Molecular Biology and Evolution

Dynamics of Actin Evolution in Dinoflagellates

Sunju Kim,†,[1] Tsvetan R. Bachvaroff,*,[1] Sara M. Handy,‡,[2] and Charles F. Delwiche[2,3]

[1]Smithsonian Environmental Research Center, Edgewater, Maryland

[2]Department of Cell Biology and Molecular Genetics

[3]The Maryland Agricultural Experiment Station, University of Maryland

†Present address: Department of Life Science, Gongju National University, Gongju, Chungnam, Republic of Korea.

‡Present address: U.S. FDA Center for Food Safety and Applied Nutrition, College Park, Maryland.

***Corresponding author:** E-mail: bachvarofft@si.edu

Associate editor: Andrew Roger

Abstract

Dinoflagellates have unique nuclei and intriguing genome characteristics with very high DNA content making complete genome sequencing difficult. In dinoflagellates, many genes are found in multicopy gene families, but the processes involved in the establishment and maintenance of these gene families are poorly understood. Understanding the dynamics of gene family evolution in dinoflagellates requires comparisons at different evolutionary scales. Studies of closely related species provide fine-scale information relative to species divergence, whereas comparisons of more distantly related species provides broad context. We selected the actin gene family as a highly expressed conserved gene previously studied in dinoflagellates. Of the 142 sequences determined in this study, 103 were from the two closely related species, *Dinophysis acuminata* and *D. caudata*, including full length and partial cDNA sequences as well as partial genomic amplicons. For these two *Dinophysis* species, at least three types of sequences could be identified. Most copies (79%) were relatively similar and in nucleotide trees, the sequences formed two bushy clades corresponding to the two species. In comparisons within species, only eight to ten nucleotide differences were found between these copies. The two remaining types formed clades containing sequences from both species. One type included the most similar sequences in between-species comparisons with as few as 12 nucleotide differences between species. The second type included the most divergent sequences in comparisons between and within species with up to 93 nucleotide differences between sequences. In all the sequences, most variation occurred in synonymous sites or the 5′ UnTranslated Region (UTR), although there was still limited amino acid variation between most sequences. Several potential pseudogenes were found (approximately 10% of all sequences depending on species) with incomplete open reading frames due to frameshifts or early stop codons. Overall, variation in the actin gene family fits best with the "birth and death" model of evolution based on recent duplications, pseudogenes, and incomplete lineage sorting. Divergence between species was similar to variation within species, so that actin may be too conserved to be useful for phylogenetic estimation of closely related species.

Key words: actin, birth-and-death evolution, dinoflagellate, phylogeny, *Dinophysis*, dinokaryon.

Research article

Introduction

Dinoflagellates are important primary producers in the oceans and are infamous for creating toxins. They also have a proclivity for endosymbiosis and have adopted plastids of almost every major pigment type (Delwiche 1999; Schnepf and Elbrächter 1999). The dinoflagellate genus *Dinophysis* is a good example of both traits because *Dinophysis* species produce diarrheic shellfish poison, and some species seem to have adopted plastids from cryptophyte algae (Schnepf and Elbrächter 1988). The *Dinophysis* plastid has been fairly well studied based on pigment analysis, ultrastructure, plastid gene phylogeny, and finally by culturing on cryptophtye fed ciliates (Lucas and Vesk 1990; Takishita et al. 2002; Park et al. 2006).

Dinophysis species also contain the characteristic dinoflagellate nucleus, so-called dinokaryon (Spector 1984). The dinokaryotic nucleus differs from other eukaryotes because it is packed with massive condensed chromosomes, visible under light microscopy throughout the cell cycle. When examined with transmission electron microscopy, these chromosomes contain banded DNA fibrils without the nucleosomal-histone DNA packaging typical of most other eukaryotes. Dinoflagellate nuclei are also notable because of large genome size with some dinoflagellates containing hundreds of picograms of DNA per nucleus (100 pg $\approx 10^{11}$ bases), and the DNA contains a significant fraction of the modified base 5-hydroxymethyluracil in place of thymine (Rae 1976).

