Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic *Micromonas* Populations

Melinda P. Simmons,[†,1,2] Charles Bachy,[†,1] Sebastian Sudek,[1] Marijke J. van Baren,[1] Lisa Sudek,[1] Manuel Ares Jr,[3] and Alexandra Z. Worden[*,1,2,4]

[1]Monterey Bay Aquarium Research Institute (MBARI), Moss Landing, CA

[2]Department of Ocean Sciences, University of California Santa Cruz

[3]Department of Molecular, Cell & Developmental Biology, University of California Santa Cruz

[4]Integrated Microbial Biodiversity Program, Canadian Institute for Advanced Research, Toronto, ON, Canada

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: azworden@mbari.org.

Associate editor: Hongzhi Kong

Abstract

Spliceosomal introns are a hallmark of eukaryotic genes that are hypothesized to play important roles in genome evolution but have poorly understood origins. Although most introns lack sequence homology to each other, new families of spliceosomal introns that are repeated hundreds of times in individual genomes have recently been discovered in a few organisms. The prevalence and conservation of these introner elements (IEs) or introner-like elements in other taxa, as well as their evolutionary relationships to regular spliceosomal introns, are still unknown. Here, we systematically investigate introns in the widespread marine green alga *Micromonas* and report new families of IEs, numerous intron presence–absence polymorphisms, and potential intron insertion hot-spots. The new families enabled identification of conserved IE secondary structure features and establishment of a novel general model for repetitive intron proliferation across genomes. Despite shared secondary structure, the IE families from each *Micromonas* lineage bear no obvious sequence similarity to those in the other lineages, suggesting that their appearance is intimately linked with the process of speciation. Two of the new IE families come from an Arctic culture (*Micromonas* Clade E2) isolated from a polar region where abundance of this alga is increasing due to climate induced changes. The same two families were detected in metagenomic data from Antarctica—a system where *Micromonas* has never before been reported. Strikingly high identity between the Arctic isolate and Antarctic coding sequences that flank the IEs suggests connectivity between populations in the two polar systems that we postulate occurs through deep-sea currents. Recovery of Clade E2 sequences in North Atlantic Deep Waters beneath the Gulf Stream supports this hypothesis. Our research illuminates the dynamic relationships between an unusual class of repetitive introns, genome evolution, speciation, and global distribution of this sentinel marine alga.

Key words: **introns, marine algae, polar systems, phytoplankton, repetitive introns, Introner Elements.**

Introduction

Spliceosomal introns are distinctly eukaryotic gene features whose origins remain mysterious. Intimately linked with eukaryote evolution, introns interrupt coding information and must be removed from primary RNA transcripts by splicing. Introns and splicing are thought to provide eukaryotes with mechanisms for diversifying mRNA molecules from a gene after transcription. Mutations that create new mRNA splicing patterns can convey advantages, in particular, by encoding novel proteins (Gilbert 1978; Koonin 2006), or even generating multiple functionally distinct protein products from an individual gene (Modrek and Lee 2002). Splicing can also modulate gene expression through mechanisms such as nonsense-mediated decay and intron-mediated enhancement (Brogna and Wen 2009; Parra et al. 2011).

Although introns play influential roles in eukaryotic biology, the molecular and evolutionary processes that structure their distributions in genomes remain difficult to trace (Curtis and Archibald 2010; Rogozin et al. 2012). In divergent taxa such as mammals and plants, introns can be found in homologous positions of orthologous genes where they are neutrally evolving and lack sequence homology (Sverdlov et al. 2007; Rogozin et al. 2012). This has been taken as evidence that the last eukaryotic common ancestor (LECA) had an intron-rich genome, that introns were a part of genes since the earliest evolutionary stages of life, and that the lack of spliceosomal introns in present-day archaea and bacteria resulted from subsequent streamlining (Roy 2003; Koonin 2006; Roy and Gilbert 2006). An alternative hypothesis holds that introns appeared and spread with the emergence of eukaryotes, with intron loss being the dominant process since descent from ancestral eukaryotes (Koonin 2006; Roy 2006; Csuros et al. 2011).

Article

Fast Track

Several studies on closely related taxa have recently suggested that intron gain is an important and ongoing process. These studies have uncovered situations in which only one of a pair of orthologous genes has an intron, representing either recent insertion or precise deletion of the intron (e.g., Llopart et al. 2002; Li, Tucker, et al. 2009). There are more than two dozen such intron presence–absence polymorphisms between two isolates of the crustacean *Daphnia pulex*, suggesting that parallel intron gains have recently occurred at homologous positions within orthologous genes (Li, Tucker, et al. 2009). Short repeats (5–12 nt) flanking the inserted introns led to the hypothesis that polymorphic *D. pulex* introns are derived from double-strand break repair (Li, Tucker, et al. 2009), although the origin of the donor intron sequence (needed for gain) is not clear.

As genomes from closely related organisms are sequenced, examples of new intron types are also emerging. These involve intron presence–absence polymorphisms where identical or nearly identical introns are present in one genome but are seemingly absent from related taxa, suggesting that these repetitive introns act as transposable elements to propagate across a given genome (Worden et al. 2009). Originally reported in the unicellular green alga *Micromonas*, an ecologically important genus of marine prasinophyte algae, these repeated intron families (intron elements, IEs) have three properties in common (Worden et al. 2009). First, members of a single family of IE have highly similar sequences, for example, a family called IE3 (Worden et al. 2009) has 32 members with identical nucleotide sequences scattered about the *Micromonas pusilla* CCMP1545 genome, and many more members that differ only at a single position. Second, each individual IE resides within a transcription unit in the sense orientation, and is removed after transcription by the spliceosome during mRNA processing. Third, IEs display intron presence–absence patterns characteristic of intron gain by repeat expansion in the genomes where they are abundant. Examples of repetitive introns also appear in the larvacean tunicate *Oikopleura dioica* (Denoeud et al. 2010), and a number of terrestrial fungi (Torriani et al. 2011; van der Burgt et al. 2012). Notably, no repetitive intron family described thus far appears to encode a protein that could promote selective reverse splicing or transposition of these unusual introns.

IEs provide an interesting case study because *Micromonas* appears to have large effective population sizes with periodic isolation and reduction on short time scales (seasonal) as well as long-term isolation influenced by changes in glaciation, land mass organization, and ocean circulation. Although *Micromonas* has low intron numbers relative to other Viridiplantae, such as chlorophytes and land plants (Worden et al. 2009; Blanc et al. 2010), the 22 Mb genome of *M. pusilla* CCMP1545 is 1 Mb larger than that of *Micromonas* sp. RCC299 due almost entirely to the presence of four IE families (IE1–IE4) that collectively have over 6,000 members (Worden et al. 2009). None of these families is found in RCC299, which instead contains a small, distinct IE family of approximately 221 members (Verhelst et al. 2013). These two isolates share at most 90% of their protein-encoding genes, and represent two of six known *Micromonas* clades, each thought to represent different species (Slapeta et al. 2006; Worden 2006; Worden et al. 2009). It remains an open question whether IEs are present in other *Micromonas* clades or are an atypical feature that is peculiarly abundant in the genome of CCMP1545, a North Atlantic strain isolated in the 1950s.

We systematically searched *Micromonas* isolates from around the world that represent the five established cultured clades to determine whether IEs are present in multiple clades. Combined with metagenomic analyses, our results reveal new IE families, expanding our understanding of these curious elements. Furthermore, we find that a newly delineated *Micromonas* clade containing the Arctic isolate CCMP2099 is widespread in the Southern Ocean, where *Micromonas* has not previously been reported. Environmental polymerase chain reaction (PCR) and cloning-based studies demonstrate the presence of this clade in the deep current that transports Arctic waters to the Southern Ocean, as well as polymorphic insertions of other IE families in Pacific Ocean populations. Our studies highlight the utility of IE families to track global distributions of *Micromonas* species. Moreover, by comparing the new IE families an unusual structural feature was identified that is potentially relevant to the spread of IEs. These results lead us to propose a novel model for the mechanism of intron transposition and to postulate that IEs influence the diversification of *Micromonas* lineages.

Results

Micromonas Clades Distinguished by Cultured Isolates and Environmental Clones

The first goal of our study was to evaluate the presence and distribution of IEs in Clade D isolates apart from CCMP1545 (representing the species *M. pusilla*), as well as in other *Micromonas* clades. Because this effort relies on robust clade discrimination, *Micromonas* isolates were analyzed using new and existing 18S rRNA gene data from cultured strains and environmental clone libraries. The phylogenetic reconstruction established the existence of seven clades (fig. 1*a*) designated here largely according to a previously established lettered naming system for cultured clades and Clade -.IV for the uncultured clade (Slapeta et al. 2006; Worden 2006) (table 1, supplementary table S1, Supplementary Material online). In the 18S rRNA gene phylogeny the key node separating *Micromonas* Clades A and B did not acquire bootstrap support (fig. 1*a*); however, these clades were distinguished using a concatenation of four protein-encoding genes (supplementary fig. S1, Supplementary Material online, see below). Different from prior phylogenies, Clade E members formed two distinct clades due to incorporation of new data (fig. 1*a* and supplementary fig. S1, Supplementary Material online). We termed these Clades E1 and E2, with the latter containing the Arctic isolate CCMP2099.

We next evaluated prasinophyte orthologs of four genes carrying IEs in CCMP1545. The selected genes were

