Stadler, P.F., DeSalle, R., and Schierwater, B. (2018). Trichoplax genomes reveal profound admixture and suggest stable wild populations without bisexual reproduction. Sci. Rep. *8*, 11168.
31. Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microRNA annotation. RNA *9*, 277–279.
32. Fromm, B., Billipp, T., Peck, L.E., Johansen, M., Tarver, J.E., King, B.L., Newcomb, J.M., Sempere, L.F., Flatmark, K., Hovig, E., and Peterson, K.J. (2015). A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. Annu. Rev. Genet. *49*, 213–242.
33. Campo-Paysaa, F., Sémon, M., Cameron, R.A., Peterson, K.J., and Schubert, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. Evol. Dev. *13*, 15–27.
34. Hassett, B.T., López, J.A., and Gradinger, R. (2015). Two new species of marine saprotrophic sphaeroformids in the Mesomycetozoea isolated from the sub-arctic Bering Sea. Protist *166*, 310–322.
35. Chong, M.M.W., Zhang, G., Cheloufi, S., Neubert, T.A., Hannon, G.J., and Littman, D.R. (2010). Canonical and alternate functions of the microRNA biogenesis machinery. Genes Dev. *24*, 1951–1960.

36. Tarver, J.E., Donoghue, P.C.J., and Peterson, K.J. (2012). Do miRNAs have a deep evolutionary history? BioEssays *34*, 857–866.

37. Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K., et al. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science *329*, 223–226.

38. Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G., and Ruiz-Trillo, I. (2012). Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci. Signal. *5*, ra35.

39. Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., and van Oudenaarden, A. (2015). Gene expression. MicroRNA control of protein expression noise. Science *348*, 128–132.

40. Jøstensen, J.-P., Sperstad, S., Johansen, S., and Landfald, B. (2002). Molecular-phylogenetic, structural and biochemical features of a cold-adapted, marine ichthyosporean near the animal-fungal divergence, described from *in vitro* cultures. Eur. J. Protistol. *38*, 93–104.

41. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

42. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

43. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Macmanes, M.D., et al. (2014). De novo transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity. Nat. Protoc. *8*, 1–43.

44. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*, 46–53.

45. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

46. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

47. Katoh, K., and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics *26*, 1899–1900.

48. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

49. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.

50. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

51. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

52. Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664.

53. Shi, H., Tschudi, C., and Ullu, E. (2006). An unusual Dicer-like1 protein fuels the RNA interference pathway in *Trypanosoma brucei*. RNA *12*, 2063–2072.

54. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Please cite this article in press as: Bråte et al., Unicellular Origin of the Animal MicroRNA Machinery, Current Biology (2018), https://doi.org/10.1016/j.cub.2018.08.018

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Marine Broth	Difco	Cat# 279110
Trizol	Life-Technologies	Cat# 15596026
Illumina Truseq small RNA seq kit	Illumina	NA
mirPremier microRNA Isolation Kit	Sigma-Aldrich	SNC50
Terminator 5′-Phosphate-Dependent Exonuclease	Epicenter	NA
Tobacco Acid Pyrophosphatase	Epicenter	T19050
Deposited Data		
Unprocessed small RNA and mRNA reads, and novel gene sequences used in this study.	This paper	ENA: PRJEB21207
Experimental Models: Organisms/Strains		
Sphaeroforma arctica	Iñaki Ruiz-Trillo's lab. Original reference [40]	Strain JP610
Sphaeroforma sirkka	Brandon Hassett [34]	Strain B5
Sphaeroforma napiecek	Brandon Hassett [34]	Strain B4
Capsaspora owczarzaki	ATCC nr. 30864	N/A
Creolimax fragrantissima	Iñaki Ruiz-Trillo's lab (available from ATCC nr. PRA-284)	N/A
Software and Algorithms		
Trimmomatic v0.35	[41]	http://www.usadellab.org/cms/?page=trimmomatic
Trinity v2.0.6	[42]	http://trinityrnaseq.github.io/
Transdecoder v3.0.0	[43]	http://transdecoder.github.io/
Cufflinks v2.1.1	[44]	http://cole-trapnell-lab.github.io/cufflinks/
Blastp	[45]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
InterProScan	[23]	https://www.ebi.ac.uk/interpro/interproscan.html
CD-search	[24]	https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?
Geneious R9	[46]	https://www.geneious.com/
Mafft v.7	[47]	https://mafft.cbrc.jp/alignment/software/
Phyre2 web server	[48]	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
PhyloBayes-MPI v1.5	[49]	http://megasun.bch.umontreal.ca/People/lartillot/www/old/
RAxML v8.0.26	[50]	https://sco.h-its.org/exelixis/web/software/raxml/index.html
TopHat v2.0.14	[51]	https://ccb.jhu.edu/software/tophat/index.shtml
Blat v3.5	[52]	https://genome.ucsc.edu/FAQ/FAQblat
Other		
Acropora digitifera genome assembly	NCBI Genome	Adig_1.1. ID: 10529
Nematostella vectensis genome assembly	NCBI Genome	ASM20922v1. ID: 230
Trichoplax adhaerens genome assembly	NCBI Genome	v1.0. ID: 354
Amphimedon queenslandica genome assembly	NCBI Genome	v1.0. ID: 2698