Although genome size makes complete sequencing difficult, cDNA sequencing has been done for a few dinoflagellate species (Bachvaroff et al. 2004; Hackett et al. 2004; Lidie et al. 2005; Patron et al. 2005, 2006). Analysis of highly expressed genes and their genomic complements suggests that some genes are in large multicopy gene families often with subtle nucleotide and amino acid variation between copies (Rowan et al. 1996; Reichman et al. 2003; Bachvaroff et al. 2004, 2009; Zhang et al. 2006 ; Bachvaroff and Place 2008; Moustafa et al. 2010). However, the evolution of the multicopy gene families in dinoflagellates is not well

understood, with possible mechanisms underlying the diversity including whole-genome duplication, concerted evolution, and single gene duplication. In general, gene family evolution can be fit to two competing models: concerted evolution versus the birth and death model.

Under concerted evolution, gene duplication and revision are driven by large-scale events at the level of whole genes, such as gene conversion via crossing over and recombination, as reviewed by Nei and Rooney (2005). These events are predicted to occur within and between the populations of gene copies in each cell. Therefore, changes between gene copies should occur at the scale of a whole gene and tend to co-occur and spread by gene conversion across many gene copies in "concert." Concerted evolution helps to explain distinct lineage sorting, particularly of neutral sites such as ribosomal RNA (rRNA) Internal Transcribed Spacer regions between sibling species, and fits the data well when large numbers of similar or identical gene copies are maintained in the genome. In contrast, under the birth and death model, after "birth" (or duplication) individual gene copies are expected to vary independently across the genome and differences between copies can accumulate at individual nucleotide sites. After many changes accumulate, some gene copies can lose proper function, become pseudogenes, and effectively "die." Birth and death models explain observations such as incomplete lineage sorting, pseudogenes, and changes between gene copies primarily in synonymous or neutral sites.

Here, we examine variation in the actin multicopy gene family in two closely related *Dinophysis* species to better understand the evolution of gene families in dinoflagellates and to assess the utility of this molecule for phylogenetic studies. Actin is a conserved and ubiquitously expressed multicopy gene that has already been deeply sequenced in another dinoflagellate (Bachvaroff and Place 2008), and we here generated small data sets from another five dinoflagellate species as well as larger data sets from two *Dinophysis* species. These latter two species, *Dinophysis acuminata* and *D. caudata* are morphologically distinct but closely related with 99.4% similar small subunit ribosomal DNA sequences (Handy et al. 2009). Of particular interest in this context was the question of whether the different actin gene copies would be clearly distinct between the two closely related species and change in a concerted manner within each species. Alternately, if sequences from the two species would intermingle in phylogenetic trees incomplete lineage sorting and a birth and death mode of evolution would be favored (Rooney 2004). The results provide insight into the tempo and mode of gene family evolution in *Dinophysis* and other dinoflagellates.

Materials and Methods

Sample Collection and Cultures

Dinophysis acuminata and *D. caudata* cell cultures were generously donated by Prof. Myung G. Park at Chonnam National University in the Republic of Korea. *Dinophysis acuminata* and *D. caudata* cells were taken a week after being fed on the prey ciliate *Myrionecta rubra*, which in turn had been raised on cryptophyte prey (Park et al. 2006), and harvested by centrifugation at 1,200 × g for 10 min. The pellets were stored in 1.5 ml microfuge tubes containing RNA*later* (Ambion AM7020, Austin, TX) at −80 °C for 1 week. The cells were washed with Tris–ethylenediaminetetraacetic acid buffer (10 mM Tris-HCl pH 8.0, 1 mM ethylenediaminetetraacetic acid [EDTA]) four times to remove RNA*later* using centrifugation and aspiration of the supernatant to exchange the washing fluid.