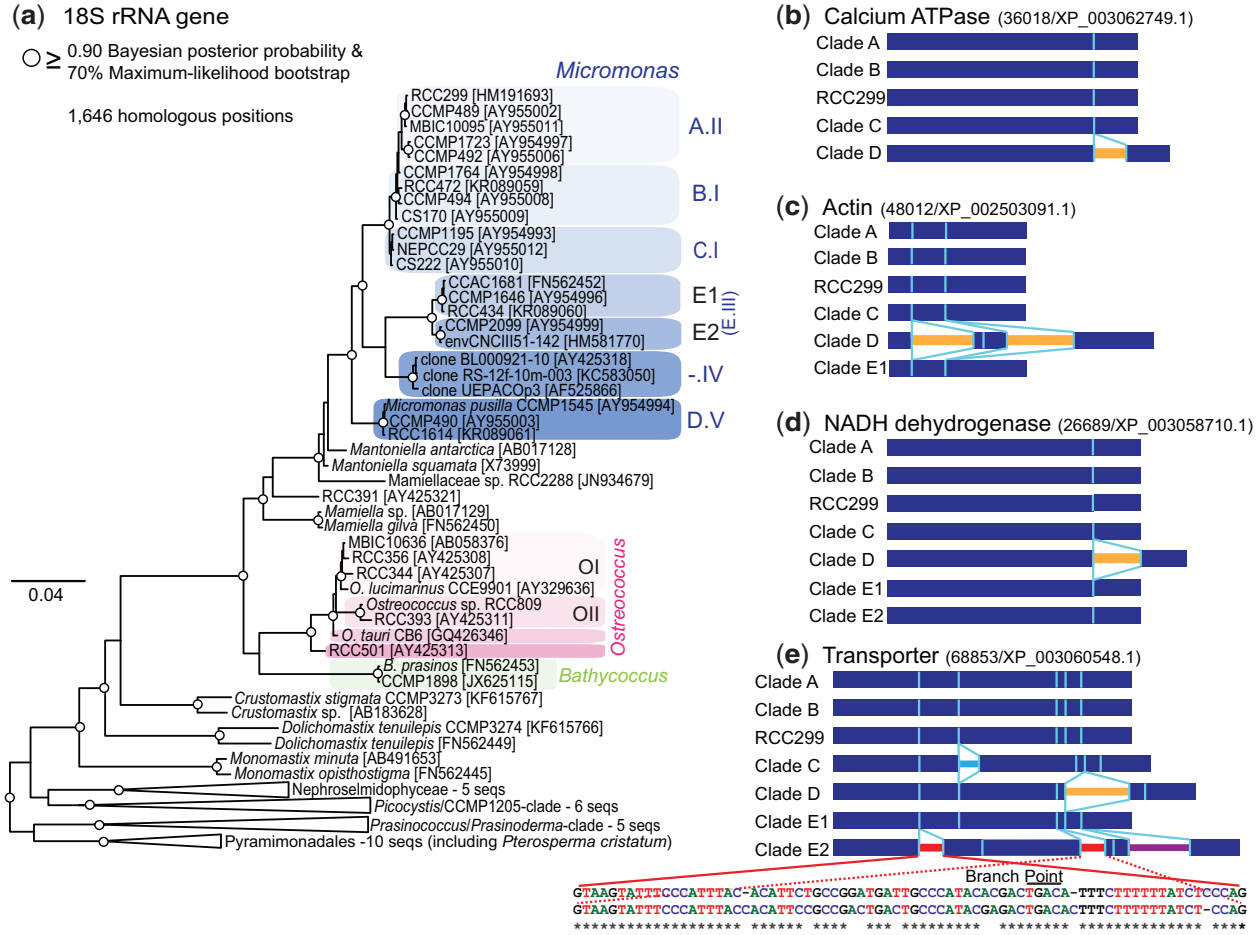

Fig. 1. Molecular phylogeny of *Micromonas* and insertion sequences in gene homologs from cultured clades. (*a*) Bayesian reconstruction of the 18S rRNA gene sequences from the Mamiellophyceae and other prasinophytes, using 1,646 unambiguously aligned positions and the GTR + Γ + I model of substitution. *Micromonas* clades (blue) are highlighted. Clade names are designated with letters, as in Slapeta et al. (2006) and roman numerals, as in Worden et al. (2009). Differentiation of Clade E.III to Clades E1 and E2 (black labeling) was achieved herein using new data. Sequences from environmental clone library studies were included for Clade -.IV (an uncultured clade) and groups with sparse representation in culture collections, such as the E2 Clade. Other widespread Mamiellophyceae genera shown, *Ostreococcus* (pink) and *Bathycoccus* (green), also have genome-sequenced representatives used in primer design for the IE PCR study. The tree is rooted by the prasinophyte *Pycnococcus*-clade for display purposes. (*b–e*) Architecture of amplified regions of protein-encoding genes investigated in cultured *Micromonas* clades (table 1). Thick bars (blue) represent exons, vertical turquoise lines denote loci where introns are present (accompanied by thin horizontal intron lines) or absent (vertical line only). Thin horizontal lines represent Clade D IEs (yellow) and newly identified introns in Clade C (blue) and Clade E2 (red, purple) homologs of the Transporter. The first two Clade E2 introns (red) are highly identical (alignment under panel [*e*]). Note that E1 and E2 data are lacking for the ATPase, as was Actin for E2 presumably due to primer mismatches later identified using transcriptome assemblies.

identified in genome sequences available from two *Micromonas* and from other members of the prasinophyte class Mamiellophyceae, specifically two *Ostreococcus* strains (fig. 1*a*). The genes encode a putative Calcium ATPase (hereafter the gene is referred to as ATPase), a putative NADH dehydrogenase [ubiquinone] flavoprotein 1 (hereafter Dehydrogenase), Actin, and an ATP-binding cassette transporter (hereafter Transporter). PCR primers were designed against available genome sequenced prasinophytes to amplify gene regions spanning one or more known IE in CCMP1545 (supplementary table S2, Supplementary Material online). *Micromonas* clade representatives from each cultured clade were then selected, grown, and these regions examined using PCR-based sequence data (table 1).

Micromonas Clade D representative CCMP490, isolated in the western North Atlantic, had IEs at the same insertion positions and phases as in the CCMP1545 orthologs (fig. 1*b–e*). Identities for IEs at homologous positions were 100% and 99% for the two Actin IE1s and 100% (IE2), 98% (IE1) and 97% (IE3) for those in the ATPase, Transporter and Dehydrogenase, respectively. Three of these were phase-0 (i.e., introns located between two codons), whereas the 5′-most IE in the Actin gene fragment was phase-1 and the Transporter IE was phase-2. Splice sites (ss) were canonical, with GT/AG ss in the Actin 3′-IE, ATPase and Dehydrogenase or GC/AG ss for the Actin 5′-IE and Transporter IE (supplementary table S3, Supplementary Material online). However, *Micromonas* Clade D IEs were not found in Clades A, B, C, and E1 or E2

Table 1. *Micromonas* Isolates Grown and Number of Assembled Sequences Obtained from Clones for Each Gene Homolog Investigated.

Isolate	Clade	Actin	ATPase	Transporter	Dehydrogenase
RCC299	A[a]	2	2	1[b]	2
CCMP492	A	2	2	0[b]	2
CCMP1764	B	2	2	2	2
NEPCC29	C	2	0[c]	2	2
CS222	C	2	2	2	2
CCMP1195	C	2	2	2	2
CCMP490	D	2	2	2	2
CCMP1545	D	2	2	2	2
CCMP1646	E	2	0[d]	2	1
CCMP2099	E	0[‡]	0[d]	2	2

NOTE.—Clade designations based on Slapeta et al. (2006).

[a]RCC299 was not included in Slapeta's analyses; therefore this assignment is based on phylogenetic analyses herein.

[b]The primers produced sequences from a different predicted ABC Transporter in Clade A strain CCMP492 and in RCC299; the correct RCC299 gene homolog was obtained from the sequenced genome and the CCMP492 amplicon was discarded from further analyses.

[c]While successful for other Clade C strains, the correct ATPase homolog was not retrieved in cloned NEPCC29 sequences.

[d]Comparison to transcript sequences, obtained later from Clade E2 isolate CCMP2099, revealed extensive primer mismatches for these genes, likely explaining unsuccessful PCR results.

orthologs (fig. 1*b*–*e*) and were therefore termed D-IEs (table 2).

Discovery of New IE Families with Lineage-Specific Distributions

Although D-IEs were not present in the other *Micromonas* clades, new IEs were discovered. Three novel introns were identified in the Transporter gene of *Micromonas* Clade E2 representative CCMP2099, but not in the E1 isolate (fig. 1*e*). These were in nonhomologous insertion positions to D-IEs from Clade D isolates. The two 5′-most introns in the CCMP2099 Transporter have 89% identity to one another (fig. 1*e* alignment), higher than expected for regular spliceosomal introns (RSIs). The 5′-most novel intron is phase-1, the next phase-0, and the last phase-1. The latter, located near the 3′-end of the Transporter PCR product, is longer (185 nt) than the upstream introns (74 and 75 nt). BLASTn queries against the RCC299 and CCMP1545 genome assemblies, as well as CCMP1764 genomic DNA reads did not recover significant hits. Thus, the newly identified CCMP2099 introns do not appear to be similar to IEs in other *Micromonas* clades with sequenced genomes.

Given the lack of additional genome data, it remained unclear whether the new introns were present in other CCMP2099 genes, as expected of an IE family. Therefore, environmental data were searched and multiple copies of the shorter CCMP2099 Transporter intronic sequences (type 1) were recovered in metagenomes from the Antarctic (Ace Lake, Southern Ocean, Ross Sea; fig. 2*a*). The detected sequences interrupted multiple different protein-encoding genes according to BLASTx analyses of the metagenomic reads against NCBI's nr database. Those with known functions included an autophagocytosis-associated protein, a putative aminopeptidase, transcriptional repressors, and an early light-induced protein (ELIP). These CCMP2099 "type 1" intronic sequences were identical or highly identical to those in metagenomes (fig. 2*b* and *c*) enabling the identification of a conserved intron motif using over 100 environmental sequences (supplementary fig. S2*a*, Supplementary Material online). This indicates that the two 5′-most introns (i.e., type 1) from the CCMP2099 Transporter represent an IE family present in Clade E2 that is not present in data from other *Micromonas* clades or in nonpolar environmental data. We therefore termed this new repetitive intron family E2-IEt1.

As seen for E2-IEt1, the distinct Clade E2 Transporter intronic sequence (type 2) garnered multiple hits in Antarctic metagenomic data, revealing another IE family (E2-IEt2; fig. 2*a*). IE flanking regions of the recovered reads represented yet other protein-encoding genes, including a putative amidophosphoribosyl transferase, a pre-mRNA-processing-splicing factor, cytochrome P450 monooxygenase, eukaryotic translation initiation factor 6, and transcription initiation factor TFIID sub.10. The longer E2-IEt2s averaged 186 ± 8 nt (supplementary fig. S2*b* and *c*, Supplementary Material online, ranging up to 207 nt) and were always geographically colocated with E2-IEt1 although the latter were found at additional sites. Detection efficiency for E2-IEt2s is likely lower than for E2-IEt1s because they are less likely to be captured completely within approximately 300-bp 454-metagenomic reads. Both E2-IE types were detected in many surface samples, but were also found below the photic zone at two sites at 330 m (at −66.57, 142.32) and 1,320 m (−67.07, 145.20, E2-IEt1 family only).

RNA-seq transcriptome assemblies from CCMP2099 (McRose et al. 2014) revealed further similarities between this Arctic isolate and the Antarctic metagenomic sequences. CCMP2099 transcripts matched predicted exonic coding regions from Antarctic E2-IE-containing metagenomic reads (see e.g., flanking sequence; fig. 2*c* and supplementary fig. S2*c*, Supplementary Material online). Three hundred and sixty E2-IEt1-containing metagenomic reads with exonic regions present on both sides of the E2-IEt1 hit came from 165 different proteins also present in CCMP2099 transcriptome assemblies. Thirty-one and 248 of these metagenomic reads shared 100% and 99% coding sequences (CDS) identity with the matching CCMP2099 transcript, respectively, whereas the overall average was 98 ± 2% for all 360 sequences. The results experimentally confirmed splicing of E2-IEs as well as exclusive use of canonical GT/GC donor and AG acceptor sites. The results also demonstrate that *Micromonas* resides in Antarctica, and that Arctic and Antarctic populations have high nucleotide conservation.

The North Atlantic Deep Waters (NADW) provide a connection between Arctic and Antarctic waters because they form and sink in the Labrador and Nordic Seas, then flow in a deep (> 1,500–2,000 m depth), thick layer through the North Atlantic and South Atlantic to the Southern Ocean where some upwelling occurs in the Wedell Sea (Broecker 1991; Morozov et al. 2010; Talley 2013). The NADW signature is strong on the western side of the North Atlantic basin.

Table 2. IE Families and Their Distribution in the *Micromonas* Clades.