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Sycon ciliatum genome assembly	http://www.compagen.org	SCIL_WGA_130802
Mnemiopsis leidyi genome assembly	NHGRI	https://research.nhgri.nih.gov/mnemiopsis/download/genome/MlScaffold09.nt.gz
Pleurobrachia bachei genome assembly	Neurobase	https://neurobase.rc.ufl.edu
Acanthoeca spectabilis transcriptome data	NCBI SRA	SRX956664
Acanthoeca sp.	Data Commons	N/A
Monosiga brevicollis genome assembly	NCBI Genome	v1.0. ID: 713
Salpingoeca pyxidium transcriptome data	NCBI SRA	SRX956675
Salpingoeca rosetta genome assembly	NCBI Genome	Proterospongia_sp_ATCC50818. ID: 24391
Capsaspora owczarzaki genome and transcriptome assembly	Figshare	v03
Ministeria vibrans transcriptome data	NCBI SRA	SRX096927, SRX096925
Abeoforma whisleri transcriptome data	NCBI SRA	SRX377508
Amoebidium parasiticum transcriptome data	NCBI SRA	SRX179384, SRX096923, SRX096918
Creolimax fragrantissima genome and transcriptome assembly	Figshare	https://figshare.com/articles/Creolimax_fragrantissima_genome_data/1403592
Ichthyophonus hoferi transcriptome data	NCBI SRA	SRX738222
Pirum gemmata transcriptome data	NCBI SRA	SRX377507
Sphaeroforma arctica genome and transcriptome assembly	NCBI Genome, this study	Spha_arctica_JP610_V1. ID: 11004
Sphaerothecum destruens transcriptome data	NCBI SRA	SRX737879
Corallochytrium limacisporum transcriptome data	NCBI SRA	SRX738098, SRX732498
Dictyostelium discoideum genome assembly	NCBI Genome	dicty_2.7. ID: 56
Fonticula alba genome assembly	NCBI Genome	Font_alba_ATCC_38817_V2. ID: 12936
Nuclearia sp. transcriptome data	NCBI SRA	SRX737107
Allomyces macrogynus genome assembly	NCBI Genome	A_macrogynus_V3. ID: 327
Mortierella verticillata genome assembly	NCBI Genome	Mort_vert_NRRL_6337_V1. ID: 801
Rozella allomycis genome assembly	NCBI Genome	Rozella_k41_t100. ID: 12422
Spizellomyces punctatus genome assembly	NCBI Genome	S_punctatus_V1. ID: 344

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kamran Shalchian-Tabrizi (kamran@ibv.uio.no).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sphaeroforma arctica JP610, *S. sirkka* (strain B5), *S. napiecek* (strain B4) and *Creolimax fragrantissima* (CCCM101) were grown on Marine Broth (Difco BD, NJ, US; 37.4g/L) at 12°C and no light. *S. arctica* was also grown on ATCC MAP medium at 16°C with no light. *Capsaspora owczarzaki* (ATCC30864) was cultured on ATCC 803 M7 medium at 23°C with no light.

METHOD DETAILS

Identification of genes related to the miRNA processing machinery

In order to search for the presence of genes involved in miRNA processing and function across the supergroup Opisthokonta (Holozoa (i.e., animals, Choanoflagellata, Filasterea and Ichthyosporea) and Holomycota (i.e., fungi plus their unicellular relatives)) we searched available transcriptomes and proteomes from a wide range of deeply diverging opisthokont species covering basal Holozoa and Holomycota (Table S2). For species from which an assembled transcriptome was not available, raw reads were downloaded from the NCBI SRA database, quality trimmed using Trimmomatic v0.35 [41] (minimum phred score 20-28 depending on read quality) and assembled using Trinity v2.0.6 [42] (with the–normalize_reads option set, otherwise default settings) and Transdecoder v3.0.0 [43] (TransDecoder.LongOrfs program with default settings) for transcriptomes where no reference genome was available and the TopHat v2.1.1 + Cufflinks v2.1.1 [44] pipeline for transcriptomes when a reference genome was available. Genes were identified using three complementary strategies; reciprocal Blast, domain identification and secondary structure analysis:

Reciprocal Blast

As query genes we used Dicer, Drosha, Pasha, Argonaute (Ago) and Exportin 5 (Xpo5) from *Homo sapiens*, *Drosophila melanogaster*, *Nematostella vectensis* and *Amphimedon queenslandica* and Dicer, Ago and Xpo5 from the fungus *Neurospora crassa*. Accession numbers of the query genes are listed in Table S3. Blast was performed by searching the query sequences against each individual target genome/transcriptome using Blastp [45] (BLOSUM45 scoring matrix, min e-value 0.01 and max target hits 30). Each blast hit was then verified by reciprocal blast searches against a database consisting of the genomes and proteomes of the query organisms (i.e., *H. sapiens*, *D. melanogaster*, *N. vectensis, A. queenslandica, S. arctica* and *N. crassa*). All blast hits were sorted by increasing e-value. Only genes ranked as top hit in both reciprocal Blast runs were retained. These hits were further verified by Blast search against the UniProt database (same search parameters as above) and annotated as potential microRNA processing genes only when the UniProt search provided the same gene type match (as the query sequence) as the best hit. Further Blast verification was usually performed against the GenBank nr database.

Protein sequence classification and domain annotation

Genes retrieved as related to the miRNA processing machinery were thereafter classified and annotated by using InterProScan [23], CD-search [24] and sequence comparison with multiple sequence alignments. We defined miRNA-related genes on the basis of the identified domains as follows; ***Ago***: both PAZ and PIWI domains present, ***Dicer*** and ***Drosha***: two RNase III domains present, ***Pasha***: two double stranded RNA-binding domains (dsRBD), ***Xpo5***: contains no conserved domains and was only identified with the reciprocal Blast strategy.

Incompletely assembled gene fragments

A few identified sequences were short and incompletely assembled gene fragments, which made robust identification difficult. For *Pirum gemmata* and *Ichthyophonus hoferi* we could not identify *Dicer* genes with double RNase III domains, but only short sequences containing a single RNase III domain which all gave Blast hits to *Dicer* genes. Likewise, for *S. napiecek* we discovered a *Drosha* homolog with high similarity to the other ichthyosporean *Drosha* sequences and which gave *Drosha* as the best Blast hit, but this was incomplete and did not cover an RNase III domain (Figure 2). All these short or fragmented sequences were not included in the phylogenetic analyses described below. The *Drosha* sequence discovered in *S. arctica* was not fully assembled in the *de novo* transcriptome assembly, but by mapping the mRNAs to the genome we confirmed that the gene was expressed as a single fragment consisting of the genes SARC_08310 and SARC_15010. Likewise, for one of the *Ago* genes in *S. arctica* we also needed to map the mRNAs to the genome to confirm its expression as it was not completely assembled *de novo*. All blast searches and domain annotations were done using Geneious R9 [46], except for the UniProt and GenBank blast searches which were performed on the UniProt and NCBI web sites. Additional domain annotations were also performed using the InterProScan and CD-search web interfaces (https://www.ebi.ac.uk/interpro/ and https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?).

Detecting the RNase III-A domain in *Sphaeroforma* sp. and *C. fragrantissima*

Only two gene families contain double RNase III domains and these comprise the *Drosha* and *Dicer* genes (i.e., class 3 and 4 of the RNase III gene family [25]). For most of the ichthyosporean sequences obtained here the two RNase III domains were identified by conventional approaches described above, but for a few genes from *Sphaeroforma* sp. and *C. fragrantissima* we identified only one of the two RNase III domains located in the C-terminal region (i.e., the B domain). We aligned these sequences to the RNase III-A and B domains of other animal and fungal *Dicers* and *Drosha* proteins, as well as the bacterial *Aquifex aeolicus* RNase III domain. The alignment was done by splitting the sequences into parts consisting of only the RNase III-A or B domain. For sequences without an annotated RNase III domain these putative domains were identified by aligning the sequence to the annotated domains of the *H. sapiens* and *N. vectensis Dicer* and *Drosha* sequences. Then all RNase III-A and B domains were aligned together. All alignments were done using Mafft v.7 [47] with the L-INS-I algorithm with the BLOSUM45 scoring matrix. Aligning the genes to known *Dicer* and *Drosha* genes confirmed that the *Dicers* from *Sphaeroforma* contain a divergent RNase III-A domain, similar to what has been found for other taxa [53], while *C. fragrantissima* lack the same domain.