Dinophysis caudata cells collected with a 30 μm plankton net from Ft. Pierce Inlet Pier (FPIP), Florida, in the United States (27°27′550″N, 80°19′067″W) were individually isolated using capillary pipette, washed six times with 0.45 μm filtered seawater, and filtered onto 5 μm Millipore polycarbonate filters. The filters were placed into 2.0 ml centrifuge tubes containing a nonionic detergent solution (Galuzzi et al. 2004) and stored at −20 or 4 °C until processed. *Blastodinium crassum* infecting the copepod *Paracalanus parvus* was isolated from the waters off of La Paz in Baja California South, Mexico (Coats et al. 2008). The dinoflagellate was preserved in nonacid Lugol's solution (4% w/v iodine + 6% w/v potassium iodide) and stored at 4 °C until processed. The Lugol's preserved cells were washed six times with distilled water and placed into a 1.5 ml microfuge tube with 100 μl distilled water.

Gymnodinium catenatum, *Karlodinium veneficum*, and *Katodinium rotundatum* were grown in natural seawater with a salinity of 15 supplemented with f/2 nutrients at 20 °C with light:dark cycle of 14:10 under approximately 100 $\mu Ein/m^2/s^1$. Fifty to hundred milliliter volumes of exponential phase cultures (approximately 10,000 cells/ml) were harvested by centrifugation in a clinical centrifuge at 3,000 × g for 10 min. The resulting pellets were stored at −80 °C until processed.

Extraction of RNA and Genomic DNA

Total RNA and genomic DNA of *D. acuminata* and *D. caudata* were extracted from cell pellets using TRIzol Reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The cells were lysed in Trizol, chloroform was added, and the aqueous phase was removed for RNA and the interphase retained for DNA extraction (see below). The RNA was precipitated from the aqueous phase with two volumes of isopropanol, and the RNA pellet was further purified using lithium precipitation. One-fifth volume of 12 M LiCl was added to the RNA and the mixture was incubated at −20 °C for 30 min, pelleted by centrifugation, washed with 70% ethanol, and resuspended in RNAase free water. The DNA was isolated from the Trizol aqueous:organic interphase. The remaining aqueous and organic phases were completely removed and DNA was precipitated from the aqueous:organic interphase with two volumes of ethanol. The isolated DNA pellet was washed four times with a solution of 0.1 M trisodium citrate and 10% ethanol, followed by a single wash with 75% ethanol. The DNA-pellet was resuspended in 8 mM NaOH adjusted to pH 8.4 using 4-(2-HydroxyEthyl)-1-PiperazineEthaneSulfonic acid

(free acid). The quantity and quality of DNA and RNA were determined using a spectrophotometer (NanoDrop, ND-1000, Thermo Scientific). When the DNA quality was low based on 260:280 nm absorbance ratios, an additional chloroform purification and ethanol precipitation were done.

For the uncultured dinoflagellates *D. caudata* FPIP and *B. crassum*, the preserved cells were washed in deionized fresh water, briefly sonicated using a probe tipped sonicator (Heat Systems Ultrasonic, Plain view, NY) as previously described in Handy et al. (2009), and the resulting sonicate was used as polymerase chain reaction (PCR) template. Genomic DNA from cultures of *G. catenatum*, *Kar. veneficum*, and *Kat. rotundatum* was extracted using Cetyltrimethylammonium bromide (CTAB) buffer (100 mM Tris-HCL pH 8.0, 0.7 M NaCl, 2% [w/v] CTAB [Sigma, St Louis, MO], 20 mM EDTA) (Doyle and Doyle 1987). Each cell pellet was resuspended in 1 ml of CTAB buffer and incubated at 50 °C for 10 min. The DNAs were purified by a single extraction with chloroform, precipitated with isopropanol, washed with 70% ethanol, and resuspended in 50–100 μl of water.

Reverse Transcription of RNA and PCR

Total RNA was reverse transcribed with a poly T primer, RTpr (5′-CGAATTGTCGACTAGTACTTTTTTTTTTTTTT-TT-3′) using the AccuScript High Fidelity Reverse Transcriptase-PCR System (Stratagene, La Jolla, CA) according to the manufacturer's instructions but modified by skipping 65 °C incubation prior to the synthesis step at 42 °C and adding 1 μl (40 U) of RNase OUT (Invitrogen, Carlsbad, CA). The additional sequence before the polyT sequence in the primer provided a 3′ UTR priming site.