This Study	Worden (2009)	Verhelst (2013)	Strain or Metagenomic Read
D-IE1	IE1	IEA1	CCMP1545, CCMP490, temperate & tropical metagenomes
D-IE2	IE2	IEA2	CCMP1545, CCMP490, temperate & tropical metagenomes
D-IE3	IE3	IEA3	CCMP1545, CCMP490 (metagenomes not searched)
D-IE4	IE4	IEA4	CCMP1545 (CCMP490 & metagenomes not searched)
Unconf.	Not reported	IEB	CCMP1545
Unconf.	Not reported	IED	CCMP1545
ABC-IE	Not reported	IEC, seen in RCC299	NEPCC29, CS222, CCMP1195, CCMP1764, RCC299, temperate metagenomes
E2-IEt1	Not reported	Not reported	CCMP2099, NADW, Antarctic metagenomes
E2-IEt2	Not reported	Not reported	CCMP2099, NADW, Antarctic metagenomes

NOTE.—Families IEB and IED reported in Verhelst et al. (2013) are considered unconfirmed; these groups have very few members that are very diverged and most are not spliced as annotated in RNA-seq data.

Therefore, we extracted DNA from depth profile samples taken in the region of the Gulf Stream Current, several of which showed temperature and salinity characteristics of the NADW (supplementary table S4, Supplementary Material online). We also designed primers specific to the flanking sequence of the CCMP2099 Transporter E2-IEt1, applied them to these samples, and obtained a PCR product of the anticipated 200 bp size in the 3,000 m sample. Bands with product sizes $\geq$300 bp were present in the 500, 2,000, and 4,000 m, as well as 90 m from a subtropical North Atlantic cast. Sequences from clones of the larger products came from bacteria (Verrucomicrobia, Planctomycetes, and Actinomycetes) and were unlike the E2-IEt1 sequence, whereas sequences from the cloned product of the 3,000 m sample came from CCMP2099 (98–100% nt identity).

Other new introns were found in *Micromonas* Clade C isolates. These were in the Transporter gene but at a non-homologous position to both Clade D and Clade E2 IEs (fig. 1*e*). These phase-0 introns had one nucleotide polymorphism between each Clade C strain. PCR products from Clade A and B orthologs did not contain introns, but 64 hits (*E* values 10^{-5}–10^{-8}, with nucleotide identity 88–94%) were retrieved when the Clade C intron was used as a query against genomic DNA sequences from Clade B isolate CCMP1764 (see supplementary fig. S3*a*, Supplementary Material online). Hits were also recovered in RCC299 and the two best of these represented confirmed introns, one in a putative Calmodulin-binding protein (JGI Prot. ID 55550) and the other in Ribonuclease H (JGI Prot. ID 105055) (supplementary fig. S3*b*, Supplementary Material online). Across Clades A, B, and C, the identified intron sequences shared higher sequence identity (~80%) than observed for RSIs (<50%, see Materials and Methods). Highly similar sequences were also present in metagenomic data (fig. 2*a*) with flanking regions that encode different proteins, for example, a putative intraflagellar transport protein and a putative superfamily I helicase. Ss were confirmed by aligning metagenomic reads with transcripts from Clade C isolate NEPCC29 (e.g., supplementary fig. S3*c*, Supplementary Material online) and compositional features evaluated using an alignment of metagenomic and culture-derived sequences (supplementary fig. S3*d*, Supplementary Material online). The presence of highly similar intron sequences in the Clade C Transporter gene, multiple *Micromonas* Clade A and B genes, and in metagenomes, identified them as yet another IE family, termed here ABC-IE (table 2).

As observed for cultures, intron phase varied for IEs in metagenomic reads. ABC-IEs in manually curated metagenomic sequences that captured both 5′ and 3′ ss ($n = 13$) were phase-0 (7) and phase-1 (6). For a manually curated subset of E2-IEt1s that met the same criteria ($n = 10$), phase-0 (6), phase-1 (3) and phase-2 (1) were observed. E2-IEt2s were also found at all three phases, but distributed as phase-0 (3), phase-1 (5) and phase-2 (2). Additionally, an E2-IE (E2-IEt1.10) in the Antarctic metagenomic data was in the same *ELIP* gene and codon as a D-IE1 in CCMP1545, but in a different phase (supplementary fig. S4, Supplementary Material online).

Natural Variation in Polymorphic Introns Detected by Environmental Cloning

Micromonas Clade D is also found in the North Pacific Ocean (Worden 2006), but IE sequence similarities to the Atlantic Clade D isolates CCMP1545 and CCMP490, or indeed IE presence, are unknown. Two of the PCR primer sets targeting both *Micromonas* and *Ostreococcus* were used to construct environmental clone libraries from samples collected in spring and autumn in the eastern North Pacific Ocean. Of the total clones successfully sequenced, 122 (ATPase) and 294 (Actin) came from the targeted homologs and comprised intron-bearing (24 ATPase; 160 Actin) and intronless (98 ATPase; 134 Actin) sequences. The latter came from *Micromonas* (although not Clade D), *Ostreococcus* and sometimes more distant taxa. Apart from determining their clade assignment, intron-less sequences were not further analyzed.

The majority of Pacific environmental *Micromonas* Clade D sequences contained IEs in homologous positions as in the Clade D cultures, and all had canonical ss (fig. 3*a* and *b* and supplementary table S3, Supplementary Material online). Sequence clustering showed that, for both genes, some environmental clones were identical to CCMP1545 and CCMP490 throughout the gene and IE, whereas others had nucleotide differences (fig. 3*c* and supplementary fig. S5 and datafile S1, Supplementary Material online). For environmental clones

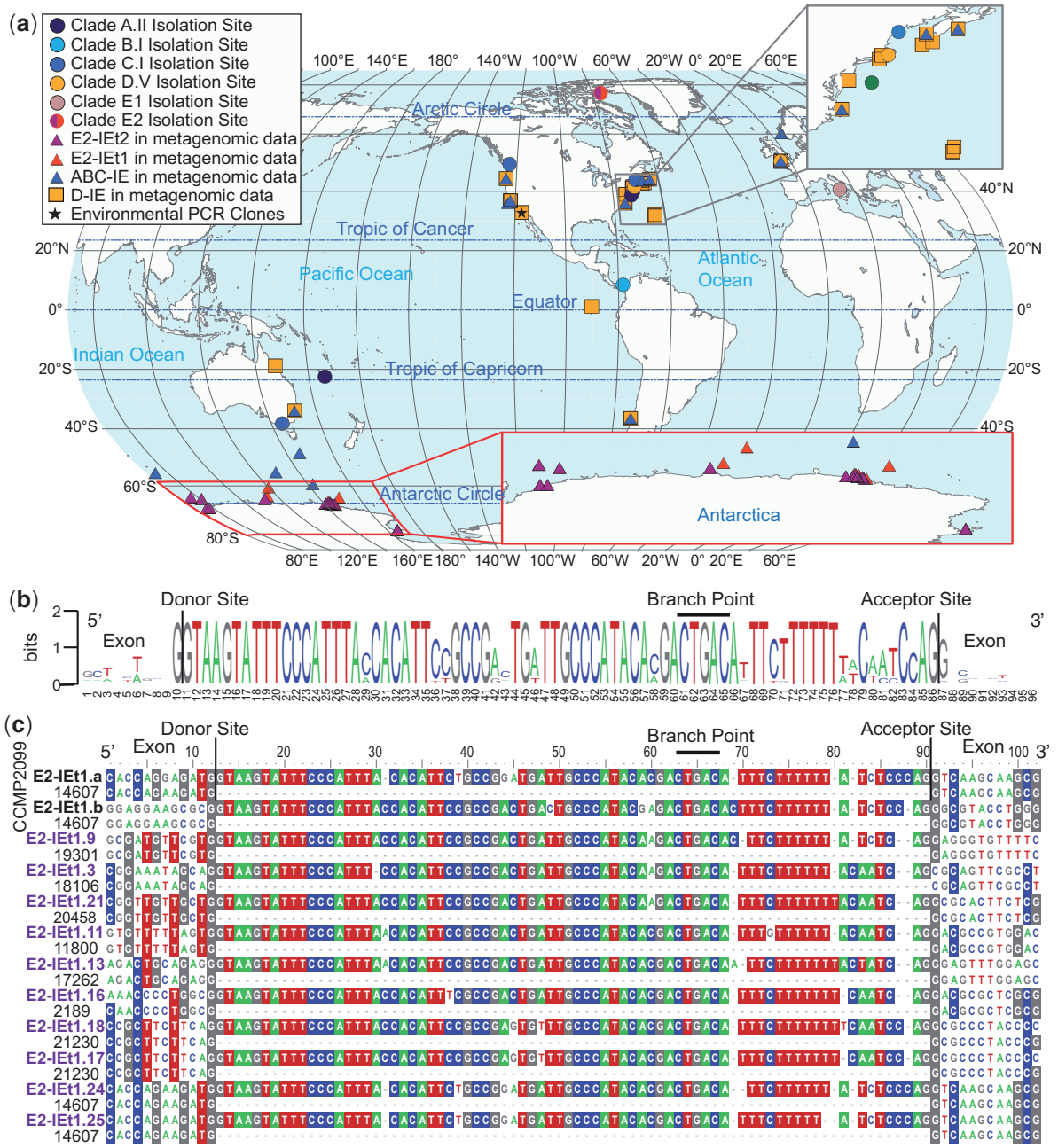

Fig. 2. The global distribution of *Micromonas* introns in available metagenomes and discovery of new IE families. (*a*) Isolation sites for cultured *Micromonas* strains (circles), the sample site for environmental clone libraries generated herein (star), and sites where multiple BLASTn hits were recovered in public metagenomic data (symbols and color-codes as indicated on legend). Inset borders are color-coded to show corresponding map regions. Note that red triangles (representing E2-IEt1) lay beneath every purple triangle (E2-IEt2 sequences) and the location of the deep profile (supplementary table S3, Supplementary Material online) is not shown. (*b*) E2-IEt1 consensus sequence from Antarctic metagenomic reads encoding eight different proteins. (*c*) Aligned E2-IEt1 (and 12 exonic flanking nucleotides at each end) from Antarctic reads, including two from the same gene present in different samples (2 such examples, bottom 4 E2-IEs; excluded from [*b*] to avoid overrepresenting element conservation) and from the CCMP2099 Transporter gene. CCMP2099 transcript contigs (nonbold numbers) are shown beneath each DNA sequence. Regions flanking the Arctic CCMP2099 E2-IEt1.a and Antarctic metagenomic E2-IEt1.24 and E2-IEt1.25 (from different Antarctic samples) in the Transporter gene were identical, as were the E2-IEs themselves except a single "T" at different positions in E2-IEt1.24 and E2-IEt1.25 (potentially representing 454 homopolymer accuracy issues).

containing both Actin D-IEs, the 5′- and more 3′-IEs had 97–100% and 98–100% nucleotide identity, respectively.