Tertiary structure analysis

We also used secondary and tertiary structure comparisons of the *Dicer* and *Drosha* candidates to see whether we could identify the other RNase III domain (i.e., the A-domain) and structures unique for Dicer or Drosha. For tertiary structure modeling we used the Phyre2 web server [48] for template based modeling. Phyre2 was run in "Normal" modeling mode to first search for homologous sequences and to create an evolutionary sequence profile to account for variation across sites. The resulting sequence profile was then compared against known tertiary structures and the query sequences were modeled against the best fitting tertiary sequence model. The *Pasha* sequences were also analyzed in this way to test which sequence was identified as the most similar based on structural similarity.

Phylogenetic annotation of miRNA processing proteins

A multiple sequence alignment containing known *Dicer* and *Drosha* sequences from animals, fungi and *Dictyostelium discoideum*, as well as the *Dicer* and *Drosha* sequences of ichthyosporeans identified in this study was generated using Mafft v7.3. First, all full-length *Dicer* and *Drosha* sequences were globally aligned using the E-INS-i algorithm and the BLOSUM45 scoring matrix, then shorter and incomplete sequences were added sequentially using the–addFragments option (all *Drosha* sequences were trimmed from the N-terminal to exclude unannotated regions where no conservation between sequences was detected). Obvious erroneously inserted end gaps (a common problem with Mafft alignments) were either manually realigned or removed. The *Sphaeroforma Dicer* and *Drosha* sequences were manually aligned according to domain annotations. All domains and inter-domain regions were subsequently realigned individually using Mafft L-INS-I algorithm. Finally, alignment columns containing $\geq$ 98% gaps were masked. See Table S3 for list of accession numbers used in the analysis. Bayesian analysis was performed with PhyloBayes-MPI v1.5 [49]. Two chains were run with the parameters -gtr and -cat and stopped when the maxdiff was 0.078 and the meandiff 0.0007 with a 15% burnin. Maximum likelihood (ML) analysis was run using RAxML v8.0.26 [50] with the LG protein substitution model determined by invoking the autoMRE option. The topology with the highest likelihood score out of 10 heuristic searches was selected as the final topology. Bootstrapping was carried out with 950 pseudo replicates under the same model. The values from the ML bootstrapping and the Bayesian posterior probabilities were added to the ML topology with the highest likelihood.

To investigate the evolutionary affiliation of the annotated *Pasha* sequences we created a multiple sequence alignment including full-length seed sequences from the double-stranded RNA binding motif (DSRM) family in the Pfam database (PF00035) [27] (DSRM is equivalent to the dsRBD notation used by InterPro). In addition, we included reference *Pasha* sequences from certain animal lineages. These included *Drosophila melanogaster*, *Nematostella vectensis*, *Caenorhabditis elegans* and *Amphimedon queenslandica*. The *Pasha* and Pfam DSRM containing protein sequences were aligned together with the ichthyosporean *Pasha* candidates with Mafft (L-INS-i algorithm and BLOSUM45 scoring matrix) implemented in Geneious v11.0.3. Further, positions in the alignment containing > 95% gaps were masked. The alignment was analyzed using ML and Bayesian analyses as described above (except that the VT model and 550 pseudo-replicates were used in the ML analysis). In the Bayesian analysis the two chains came close to convergence (burn-in 25%, maxdiff = 0.30, meandiff = 0.014). The values from the ML bootstrapping and the Bayesian posterior probabilities were added to the ML topology with the highest likelihood.

Culturing and RNA sequencing

We first cultured and sequenced small RNAs from *S. arctica* (cultured on Marine Broth), *C. fragrantissima* and *C. owczarzaki*. Total RNA was isolated from all cultures using Trizol (Life Technologies, Carlsbad, CA, USA). Small RNA libraries were prepared using the Illumina Truseq small RNA seq kit (Illumina, San Diega, CA, USA). The samples were run on an GAIIx Illumina sequencer at the University of Bristol Transcriptomics facility with 36 bp single read sample.