PCR Amplification

The quality of cDNA was then checked using a PCR with the spliced leader primer SL1 (5′-TCCGTAGCCATTTTG-GCTCAA-3′) and the reverse transcription primer NDTRN that is included in the polyT reverse transcription primer (5′-CGAATTGTCGACTAGTACTTT-3′). The spliced leader is a conserved sequence found at the 5′ end of a diverse array of dinoflagellate mRNAs (Lidie and Van Dolah 2007; Zhang et al. 2007). The PCR conditions were initial denaturation at 94 °C for 2 min; 35 cycles of 94 °C for 30 s, 57 °C for 30 s, and 72 °C for 2 min; and final extension at 72 °C for 5 min. Reactions were run in total volume of 20 μl containing 500 mg/ml bovine serum albumin (Sigma A2053), 50 mM Tris-HCl (pH 8.3), 3 mM Mg, 10 μM deoxyribonuclotides, and 0.12 units of Promega Go-Taq. Two microliters of the first-strand cDNA reaction were added as template. Agarose gel electrophoresis followed by ethidium bromide staining was used to see if reverse transcription and amplification were successful.

Partial actin cDNA (approximately 774 nt) was amplified with the forward primer ACT AF2 (5′-ATGACKCAGA-TYATGTTYGA-3′) and reverse primer ACT OR1 (5′-TCA-GAAGCACTTCCTGTGCAC-3′) (fig.1) using 2 μl of the first stand of cDNA reaction in the same PCR programs above. The complete coding region (approximately 1,199–1,248 nt) was amplified from cDNA with the general spliced leader primer SL1 and gene-specific reverse primer OR1 (fig. 1) using the same PCR program as above. The same actin primers and PCR conditions were used to amplify genomic versions of actin using either DNA purified from cultures (50 ng of DNA added) or sonicated cells (4 μl of sonicate) as template.

The PCR products were visualized with agarose gel electrophoresis followed by ethidium bromide staining. Most PCR products were purified using PolyEthylene Glycol (PEG) precipitation (20% w/v PEG, mw 8000, 2.5 M NaCl solution) (Morgan and Soltis 1995). When multiple bands were present, only correctly sized products were purified with a QIAEX II agarose gel extraction kit (QIAGEN, Valencia, CA). Purified PCR products were cloned with the pGEM-T Easy Vector and competent cells (Promega, Madison, WI). White colonies were randomly picked for PCR with M13 vector primers (as above) and sequenced with Big Dye Terminator Cycle Sequencing Kit version 3.1 (Applied Biosystems, Foster, CA) on an ABI 3730 sequencer.

Sequence Analyses and Alignments

Sequencher v.4.8 (Genecodes, Ann Arbor, MI) was used to trim vector and ambiguous sequences and to assemble bi-directional sequencing reads for each clone. The resulting nucleotide sequences were verified by Blast searches in National Center for Biotechnology Information (NCBI) and screened for potential chimeric sequences with the Bellerophon program (Huber et al. 2004).

The actin sequences generated in this study were aligned with the actin sequences from GenBank using MacClade v.4.08 (Maddison and Maddison 2002). The nucleotide alignment contained 277 sequences and was trimmed to 774 nucleotide comparable positions. Four separate data sets were used for phylogenetic analysis: one, *D. acuminata* alone; two, *D. caudata* alone; three, *D.acuminata* and *D. caudata* together; and four, all eight dinoflagellate species. The aligned nucleotide data set was also translated into amino acids (257 amino acids) with MacClade. Pairwise distances between gene copies were calculated with PAUP* v. 4b10. Different taxon categories were extracted using perl scripts. Histograms of the distances were calculated with IGOR Pro v. 5.04B software (Oregon) and graphed using Sigma Plot 11 (Systat software, San Jose, CA). The Guanine and Cytosine (GC) content of all positions and third codon positions were analyzed with codonw (http://mobyle.pasteur.fr/cgi-bin/postal.py?form=codonw). The number of synonymous substitutions per synonymous site (d_S) and nonsynonymous substitutions per nonsynonymous site (d_N) (Nei and Gojobori 1986) were estimated using PAML3.14 (Yang et al. 1997).