Intron-bearing clones with the least conservation to the ATPase and Actin genes from D-lineage cultures also had deviant intron numbers. Spring and autumn clones (P3I4_28 and P5I4_80, respectively) lacked the Actin 5′- and 3′-IEs, but contained a phase-0 intron at a nonhomologous intervening position (fig. 3*b*, Env. Clade D-like). These two nearly identical clones (one mismatch) have higher CDS identity to Clade D isolates (96%) than to other cultured *Micromonas* clades (~90–91%), and appear to represent a basal Clade D group. High identity across the amplified

Actin CDS region precluded further resolution by phylogenetic approaches and because similar sequences were not detected in available *Micromonas* genomic or metagenomic data, we concluded that this intron is a polymorphic RSI. However, introns in the other divergent Clade D environmental clusters were D-IEs. Despite sharing 99% CDS identity, Actin spring clone P3I4_VIIII47 (Cluster S4; supplementary fig. S6, Supplementary Material online) lacked the more 5′ D-IE1 and the 3′ D-IE1 had relatively low identity (93%) to homologously positioned D-IE1s (fig. 3*b* and supplementary fig. S5, Supplementary Material online). Notably, a phase-0 intron with canonical ss was also found in Actin from the prasinophyte *Pterosperma cristatum*, at the same codon as the more 3′ D-IEs in *Micromonas* Clade D Actin genes (also phase-0; supplementary fig. S6, Supplementary Material online). Finally, in the ATPase Cluster B (6 clones), a D-IE1 (phase-0) was identified upstream of the D-IE2 shared with cultures and other environmental Clade D clusters (fig. 3*a* and *c*). Cluster B CDS (and the more 3′ D-IE) nucleotide differences were also distinct (fig. 3*c*) and phylogenetic analysis placed it in a supported position basal to the *Micromonas* Clade D-lineage, with low evolutionary distance from the cultured strains (supplementary fig. S7, Supplementary Material online). Collectively, the intron presence–absence polymorphisms observed in wild Pacific *Micromonas* closest to Clade D Atlantic isolates and the nucleotide polymorphisms present in environmental clusters are suggestive of a dynamic IE landscape that influences the development of discrete *Micromonas* populations.

Structural Features of IE Families

We next searched for RNA structural features that might hint at mechanisms of transposition. Previous comparisons of IE sequences have been limited to a few families in the same genome (Worden et al. 2009; Verhelst et al. 2013), where multiple expansion episodes have created sequence-related sets of elements within which conservation of mechanistically required features are difficult to discern. Here, we chose IE groups from within the ABC, D and E2 lineages that were present in multiple exact copies as these likely represent recently active elements. Three short motifs conserved between these groups reflect function in splicing (fig. 4, see also fig. 2*b* and *c*, and supplementary figs. S2 and S3 and table S5, Supplementary Material online): GYRaGu represents the 5′ ss, GacUGAC contains the intron branchpoint (underlined), and CAG is at the 3′ ss (not shown in fig. 4). Other than these regions and 1-2 pyridimine (mostly U) rich segments, no primary sequence similarity was detected between ABC-IE (fig. 4*a* and *b*), D-IE (fig. 4*c*), and E2-IEt1 (fig. 4*d*). However, all had a sequence complementary to their respective 5′ ss within a few tens of nucleotides downstream.

Discussion

The discovery of repetitive elements that bear properties of spliceosomal introns has raised a number of questions about the origin of introns and their influence on eukaryotic genomes. Until now, such introns had only been observed in *Micromonas* Clade D (Worden et al. 2009) and more minimally in Clade A (Verhelst et al. 2013) as well as in several fungi (Torriani et al. 2011; van der Burgt et al. 2012). They appear to be absent from closely related prasinophytes, plants, mammals, and other taxa for which genome or gene sequences are available. Given biases in genomic resources (Keeling et al. 2014), their true distributions across the broader expanses of the eukaryotic tree of life remain unknown.

Here, we report new IE families and document polymorphic introns, including some RSIs, in cultured and wild *Micromonas*. The results demonstrate that multiple pervasive IE families exist which have no extended sequence homology to introns or genomic sequence in distant *Micromonas* clades, but instead correspond to individual lineages. We have designated these using a lettered prefix representing the *Micromonas* clade(s) they occupy, including modification of the original CCMP1545 nomenclature to D-IEs (table 2). Our results lend insight into several outstanding questions on biogeography of this algal genus as well as intron and IE evolution.

Diversification of *Micromonas*

The established *Micromonas* clades are thought to reflect species level differences and have been defined using multiple marker genes (Slapeta et al. 2006; Worden 2006; Marin and Melkonian 2010) (fig. 1*a* and supplementary fig. S1, Supplementary Material online). Clade D, in which IEs were first reported (Worden et al. 2009), is the most basal *Micromonas* group. Its divergence from other clades is estimated at 66 ± 10 Ma with evolutionary distances that appear to be greater than those between *Maize* and *Oryza* (Slapeta et al. 2006). Across the *Micromonas* clades examined here, only Clade D isolates (*M. pusilla*) contain D-IEs. D-IEs remain the best characterized repetitive introns, in part because the CCMP1545 genome has been sequenced and because they are numerous. D-IEs range in length, but the most abundant group (D-IE1, 6,112 members) is on average slightly shorter (173 nt) than the 3,553 RSIs (192 nt) in CCMP1545 (Worden et al. 2009; Verhelst et al. 2013). Multiple D-IEs in North Pacific environmental clones had 100% identity to those from the two North Atlantic Clade D isolates although these water masses last had surface water connectivity approximately 3 Ma or more ago (before the formation of the Isthmus of Panama) (Molnar 2008). Separation of the relevant populations is presumably older, given the locations where CCMP1545 and CCMP490 were isolated and circulation patterns. Together, the results indicate that the putative founder D-IE invaded the genome at or shortly after the time of Clade D separation from other *Micromonas* clades, potentially contributing to its divergence.

The more recent divergence of Clades A, B, and C, relative to other *Micromonas* clades, is underscored by ABC-IE relatedness. The ABC-IEs in the Transporter intron presence–absence polymorphism, and other CCMP1764 (Clade B) and RCC299 (Clade A) genes have high identity to each other as well as to a family of approximately 200 members reported

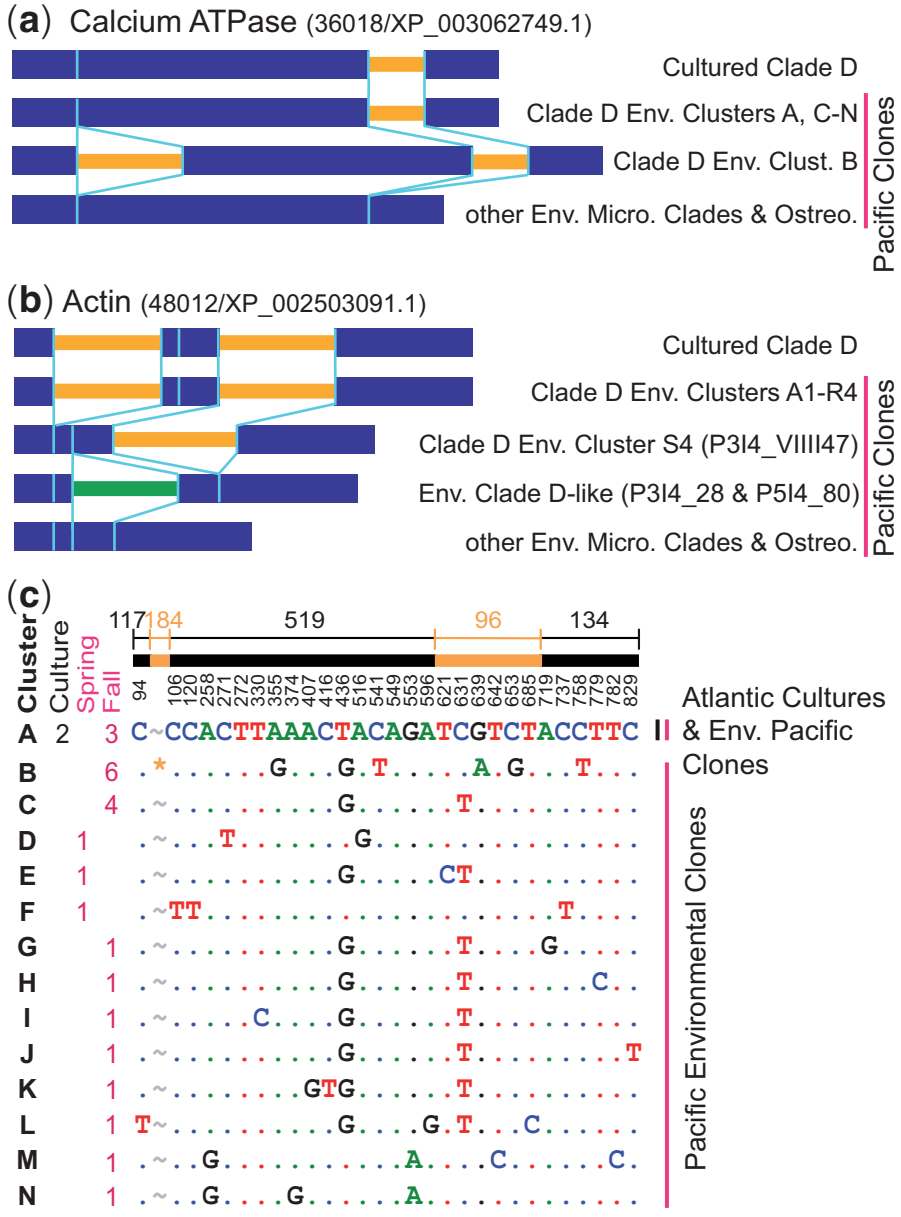

Fig. 3. Intron presence–absence patterns in Pacific Ocean environmental clones. Architecture for regions of the genes encoding (*a*) the putative Calcium ATPase and (*b*) Actin are shown. Thick bars (blue) represent exons, vertical turquoise lines denote loci where introns are present (accompanied by thin horizontal intron lines) or absent (vertical line only). Thin horizontal lines represent D-IEs (yellow, D-IEs) and a newly identified presence–absence polymorphism (green) in environmental clones similar to Clade D. ATPase Cluster B consists of six environmental sequences, whereas Actin Cluster S4 and the RSI-bearing Clade D-like type have one and two clones, respectively. (*c*) Nucleotide polymorphisms in the amplified region of IE-bearing ATPase homologs. Coding region (black) and D-IEs (orange) lengths are indicated above top bar and numbering below corresponds to SNP positions. The number of sequences (100% identical) in each cluster from cultures and environmental clones from spring or fall Pacific clone libraries is indicated. Dots represent identical nucleotides to those of the first sequence and variants denote other nucleotides. Only positions with polymorphisms are shown. The asterisk (orange) represents a 5′-IE (184 nt) in Env. Cluster B sequences, absent from all other ATPase sequences (variant nucleotide numbering does not include the Cluster B 5′-IE).

in RCC299 (Verhelst et al. 2013) (table 2, supplementary fig. S3, Supplementary Material online). This suggests that the common ancestor of these clades contained ABC-IEs, whereas their heterogeneous distribution in orthologs (e.g., fig. 1*e*) from the three clades indicates that differential colonization (gains) and/or losses are connected to subsequent clade divergence.