In a second round of sequencing we analyzed *S. sirkka* and *S. napiecek* in addition to *S. arctica* (cultured on MAP medium (18.6g/l Difco marine broth 2216, 20 g/l Bacto peptone, 10 g/l NaCl)) and *C. fragrantissima*. Total RNA was isolated by lysing the cells on a FastPrep system (MP Biomedicals, Santa Ana, CA, USA), followed by small RNA and total RNA isolation using the mirPremiere RNA kit (Sigma-Aldrich, St. Louis, MO, USA). For *S. arctica* we also performed transcription start site (TSS) sequencing by treating the total RNA with Terminator 5′-exonuclease (Epicenter, Madison, WI, USA) and resistant mRNAs (i.e., carrying a 5′CAP). The TSS samples were sequenced as two libraries; one treated with tobacco acid pyrophosphatase (TAP; Epicenter) and one untreated. All RNA samples of *S. arctica* were sequenced on Illumina HiSeq2000 machine. Library preparation and sequencing was performed by Vertis Biotechnologie AG (Freising, Germany). For *S. sirkka*, *S. napiecek* and *C. fragrantissima* miRNA libraries and mRNA libraries were prepared and sequenced on the Illumina MiSeq (miRNA: 50 nt single-end, mRNA: 300 nt paired-end) platform at the Norwegian Sequencing Centre.

Mapping of RNA reads and miRNA detection

For *S. arctica*, mapping of all RNA reads was done against the 2012 version of the *S. arctica* genome, downloaded from the Broad Institute (http://www.broadinstitute.org). Also, 100 bp poly(A)-selected RNA Illumina reads from the SRX099331 and SRX099330 *S. arctica* experiments were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The sequenced and downloaded RNA reads were trimmed for low quality nucleotides (phred score cutoff of 20) and sequencing adaptors using Trimmomatic v.0.30 [41], and trimmed for ‘N’ characters and poly(A)-tails using PrinSeq-lite v.0.20.3 [54]. Additionally, only small RNAs reads between 18-26 nts were retained. TSS reads and poly(A)-selected reads were mapped to the *S. arctica* genome using TopHat v2.0.14 [51] with default settings. Small RNAs were mapped to the genome using Blat v3.5 [52] with the options -tileSize = 6 -stepSize = 5 -minScore = 18 -minIdentity = 85 -maxGap = 0 -fine.

For *S. sirkka*, *S. napiecek* and *C. fragrantissima*, small RNAs were trimmed using Trimmomatic v.0.36 to remove adapters and nucleotides with a quality < 28. Only reads longer than 19 nts were retained. The *S. sirkka* reads were mapped to the genome downloaded from NCBI under accession LUCW01000000 and *C. fragrantissima* reads were mapped to the genome downloaded from https://figshare.com/articles/Creolimax_fragrantissima_genome_data/1403592 using Blat as described above. *S. sirkka* and *C. fragrantissima* mRNA reads were quality trimmed and mapped to their respective genomes as described *S. arctica* above.

For miRNA-detection, an adapted version of the MiRMiner pipeline [8] was used to allow for the detection of longer hairpins [Fromm et al. in prep]. For *S. napiecek* there is no genome available so we could not run the MiRMiner pipeline for novel miRNA detection. Instead we mapped the expressed small RNAs to the *de novo* assembled transcriptome (assembled using Trinity v2.0.6 [42] with the–normalize_reads option set, otherwise default settings) with Blat as described above.

The miRNA secondary structures were generated using the mfold web server (http://unafold.rna.albany.edu/?q=mfold/rna-folding-form) with default settings, but structures were constrained from basepairing in the flanking regions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Phylogenetic analyses

Details can be found in the "Phylogenetic annotation of miRNA processing proteins" section. Bayesian analysis was performed with PhyloBayes-MPI v1.5 [49]. Two chains were run with the parameters -gtr and -cat and stopped when the maxdiff was $\leq$ 0.1-0.3 and meandiff < 0.015 with a 15% burnin. Maximum likelihood (ML) analysis was run using RAxML v8.0.26 [50] with the LG model. The ML topology with the highest likelihood score out of 10 heuristic searches was selected as the final topology. Bootstrapping was carried out until the support values had converged (using the AUTO_MRE option). Only support values over 50% for ML and/or over 0.75 for BP were shown on the phylogenies (Figure 3).

Blast searches

Details can be found in the "Reciprocal Blast" section. Reciprocal Blast was performed using Blastp [45] (BLOSUM45 scoring matrix, min e-value 0.01 and max target hits 30).

DATA AND SOFTWARE AVAILABILITY

All sequence data generated in this study has been submitted to the EMBL-EBI European Nucleotide Archive (ENA); small RNA and mRNA transcriptome data, ENA: PRJEB21207; gene assembles, ENA: LS991975–LS991998; miRNAs, ENA: LS992005–LS992065. In addition, sequence alignments used in the phylogenetic analyses are available at Mendeley Data: 10.17632/h96s28wcx9.1 and the Bioportal (www.bioportal.no).