The optimal trees were found from the nucleotide and amino acid data sets with the program RAxML 7.0.4 (Stamatakis 2006) under the optimal GTR $+ I + \Gamma$ model for nucleotide alignments and the PROTGAMMAJTT model for amino acid alignments with 100 bootstrap replicates in both cases.

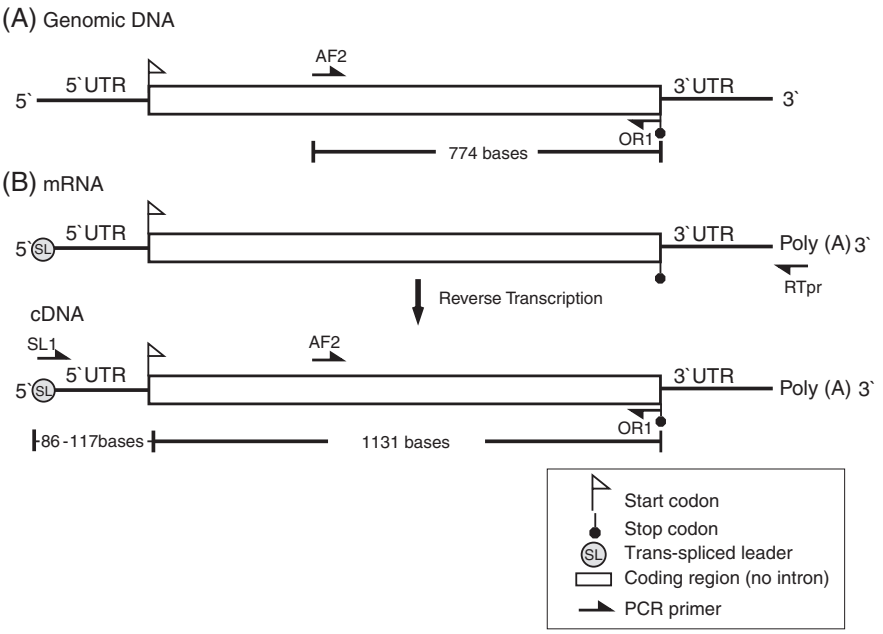

Fig. 1. Actin gene amplification schematic. The relative positions of the different primers used in the study are placed on a model of the actin gene. The genomic and mRNA versions of the gene are shown.

Results

Novel Dinoflagellate Actin Sequences

A total of 142 partial and complete actin sequences (approximately 774 and 1131 nt, respectively) were determined from five different dinoflagellate species, of which 128 were unique sequences (table 1 and supplementary table S1, Supplementary Material online): 11 from *B. crassum* (HQ391454–HQ391464), 47 from *D. acuminata* (HQ391359–HQ391384 and HQ391405–HQ391420), 56 from *D. caudata* (HQ391385–HQ391404 and HQ391421–HQ391453), 8 from *G. catenatum* (HQ391465–HQ391472), and 20 from *Kat. rotundatum* of which 14 were unique (HQ391473–HQ391486). One 275 base intron was found when aligning the genomic and cDNA sequences in the beginning of a *Kat. rotundatum* actin genomic amplicon (HQ391477).

Rarefaction Analysis of Sequences

Rarefaction analysis of the sequences cloned from both *Dinophysis* species suggested that redundancy between clones was low because the rarefaction curve was still in a near linear stage with 41 unique sequences of 47 for *D. acuminata* and 53 unique sequences of 56 for *D. caudata*. Errors due to amplification are likely to be approximately one error per 774 base amplicon (Eun 1996). The rarefaction analysis was then repeated assuming that sequences differing by less than 2 bases in pairwise comparisons were the same. Under this assumption, there were 23 of 47 unique sequences for *D. acuminata* and 43 of 56 for *D. caudata* still suggesting that redundancy was low. Increasing the number of allowable differences stepwise from 1–13 nucleotides only showed a notable decrease in the number of unique sequences between four and five nucleotides (fig. 2).