The discovery of additional IE families, such as those unique to Clade E2, lends further support to the idea that IE invasion may relate to diversification. Novel E2-IEt1s and E2-IEt2s were present in CCMP2099 and Antarctic metagenomic data, but none was found in the PCR-tested genes from Clade E1 isolate CCMP1646 (for which a genome sequence is lacking). This result could arise from Clade E1 under

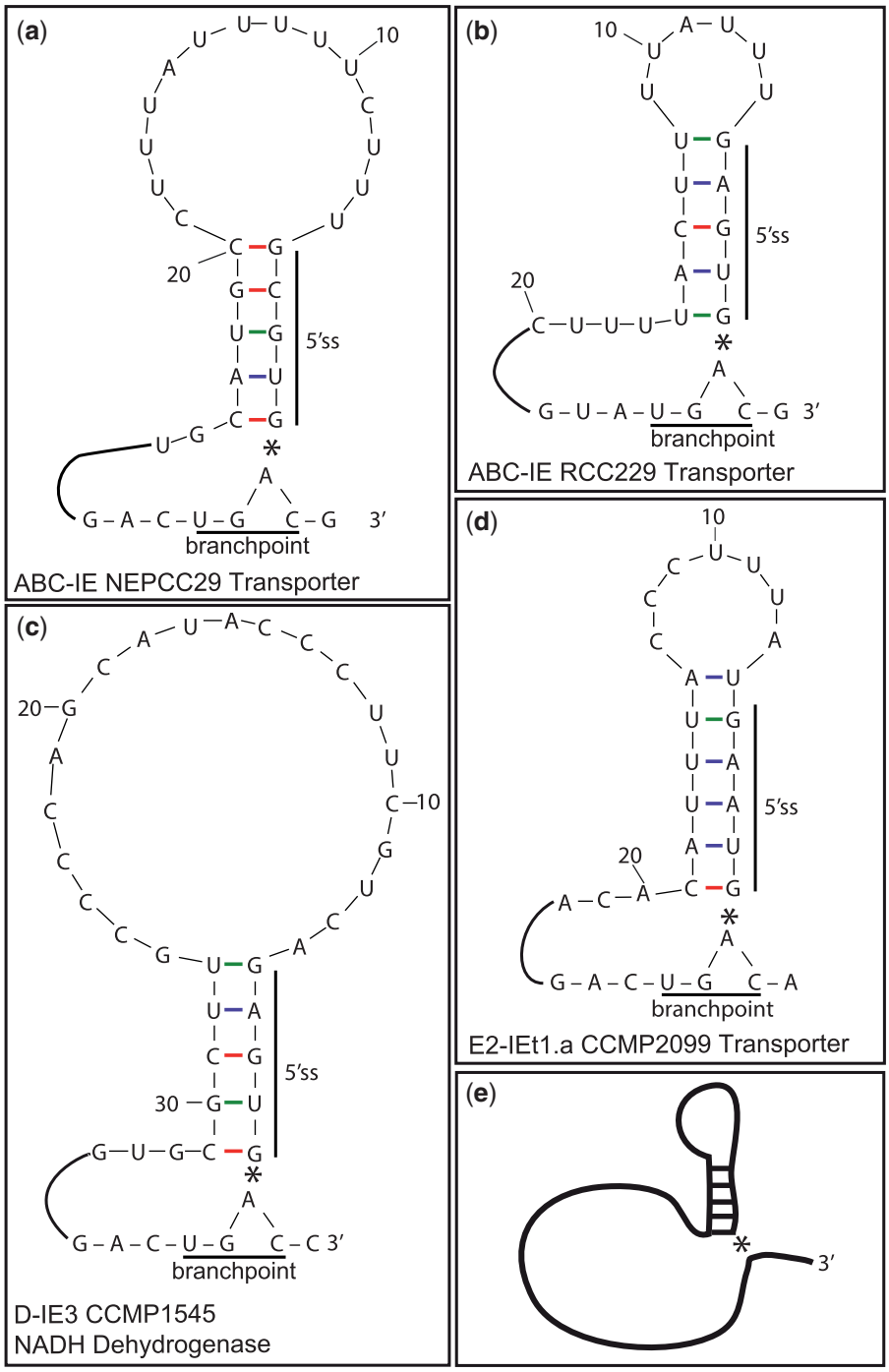

Fig. 4. Secondary structure models for Introner lariat RNAs. A sequence complementary to the 5′ splice site is found in several IE types. The feature does not appear as a conserved primary sequence element because its sequence varies to maintain pairing with the 5′ ss. 2′-5′ linkage between the branchpoint A residue and the G at the 5′-end of the intron is shown with an asterisk and bases are numbered from the beginning of the intron. Additional sequences between the 5′ splice site and the branchpoint are represented by a line, and sequences downstream from the branchpoint are not shown. (*a*) ABC-IE in the Transporter gene of NEPCC29. (*b*) ABC-IE in the Transporter gene of RCC299 (two exact copies in this genome). (*c*) An example D-IE3 from an NADH dehydrogenase subunit that is present in 32 identical copies in CCMP1545, with another 28 copies that have single base changes. The loop is larger than in panels (*a*) and (*b*) and other structures can form but a 5′ ss complementary sequence is present. (*d*) Secondary structure of the Type 1 E2-IE from the CCMP2099 Transporter gene. (*e*) A generalized structure for IE lariats showing the 5′ ss paired with the sequence downstream.

sampling. However, E2-IEs were not detected in metagenomic data from tropical sites or temperate sites similar to where CCMP1646 was isolated, supporting absence from Clade E1 (fig. 2*a*). Indeed, unless completely purged from Clade E1, E2-IEs must have been gained during or after E1 and E2 separation, potentially influencing E2 divergence through genome invasion. In line with these findings, our phylogenetic reconstruction supports separation of Clade E into Clades E1 and E2 (fig. 1*a* and supplementary fig. S1, Supplementary Material online).

Phylogenetic analysis using the concatenated gene sequences generated here provides a more robust assessment of clade composition than most previous studies (supplementary fig. S1, Supplementary Material online). The distinct sequences and insertion positions of IE family members raise possible explanations for how intron invasion might influence development of the *Micromonas* clades. The lineage-specific IE families (table 2, fig. 2 and supplementary figs. S2 and S3, Supplementary Material online) together with insertion polymorphisms of ABC-IEs (e.g., fig. 1*e*) and D-IEs (fig. 3) suggest a heterogeneous landscape in IE distributions across the genomes of extant taxa. Such IE presence–absence polymorphisms could shape diversification by impeding homologous recombination that likely occurs in nature (Worden et al. 2009). The level of CDS divergence among strains in the ABC lineage is low and may not be sufficient to impede sexual reproduction. However, the differential presence of ABC-IEs would have a significant impact on recombination. Divergence could also be influenced by other mechanisms related to differential invasion, for example, differential losses of gene function—the detrimental consequences of faulty insertions (Li, Tucker, et al. 2009); influences on protein evolution and regulatory changes related to alternative splicing; or establishment of new proteins through exon shuffling facilitated by phase-0 insertions (Koonin 2006; Huang et al. 2014). All of these would contribute to development of accessory genome components unique to each *Micromonas* lineage. Thus, although we cannot rule out presence of all IE families in a more ancestral alga that then underwent major differential losses, the more parsimonious explanation of the patterns observed is that the putative founder of each IE family invaded the respective genome at, or shortly after, separation of that clade from other *Micromonas* clades. In this scenario, IE propagation could well have contributed to the expansion and diversification of the *Micromonas* radiation, which shows greater divergence across clades than other Mamiellophyceae genera (fig. 1*a*).

Global IE Family Distributions Point to Bipolar Connectivity

The geographic patterns observed for IE families advance knowledge of global *Micromonas* biogeography (fig. 2*a*). Here, D-IEs and ABC-IEs were frequently observed in the same temperate water samples, and colocation of these clades has been reported in the English Channel and Southern California Bight which are also temperate (Worden 2006; Foulon et al. 2008). D-IEs are present in some low latitude (tropical) metagenomes and CCMP1764 (carrying ABC-IE) was isolated from the tropics. Lack of ABC-IEs in tropical metagenomic data may reflect lower numbers of this IE family across ABC-lineage genomes (contributing to lower frequency and detection in metagenomic data than D-IEs) and/or large differences in relative cell abundances at the time of sampling. Detection of ABC-IEs in Atlantic, Pacific and Indian Ocean metagenomes expands the known range of the *Micromonas* ABC lineage, as being from tropical to high latitude temperate waters just below 60°S (fig. 2*a*).

Importantly, our studies demonstrate that *Micromonas* is present in the Southern Ocean, the circumpolar waters around Antarctica (Morrison et al. 2015). E2-IEs were discovered in the Southern Ocean using query sequences from CCMP2099 which is considered endemic to the Arctic—restricted by geographic and ecological barriers (Lovejoy et al. 2007). The clear separation of Arctic and Antarctic waters, high sequence identity between protein-encoding portions of the Antarctic E2-IE-containing reads and CCMP2099 transcripts, and of E2-IEs themselves (e.g., fig. 2*b* and *c* and supplementary fig. S2, Supplementary Material online), pinpoints E2-IE gains (as well as the split of the Clade E-lineage) to before physical separation of these *Micromonas* populations. If originally endemic to the Arctic (Lovejoy et al. 2007), then how did Clade E2 come to be present in both polar systems, which have "always" been divided by warmer equatorial waters? Apart from ballast water (unlikely based on shipping routes), an effective transport mechanism would be through the NADW which is formed in the Labrador and Nordic Seas, and upwelled in the Southern Ocean (e.g., near sites where E2-IEs were detected below the photic zone). Indeed, although unaware of *Micromonas* presence in the Southern Ocean, Slapeta et al. (2006) suggested that *Micromonas* may be circulated around the world in deep-sea currents at a low metabolic state. Our amplification of E2-IEs and flanking sequence from the NADW demonstrates that Clade E2 *Micromonas*—or their intact DNA—are present in this deep-sea current and could therefore be a source for Antarctic populations, although clearly more comprehensive sampling of deep currents is warranted. It takes approximately 100 years for freshly formed NADW to reach the Southern Ocean (Talley L, Scripps Institution of Oceaongraphy, personal communication). Thus it seems possible that spores or a cell-walled life stage (Worden et al. 2009) may serve as more stable morphotypes during long periods of transport in deep-sea currents.

We also found E2-IEs and flanking sequence corresponding to CCMP2099 transcripts in Antarctica's Ace Lake which is thought to have undergone little change over the past 4,000 years (Fulford-Smith and Sikes 1996). In these samples, salinities ranged down to 22 ppm with temperatures between 0.42 and 1 °C whereas the E2-IE containing Southern Ocean samples had temperatures as low as −1.9 °C and higher salinities (e.g., 33–34 ppm). CCMP2099 grows from 0 to 12 °C and occurs at Arctic sites with salinities from 27 to 34 ppm, but its salinity range has not been experimentally characterized (Lovejoy et al. 2007; Kilias et al. 2014). Our findings extend this range significantly and indicate that considerable salinity reductions will not adversely affect these cells. This is important because *Micromonas* abundance has increased in association with climate-change enhanced ice melt and corresponding salinity reductions (to ~29 ppm) in the Canadian Arctic (Li, McLaughlin, et al. 2009). *Micromonas* is much smaller than the phytoplankton it is replacing, resulting in different food web connections and sinking rates. Hence, further increases in *Micromonas* abundance will likely have major ecosystem consequences.

Intron Stability

Intron phase is hypothesized to play a role in stability, with introns that split a codon being more likely to cause faulty splicing or intron sliding (Lynch 2002). About 50% of IEs in the two most abundant D-IE families are phase-0, the remainder being split between phase-1 and phase-2 (Verhelst et al. 2013). The statistical power of data on E2-IEs and ABC-IEs is lower, but most appear to be phase-0, except E2-IEt2s. Nevertheless, phase-2 IEs were observed for each family and these general patterns may reflect different IE stabilities within the genomes of the various *Micromonas* clades.

In several fungi, introner-like elements (ILEs) have been identified and proposed to degrade into RSIs, thereby serving as a source of RSIs (Torriani et al. 2011; van der Burgt et al. 2012). The average number of total introns per gene (1.4 ± 2) in the four best characterized of these fungi is higher than in CCMP1545 (0.9) or RCC299 (0.6) (Worden et al. 2009; Goodwin et al. 2011; de Wit et al. 2012). However, the number of ILEs per genome (372 ± 180) is akin to ABC-IEs in RCC299 and much lower than D-IEs in CCMP1545. Hence, the balance between RSIs and repetitive introns is different in these fungi than for Clade D (and potentially Clade E2) *Micromonas*. If IEs are degraded into RSIs, then the resulting introns may be less stable, given the low overall RSI numbers in *Micromonas*. Moreover, CCMP1545 RSIs are on average longer than D-IEs. Thus IEs in *M. pusilla* do not seem to fit criteria for degradation to RSIs, suggesting differences from fungal ILEs—the closest analog to IEs reported to date.

An intron-rich LECA has been used to explain occurrence of introns at homologous positions in gene orthologs from divergent eukaryotes (Koonin 2006; Roy and Gilbert 2006). An alternative hypothesis is that some regions, or types of sequence composition, are predisposed to intron-insertion. Li et al. (2009) reported parallel gains at homologous positions in independent allelic lineages of *D. pulex*. In addition to IEs in nonhomologous positions (e.g., fig. 3*a* and *b*), we observed many at homologous positions in different isolates and environmental samples. However, intron phases were sometimes not matched within the homologous codon (e.g., supplementary fig. S4, Supplementary Material online), as also observed for an intron and intein in a relative of *Micromonas*, *Bathycoccus* (Monier et al. 2013). Although these observations are too few to garner statistical support (which would require availability of either genome sequences or more gene homolog sequences from many more isolates), they are suggestive of parallel gains and potential intron insertion "hot spots" as proposed by Li et al. (2009) for *D. pulex*. If LECA derived, the ancestral intron would presumably have been replaced by the D-IE and have undergone differential losses in the other *Micromonas* clades, a less parsimonious scenario than parallel gain. Regardless, the likelihood of either type of scenario relies on the biological mechanism behind intron gains or losses. The fundamental question is: How are IEs propagated across an individual genome?

A Mechanism for IE propagation

Because IEs are always found on the coding strand, new copies have been hypothesized to arise by intron RNA transposition through reverse splicing (Tseng and Cheng 2008, 2013) into an mRNA followed by reverse transcription of the RNA to cDNA (Verhelst et al. 2013) and homologous recombination. At the same time, a well-accepted model for intron removal invokes reverse transcription of spliced mRNA followed by homologous recombination (Fink 1987). This mechanism is thought to account for the strong 5′ position bias in RSIs due to the greater representation of cDNA products arising from the 3′-end of transcripts (Fink 1987). By the same logic it follows that cDNA from a reverse spliced IE RNA would produce a net 3′ bias in the gene position of IEs gained, but this has not been observed. Furthermore, the anticipated amount of cDNA available for these competing gain and loss mechanisms would greatly favor intron removal due to the far greater abundance of spliced mRNA (Rogozin et al. 2012; Yenerall and Zhou 2012), creating a paradox. We propose an alternative hypothesis that avoids this paradox: IE insertion by reverse splicing directly into single-stranded DNA (ssDNA) of R-loops (fig. 5). R-loops are recently appreciated aberrant structures that form behind blocked or stalled RNA polymerase complexes in which the nascent RNA strand pairs back to the underwound DNA template strand behind the RNA polymerase. This displaces an ssDNA loop that can lead to local mutation at the site of R-loops, as well as genome instability (Aguilera and Garcia-Muse 2012; Chan et al. 2014).

We propose that an "armed" spliceosome (Lynch 2002) forms after the spliced mRNA has been released and retains the IE intron RNA. We further postulate that a special sequence or structure of the IE, possibly including the stem-loop in our secondary structure predictions (fig. 4), interferes with spliceosome disassembly and debranching, leading to the persistence of a splicing complex that is primed for reverse splicing. In this model, the ssDNA of the R-loop binds the armed spliceosome in the binding site formerly occupied by spliced mRNA exons, after the completion of the second step of splicing (fig. 5*b*). The 3′OH at the end of the intron lariat (the leaving group in step 2 of forward splicing) becomes the attacking group on a phosphate in the ssDNA in the reverse reaction of step 2. The leaving group in this transesterification is a DNA 3′OH bound in the spliceosome where free exon 1 would be bound between the first and second steps of forward splicing. In a second transesterification, this 3′OH attacks the phosphate at the lariat branch, with the 2′OH of the branch point adenosine (see fig. 4) acting as the leaving group, in a reverse reaction of step 1 of splicing. This inserts the intron RNA cleanly into the ssDNA. As the R-loop is repaired (Aguilera and Garcia-Muse 2012; Chan et al. 2014), either reverse transcriptase or a DNA repair polymerase (Storici et al. 2007) copies the strand containing the RNA into DNA and a new copy of the IE element is incorporated into the genome.

Our model relies on the ability of the spliceosome to catalyze reverse splicing on an ssDNA substrate. Moore and Sharp (1992) substituted the ribose moiety at the end of

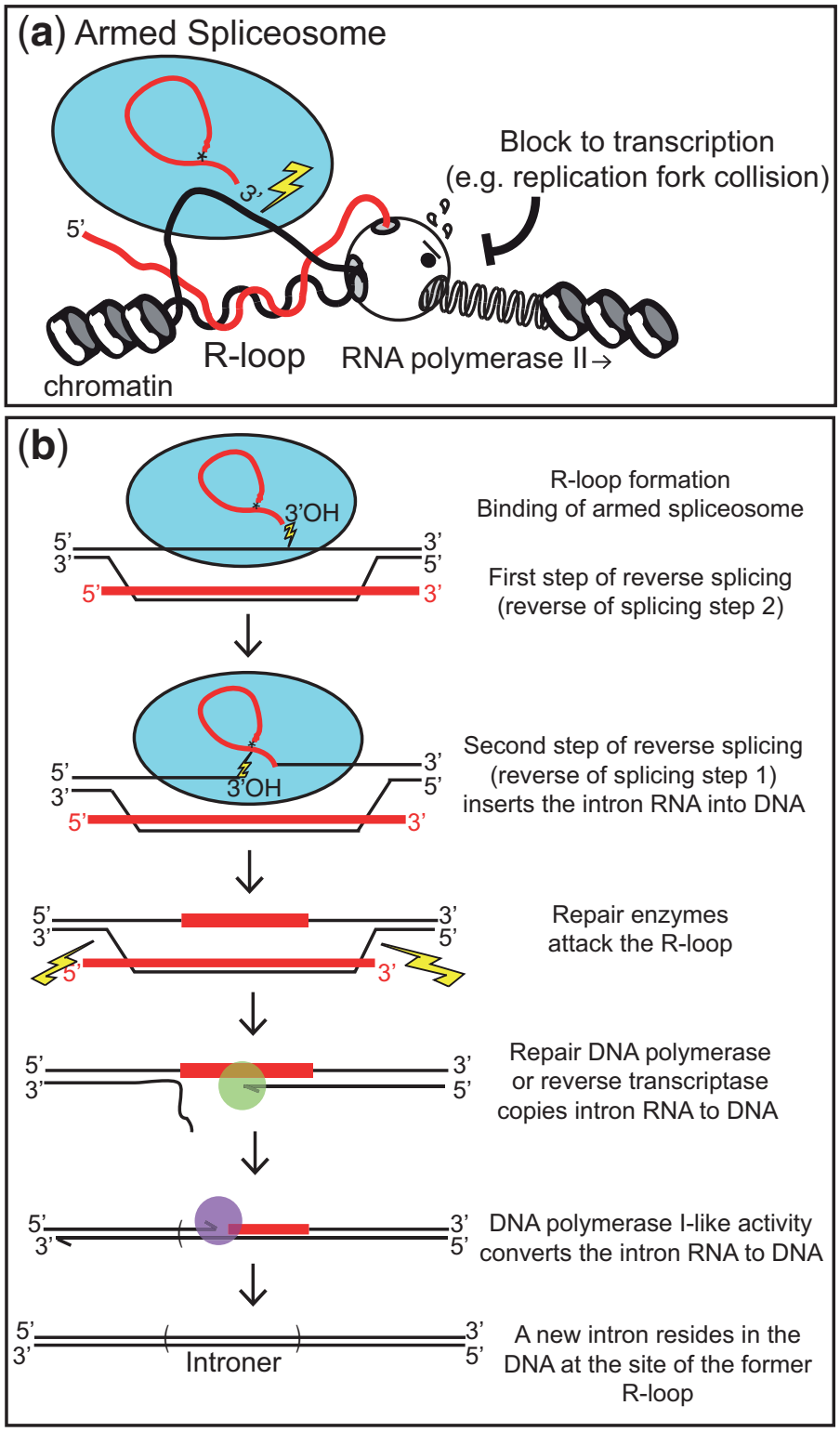

Fig. 5. Proposed model for IE reverse splicing into ssDNA generated at R-loops. (*a*) Diagram of a stalled RNA polymerase II complex behind which an R-loop has formed by pairing of the nascent transcript with the template strand of DNA. A spliceosome that carries the lariat intron product of a recent splicing event binds to the displaced nontemplate DNA strand. RNA (red) and DNA (black) are shown along with nucleosomes (discs) and spliceosome (blue oval). The lightning bolt indicates potential for where the first step of reverse splicing (the reverse of the second step of forward splicing) might occur on the DNA. (*b*) Detailed description of a possible reverse splicing mechanism for IE transposition at R-loops. See text for additional details.

exon 1 in a model pre-mRNA with deoxyribose and found that the rates of the first (where the adjacent 3′OH is the leaving group) and second (where the adjacent 3′OH is the attacking group) steps of forward splicing were not affected. The reverse reactions should be similarly unaffected, suggesting that ssDNA exons should suffice for catalysis of reverse splicing. For comparison, the retrotransposition mechanism of catalytically similar group II self-splicing introns involves

reverse splicing into DNA (Zimmerly et al. 1995; Eskes et al. 2000; Dickson et al. 2001). Although the source of reverse transcriptase activity for repair of the inserted RNA into DNA is uncertain in our model, evidence exists that cellular DNA polymerases can copy an RNA template (Storici et al. 2007).