Table 1. The Number of Actin Gene Sequences in This Study, GC content, The Average of nonsynonymous (d_N) and synonymous (d_S) Substitution Rates between Actin Gene Copies in Dinoflagellate Species.

Species	Genomic	Partial cDNA	Trans-Spliced cDNA	GC Content[a]	$d_N \pm$ SE[b]	$d_S \pm$ SE[b]	d_N/d_S[b]
Amphidinium carterae[c]	111			0.54 (0.66)	0.004 ± 0.0000	0.103 ± 0.0013	0.091 ± 0.0021
Blastodinium crassum	11			0.58 (0.80)	0.033 ± 0.0035	0.532 ± 0.0446	0.132 ± 0.0301
Dinophysis acuminata	21 (16)	14	12	0.59 (0.82)	0.007 ± 0.0002	0.090 ± 0.0043	0.232 ± 0.0074
D. caudata	36 (33)	7	13	0.61 (0.87)	0.009 ± 0.0001	0.211 ± 0.0051	0.157 ± 0.0054
Gymnodinium catenatum	8			0.59 (0.71)	0.048 ± 0.0050	0.173 ± 0.0160	0.306 ± 0.0243
Karenia brevis[c]	4			0.53 (0.60)	0.024 ± 0.0097	0.431 ± 0.1469	0.044 ± 0.0088
Karlodinium veneficum[c]	19			0.52 (0.58)	0.004 ± 0.0002	0.280 ± 0.0068	0.013 ± 0.0009
Katodinium rotundatum	20 (14)			0.63 (0.90)	0.004 ± 0.0002	0.034 ± 0.0065	0.419 ± 0.0400

[a] The mean of overall GC content with that of third codon GC content in a parentheses were estimated based on partial actin-coding region (774 nt).
[b] Synonymous (d_S) and nonsynonymous (d_N)substitutions per site and its ratio (d_N/d_S) were shown as mean ± standard error (SE) and excluding infinite values, where is $d_S = 0$. Note that the average d_N and d_S values for the two *Dinophysis* species do not reflect the complicated distributions shown in figure 3.
[C] These multiple gene copies were accessed from GenBank.

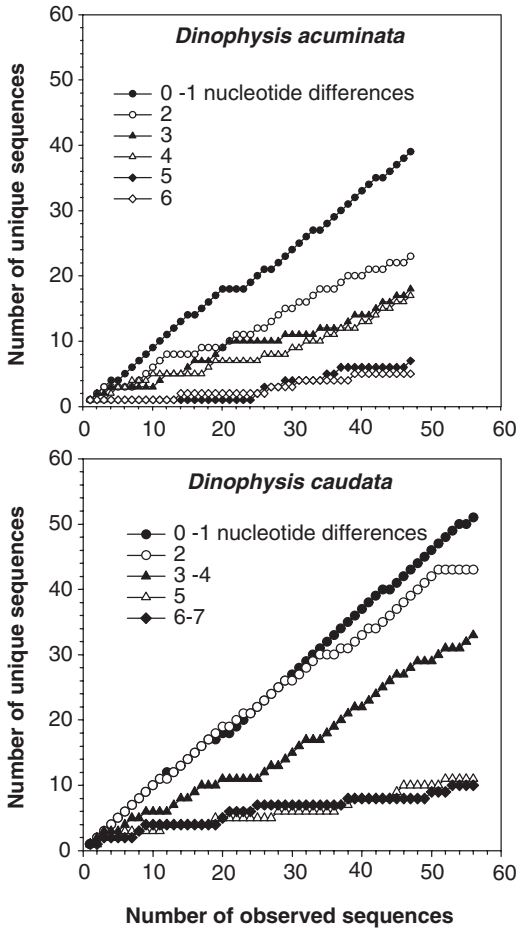

FIG. 2. Rarefaction analysis of the actin gene sequences from the two *Dinophysis* species. Individual clones from each species were compared allowing increasing numbers of differences between sequences. The number of observed sequences was plotted against the number of unique sequences.