One prediction of this mechanism is that IE insertions will be biased toward R-loop susceptible locations (Aguilera and Garcia-Muse 2012; Chan et al. 2014), rather than strictly by the transcription rate or cDNA production efficiency predicted by other models (Yenerall and Zhou 2012). This bias should hold for the fungal ILEs as well. A general tendency for R-looped regions to act as targets for intron insertion might explain why intron insertion events appear to occur near each other, but not exactly in the same location (Yenerall and Zhou 2012). R-loops suffer other kinds of mutation, and recruitment of repair proteins to R-loops may incidentally help promote reverse splicing and intron insertion. We envision that transposable IEs may arise spontaneously if the intron RNA sequence evolves so that 1) the intron is refractory to disassembly from the spliceosome and 2) the forward and reverse rates of splicing for that intron become similar. R-looping is intrinsic to transcription, thus IEs may be widespread and appear de novo in any genome.

Conclusions

Intron gains were once considered rare events. Studies based on increased genome-level taxonomic sampling reaching beyond heavily investigated multicellular eukaryotic lineages have revealed major exceptions to this rule. The *Micromonas* species are extraordinary due to the number, variety, and repetitive nature of polymorphic introns comprising unique IE families that trace speciation. After analyzing representatives of different *Micromonas* clades, Pacific Ocean clone libraries, and global metagenomes, we have laid a foundation for future research on the heterogeneity and functional implications of the *Micromonas* IE landscape. We hypothesize that invasion of distinct IE families facilitated the divergence of extant *Micromonas* lineages from their last common ancestor. This could have occurred through IE-influenced processes, including impedance of homologous recombination, differential gene losses, and protein innovations resulting in gain of new functions. Our R-loop based model for IE proliferation is generalizable to the majority of eukaryotes, thus as genome sequences become available for a greater diversity of eukaryotes, we anticipate discoveries of other rampant invasions by repetitive intronic elements. Together with analyses on potential functions of repetitive introns, such studies will provide a more comprehensive view on intron gain and its influence on the eukaryotic tree of life.

Materials and Methods

Culturing and Nucleic Acid Extraction

Ten *Micromonas* isolates were grown at 200 photon m^{-2} s^{-1} PAR (measured using a QSL2101 light meter; Biospherical Instruments Inc., San Diego CA) on a 14-h/10-h light/dark cycle (table 1). These were obtained immediately prior to the study from several culture collections (or for RCC299 and CCMP1545 from in-house) and grown in standard media and conditions (supplementary table S1, Supplementary Material online). Additionally, *Micromonas* RCC434, RCC472, and RCC1614 were grown to improve 18S rRNA gene sequence availability for three clades. Cells were harvested by centrifugation at 6,000 or 8,000 × g, the supernatant removed immediately and pelleted cells frozen at −80 °C until extraction. DNA was extracted using a QIAGEN DNeasy Kit (Germantown, MD) according to the manufacturer's instructions except for CCMP1764 which was extracted using a protocol for genome quality DNA (http://www.mbari.org/phyto-genome/Resources.html, last accessed June 8, 2015). Environmental samples were collected near the end of the Scripps Institution for Oceanography pier (32°53′N, 117°15′W) in April and October 2001 and extracted as part of a previous study (Worden 2006).

PCR, Cloning, and Sequencing

PCR primers were designed to conserved regions of four gene homologs found in the genomes of *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Micromonas* sp. RCC299, *M. pusilla* CCMP1545, and spanning IE in the latter (supplementary table S2, Supplementary Material online). Accessions for these in CCMP1545 are: Actin, XM_003061058.1; ATPase, XM_003062703.1; Transporter, XM_003060502.1; Dehydrogenase, XM_003058664.1. The three latter genes are single copy in the genome, whereas the Actin primers were specific to one of several related copies. In addition, 18S rRNA gene primers (18SEUKF: 5′-ACCTGGTTGATCCTGCCAG-3′; 18SEUKR: 5′-TGATCCTTCYGCAGGTTCAC-3′) were used to verify isolate identity as in Worden et al. (2004). DNA from each isolate was amplified in individual reactions for each of the five genes (i.e., including the 18S rRNA). Specifically, 25 µl PCR reactions consisted of 9 µl nuclease free water, 12.5 µl HotStar Master Mix (Qiagen), 500 nM each of forward and reverse primers, and 1 µl of DNA. For negative controls, an additional 1 µl of nuclease free water was used in place of DNA. PCR conditions were as follows: 30–32 cycles at 94 °C for 30 s, annealing for 30 s (see supplementary table S2, Supplementary Material online, for temperatures), extension at 72 °C for 2 min, preceded by 15-min initial denaturation at 95 °C, and followed by 10-min extension at 72 °C.

For cultures, products for the different genes were amplified separately from each of the cultures and run on a 1% agarose gel. The majority showed a single band and was purified using the QIAquick PCR Kit. For those with multiple bands (starting cultures contained bacterial contaminants which primer design did not account for), specifically the ATPase of CCMP490, CCMP1195, CS222, NEPCC29, the Transporter of RCC472, and the NADH dehydrogenase of CCMP490, NEPCC29, PCR products were excised from the gels and purified using the QIAquick Gel Extraction Kit (Qiagen). Actin, the only eukaryote specific gene investigated, had single bands for all cultures. The PCR products from each culture and gene were then independently cloned using the

TOPO TA Cloning kit (Life Technologies, Carlsbad, CA). Insert lengths ranged from 422 to 1,256 nt (supplementary tables S2 and S3, Supplementary Material online). For each culture, 2–16 colonies were picked and plasmids purified using the QIAprep Miniprep Kit. The plasmids were sequenced bidirectionally on an ABI 3100 using BigDye terminator v3.1 chemistry (Life Technologies) with M13F (5′-CTGGCCGTCGTTTTAC-3′) and M13R (5′-CAGGAAACAGCTATGAC-3′). Additional sequencing primers were used for internal regions of the 18S rDNA, 502F (5′-GGAGGGCAAGTCTGGT-3′) and EUK1174R: (5′-CCCGTGTTGAGTCAAA-3′).

For environmental samples a different approach was used for investigating potential CCMP2099 presence in NADW (Atlantic Ocean samples) than for the environmental clone libraries used to investigate IE diversity (Pacific Ocean samples, see below). For the former, primers (ABC.E2F: GGCGAACCAGCAACAACGAGAAG; ABC.E2R GCTTCGTCCTGGAGTTTCGCC) were designed to specifically amplify a 200-bp region spanning the first E2-IEt1 of the CCMP2099 Transporter (fig. 1*e*). Twenty-five µl PCR reactions consisted of 9.5 µl nuclease free water, 12.5 µl Qiagen HotStar Master Mix, 500 nM each of forward and reverse primers, and 1.5 µl of the extracted DNA template, the positive control (CCMP2099 DNA) or negative control (nuclease free water). PCR was carried out under the following conditions: 35 cycles of 94 °C for 30 s, annealing at 56.5 °C for 30 s, extension at 72 °C for 90 s, preceded by 15-min denaturation at 95 °C, and followed by extension at 72 °C (10 min). PCR products were purified using the QIAquick PCR Kit (Qiagen) and cloned using the TOPO TA Cloning kit (Life Technologies) following the instructions provided by the manufacturer. For each PCR product, 4–16 colonies were picked and cloned inserts amplified with the vector primers M13F and M13R. Inserts were sequenced unidirectionally using M13F for all clones, and bidirectionally (M13F and M13R) for 3,000-m clones on an Applied Biosystems Hitachi 3500 xL Genetic Analyzer using BigDye terminator v3.1 chemistry (Life Technologies). Clone reads were assembled in geneious v8.1 with manual curation. To validate presence of eukaryotic DNA in these samples, PCR was performed using the 18S rRNA gene primers (as above) and products verified by size on a gel.

PCR was also performed on North Pacific environmental samples using the ATPase and Actin primer sets (supplementary table S2, Supplementary Material online) as above. The reactions were performed independently for the spring and fall templates collected in Worden (2006). Two libraries were constructed for each of two genes from the spring and fall samples because for each gene amplification two different bands (one reflecting intron-less and one reflecting intron-containing sequences) were excised from a 1% agarose gene and independently cloned after clean up (i.e., eight total; results were attained for seven libraries because one non-IE bearing product was lost in processing). In total, 96 colonies were picked per library and plasmids purified according to the methods of Davis (1986) prior to sequencing using a 3730xl DNA Analyzer (Life Technologies).

Reads from each clone were assembled using DNAStar (Lasergene) and manual curation. Nonprasinophyte sequences were removed from environmental clone libraries based on an initial BLASTx and BLASTn (Altschul et al. 1997) evaluation against NCBI's nonredundant database and an in-house database of publically available genomes. CCMP1545 and CCMP490 ATPase Clusters O and P (one clone per strain) each had a one nucleotide difference from the CCMP1545 genome sequence (Worden et al. 2009) and PCR-derived sequences from this study (across the entire amplicon, including IE sequence); these were considered PCR artifacts and not analyzed further. CCMP1764 DNA was sequenced using the 454-FLX platform.

Clustering, RSI Identities, and Phylogenetics

Genomic DNA and cDNA sequences were aligned using ClustalW, or manually in DNAStar or Bioedit. Clustering of environmental clones was performed using BLASTClust (Altschul et al. 1990) with required coverage specified by both a similarity threshold of 100% and minimum length coverage of 1.0. Pairwise intron nucleotide identities were computed using Emboss water (employing the Smith–Waterman algorithm), unless otherwise specified. Sequence logos were constructed using WebLogo (Crooks et al. 2004) after manual curation of the insertion sequences alignments. RSIs identities were calculated for introns in the β-tubulin gene because sequences exist for all cultured *Micromonas* clades and three introns are present. Two of these represent two different homologous RSIs (termed 5′ and 3′ here) for which RSI-locus comparisons show 73% (5′ RSI) and 51% (3′ RSI) nucleotide identity between NEPCC29 (Clade C) and RCC299 (Clade A). RSIs at the same loci in Clades D, E1 or E2 have less than 50% identity to those in the other clades. BLASTn queries of the RCC299 and CCMP1545 β-tubulin RSIs to their respective genome sequences attain only self-hits and pairwise alignment of these sequential β-tubulin RSIs renders identities less than 50% within each strain.

For the 18S rRNA gene phylogeny, we retrieved nearly complete (>1,500 bp) 18S rDNA from Mamiellophyceae and prasinophyte sister clades from NCBI and added those generated herein (see above). Sequences were aligned using MAFFT (Katoh et al. 2005). Regions of unambiguous alignment were identified using MUST (Philippe 1993) and all gap-containing positions removed, except for ten positions (corresponding to nucleotides 645–655 in the *Micromonas pusilla* CCMP1545 sequence #AY954994) that help resolve *Micromonas* clade differences. Phylogenetic reconstructions were statistically evaluated using Bayesian inference (BI) and maximum-likelihood (ML) methods from 1,646 homologous positions. The GTR + Γ + I was used as the model of nucleotide substitution for both analyses. Phylogenetic analyses were calculated using MrBayes 3.232 for BI (Ronquist et al. 2012) and Treefinder for ML (Jobb et al. 2004). Bayesian analyses were performed with two independent runs and 1,000,000 generations per run. After a burn in of 350,000 trees per run, the remaining trees were used to reconstruct a consensus tree and to get posterior probabilities for node supports.

Bootstrap values were calculated using 1,000 replicates with the same substitution model.

For the ATPase phylogenetic analyses, introns and IEs were removed from the nucleotide sequences. Then, sequences from cultures and representative environmental sequences were aligned using MAFFT (Katoh et al. 2005). Regions with unambiguous alignment were identified using MUST (Philippe 1993), and all gap-containing positions were removed. A ML phylogeny was built from 734 homologous nucleotide positions using the TVM + G model including relaxing parameters of first, second, and third codon positions. The model was selected using Modeltest (Posada and Crandall 1998) as implemented in Treefinder (Jobb et al. 2004). The Dehydrogenase, Transporter, and Actin genes were analyzed similarly but using only sequences from cultures and MMETSP data from the same strains (Keeling et al. 2014) to gain full length information. Four resulting alignments (including the ATPase) were concatenated. CCMP490 sequences were partial and the missing data were considered as missing entries in the matrix. An ML tree was constructed from 4,612 homologous positions (in the alignments of these four genes) using the same evolution model as the ATPase gene. Bootstrap statistics were performed using 1,000 ML replicates for all these phylogenies.

Metagenome Searches

Metagenomic searches (fig. 2*a*) were performed using D-IE1.1 and the D-IE from the Transporter (PID68853) to represent Clade D elements, the Transporter-located IE in strain NEPCC29 to represent ABC-IEs, and finally E2-IE Type 1 and Type 2 sequences from the CCMP2099 Transporter, as queries in BLASTn, implemented in CAMERA (Sun et al. 2011) using the CAMERA "all metagenomic 454" data set as of March 1, 2014. The complexity filter was off and only hits with *E* value $< 10^{-5}$ were returned, those with IE typically ranged from 10^{-7} to 10^{-100}. Only metagenomic reads with flanking sequence on either side of the "hit" alignment region were further characterized. Sequences were verified as being IE-like through alignment and used as BLASTn queries against the CCMP2099 and NEPCC29 transcriptomes (McRose et al. 2014). Nucleotide identities were typically 99% between metagenomic flanking sequence and transcripts between Antarctic sequences and CCMP2099 as well as between IEC taken from model RCC299 Mipur011i11380 (Verhelst et al. 2013) and the NEPCC29 Transporter.

To confirm identities between Antarctic CDS and CCMP2099 transcripts, the CCMP2099 Transporter E2-IEt1 was reblasted against 400 metagenomic sequences, and the resulting hits were searched using MEME (Bailey and Elkan 1994) to find a common 50 nt motif (the IEs are longer but length variation is typically associated with the start of the polyU-stretch). The motif was then used to search all 400 sequences to find 402 hits in 396 sequences. E2-IEt1s (identified as G[CT]N[2-7 nt]—motif—N[7-37 nt]AG) were then excised from these sequences, and 21 sequences removed from the analysis because although the motif was present, the IE was not complete. The remaining ($n = 375$) read segments were used as BLASTn queries against CCMP2099, RCC299, and CCMP1545 sequences. All best hits were to the CCMP2099 transcriptome. The 360 best hits (15 had no hit) were used to compute nucleotide identities between CCMP2099 and protein-encoding portions of the Antarctic metagenomic reads.

Searches for RNA Structure

IE sequences were submitted to the mFold server http://mfold.rna.albany.edu/?q=mfold (last accessed June 8, 2015) and the mFold output (Zuker 2003) for different IEs was evaluated by inspection. Several structures of decreasing stability were evaluated for each IE. The most frequent common feature of folding for several IEs was the stem loop occupying the 5′ splice site sequence shown in figure 4.

Supplementary Material

Supplementary material, tables S1—S5, figures S1—S7, and datafile S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank H.M. Wilcox, E. Demir-Hilton, and D. McRose for assistance. They are grateful to G. Dick and S. Jain for blasting IEs against their deep sea metagenomes (negative results) and also to J. Sarmiento and L. Talley for kindly discussing NADW formation and movement with us. Sequences generated in this study have been deposited in NCBI database under accessions KR089059–KR089061 for 18S rRNA gene sequences from *Micromonas* RCC434, RCC472, and RCC1614; KR089139–KR089205 for the four genes studied using PCR amplification; representative (non-redundant) sequences KR089062–KR089138 (ATPase) and KR089206–KR089345 (Actin) from environmental clone libraries from the Eastern Pacific; KR152644–KR152649 for the 3,000-m NADW Transporter clones; and genomic DNA 454-FLX reads from CCMP1764 have been deposited in CAMERA under project CAM_PROJ_CCMP1764. M.A. was supported by NIH grant GM040478. This research was supported by the David and Lucile Packard Foundation, a Gordon and Betty Moore Foundation Investigator Award (GBMF3788), NSF-IOS0843119, and DOE-DE-SC0004765 grants to A.Z.W.

References

Aguilera A, Garcia-Muse T. 2012. R loops: from transcription byproducts to threats to genome stability. *Mol Cell*. 46:115–124.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.

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

Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, et al. 2010. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* 22:2943–2955.

Broecker WS. 1991. The great ocean conveyor. *Oceanography* 4:79–89.
Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol.* 16:107–113.
Chan YA, Hieter P, Stirling PC. 2014. Mechanisms of genome instability induced by RNA-processing defects. *Trends Genet.* 30:245–253.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.
Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol.* 7:e1002150.
Curtis BA, Archibald JM. 2010. A spliceosomal intron of mitochondrial DNA origin. *Curr Biol.* 20:R919–R920.
Davis LG. 1986. Plasmid "Mini-Prep" method. In: Davis LG, Dibner MD, Battey JF, editors. Basic methods in molecular biology. Elsevier Science Publishing Co, Inc, New York. p. 102–104.
de Wit PJ, van der Burgt A, Okmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, et al. 2012. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genet.* 8:e1003088.
Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Canestro C, et al. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330:1381–1385.
Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc Natl Acad Sci U S A.* 98:13207–13212.
Eskes R, Liu L, Ma HW, Chao MY, Dickson L, Lambowitz AM, Perlman PS. 2000. Multiple homing pathways used by yeast mitochondrial group II introns. *Mol Cell Biol.* 20:8432–8446.
Fink GR. 1987. Pseudogenes in yeast? *Cell* 49:5–6.
Foulon E, Not F, Jalabert F, Cariou T, Massana R, Simon N. 2008. Ecological niche partitioning in the picoplanktonic green alga *Micromonas pusilla*: evidence from environmental surveys using phylogenetic probes. *Environ Microbiol.* 10:2433–2443.
Fulford-Smith SP, Sikes EL. 1996. The evolution of Ace Lake, Antarctica, determined from sedimentary diatom assemblages. *Palaeogeogr Palaeoclimatol Palaeoecol.* 124:73–86.
Gilbert W. 1978. Why genes in pieces? *Nature* 271:501.
Goodwin SB, M'Barek SB, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A, et al. 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 7:e1002070.
Huang S, Chen Z, Yan X, Yu T, Huang G, Yan Q, Pontarotti PA, Zhao H, Li J, Yang P, et al. 2014. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat Commun.* 5:5896.
Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol.* 4:18.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.
Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12:e1001889.
Kilias ES, Nöthig E-M, Wolf C, Metfies K. 2014. Picoeukaryote plankton composition off West Spitsbergen at the entrance to the Arctic Ocean. *J Euk Microbiol.* 61:569–579.
Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct.* 1:22.
Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in *Daphnia* populations. *Science* 326:1260–1262.
Li WKW, McLaughlin FA, Lovejoy C, Carmack EC. 2009. Smallest algae thrive as the Arctic Ocean freshens. *Science* 326:539.
Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad Sci U S A.* 99:8121–8126.
Lovejoy C, Vincent WF, Bonilla S, Roy S, Martineau MJ, Terrado R, Potvin M, Massana R, Pedros-Alio C. 2007. Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *J Phycol.* 43:78–89.
Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A.* 99:6118–6123.
Marin B, Melkonian M. 2010. Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. *Protist* 161:304–336.
McRose D, Guo J, Monier A, Sudek S, Wilken S, Yan S, Mock T, Archibald JM, Begley TP, Reyes-Prieto A, et al. 2014. Alternatives to vitamin B1 uptake revealed with discovery of riboswitches in multiple marine eukaryotic lineages. *ISME J.* 8:2517–2529.
Modrek B, Lee C. 2002. A genomic view of alternative splicing. *Nat Genet.* 30:13–19.
Molnar P. 2008. Closing of the Central American Seaway and the Ice Age: A critical review. *Paleooceanogr.* 23: PA2201.
Monier A, Sudek S, Fast NM, Worden AZ. 2013. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* 7:1764–1774.
Moore MJ, Sharp PA. 1992. Site-specific modification of pre-mRNA: the 2′-hydroxyl groups at the splice sites. *Science* 256:992–997.
Morozov EG, Demidov AN, Tarakanov RY, Zenk W. 2010. Deep water masses of the South and North Atlantic. In: Abyssal channels in the Atlantic Ocean: water structure and flows. New York, Springer. p. 266.
Morrison AK, Frölicher TL, Sarmiento JL. 2015. Upwelling in the Southern Ocean. *Physics Today* 68:27–32.
Parra G, Bradnam K, Rose AB, Korf I. 2011. Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. *Nucleic Acids Res.* 39:5328–5337.
Philippe H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* 21: 5264–5272.
Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biol Direct.* 7:11.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.
Roy SW. 2003. Recent evidence for the exon theory of genes. *Genetica* 118:251–266.
Roy SW. 2006. Intron-rich ancestors. *Trends Genet.* 22:468–471.
Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet.* 7:211–221.
Slapeta J, Lopez-Garcia P, Moreira D. 2006. Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Mol Biol Evol.* 23:23–29.
Storici F, Bebenek K, Kunkel TA, Gordenin DA, Resnick MA. 2007. RNA-templated DNA repair. *Nature* 447:338–341.
Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, et al. 2011. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.* 39:D546–D551.
Sverdlov SV, Rogozin IB, Babenko VN, Koonin EV. 2007. Conservation versus parallel gains in intron evolution. *Nucleic Acids Res.* 33:1741–1748.

Talley LD. 2013. Closure of the global overturning circulation through the Indian, Pacific, and Southern Oceans: schematics and transports. *Oceanography* 26:80–97.

Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D. 2011. Evidence for extensive recent intron transposition in closely related fungi. *Curr Biol.* 21:2017–2022.

Tseng CK, Cheng SC. 2008. Both catalytic steps of nuclear pre-mRNA splicing are reversible. *Science* 320:1782–1784.

Tseng CK, Cheng SC. 2013. The spliceosome catalyzes debranching in competition with reverse of the first chemical reaction. *RNA* 19:971–981.

van der Burgt A, Severing E, de Wit PJ, Collemare J. 2012. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Curr Biol.* 22:1260–1265.

Verhelst B, Van de Peer Y, Rouze P. 2013. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biol Evol.* 5:2393–2401.

Worden AZ. 2006. Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat Microb Ecol.* 43:165–175.

Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324:268–272.

Worden AZ, Nolan JK, Palenik B. 2004. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnol Oceanogr.* 49:168–179.

Yenerall P, Zhou L. 2012. Identifying the mechanisms of intron gain: progress and trends. *Biol Direct.* 7:29.

Zimmerly S, Guo H, Eskest R, Yang J, Perlman PS, Lambowitz AM. 1995. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* 83:529–538.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.