Bellerophon identified 11 possible chimeric sequences for *D. caudata* and none from *D. acuminata*. Of the putative chimeric sequences from *D. caudata*, seven were derived from the same pair of parents, and two others were derived from one of these two parent sequences. None of the chimeras was an exact match to either parent sequence on the two sides of the putative break point, suggesting that chimera check might not accurately recover potential chimeras in actin gene families, as has been noted by Lahr and Katz (2009). Therefore, all sequences were retained for further analysis.

Gene Copy Diversity within Species

Pairwise nucleotide and amino acid differences between gene copies within each of the two *Dinophysis* species were calculated. In order to compare different size amplicons, the nucleotide alignment was trimmed to 774 bases (fig. 1). The pairwise nucleotide differences between different genomic actin copies from *D. caudata* showed a bimodal distribution with two distinct peaks around 10 and 52 nucleotides, whereas the full-length and partial cDNA sequences exhibited a less even distribution up to 92 nucleotide differences (fig. 3*A*). Similarly, the *D. acuminata* actin sequences had a bimodal pattern with two peaks around 8 and 59 nucleotide differences (fig. 3*B*). Despite the substantial differences between gene copies in nucleotide comparisons, translation into amino acids reduced the differences to a single peak centered around three to five amino acids for both species (fig. 3*C* and *D*).

The full-length spliced leader amplified cDNA sequences included the nontranslated 5′ UTR excluded from the analysis above because one of the primers was designed from the 5′ spliced leader (fig. 1). In *D. caudata*, the 5′ UTR region varied from 64 to 96 bases with a total of six sequence types differing in sequence and length (data not shown). In *D. acuminata*, all the 5′ UTR were 56 bases long, and all but one had very similar sequence (0–2 bases of differences). Clustering based on 5′ UTR length and sequence correlated well with placement on the trees described below using the partial coding region.

Gene Copy Diversity between Species

When comparing sequences between the two species in phylogenetic trees, there were three major groups, ranked here by the amount of divergence between species. First, the majority of sequences appeared to be recently duplicated and were termed type 1. These bushy clusters of sequences formed two clades corresponding to the two species. In between-species comparisons, the sequences differed by on average 55 nucleotides or 7.1% (both median and mode) between the two clades. Although the majority of sequences were tightly clustered with each other, these sequences were not the most closely related in comparisons between species (fig. 4). The type 2 sequences were the most similar copies in comparisons between species. These sequences differed by as few as 12 nucleotides between the two species (1.5% of 774 bases) and represented only a small fraction of sequences recovered (9/103). Finally, the type 3 sequences were a few very divergent copies (12/103), mostly from *D. caudata* (9/12) that differed from *D. acuminata* by up to 86 bases.

The three types of sequences were defined on the basis of between-species comparisons but were similarly distributed into three categories on the distance histograms comparing sequences within species (fig. 3*E* and *F*). For example, the labeled colored branches a and d from figure 4 can be compared with the distributions on figure 3*E* and *F*. Similarly, the type 2 and 3 sequences could be mapped onto specific branches and compared with distributions of d_N and d_S values in the within-species comparisons.

The distribution of sequences across the distance histograms (fig. 3) and tree (fig.4) seemed to be more influenced by the PCR template (cDNA vs. genomic DNA) rather than primers (gene-specific vs. general spliced leader forward primer) used for amplification. Most type 3 or divergent sequences were recovered from cDNA template, with exactly half of these sequences coming from the two different primer sets. Also, the type 2 or most similar sequences in between sequence comparisons were overwhelmingly drawn from genomic DNA (8/9). Pseudogenes were com-

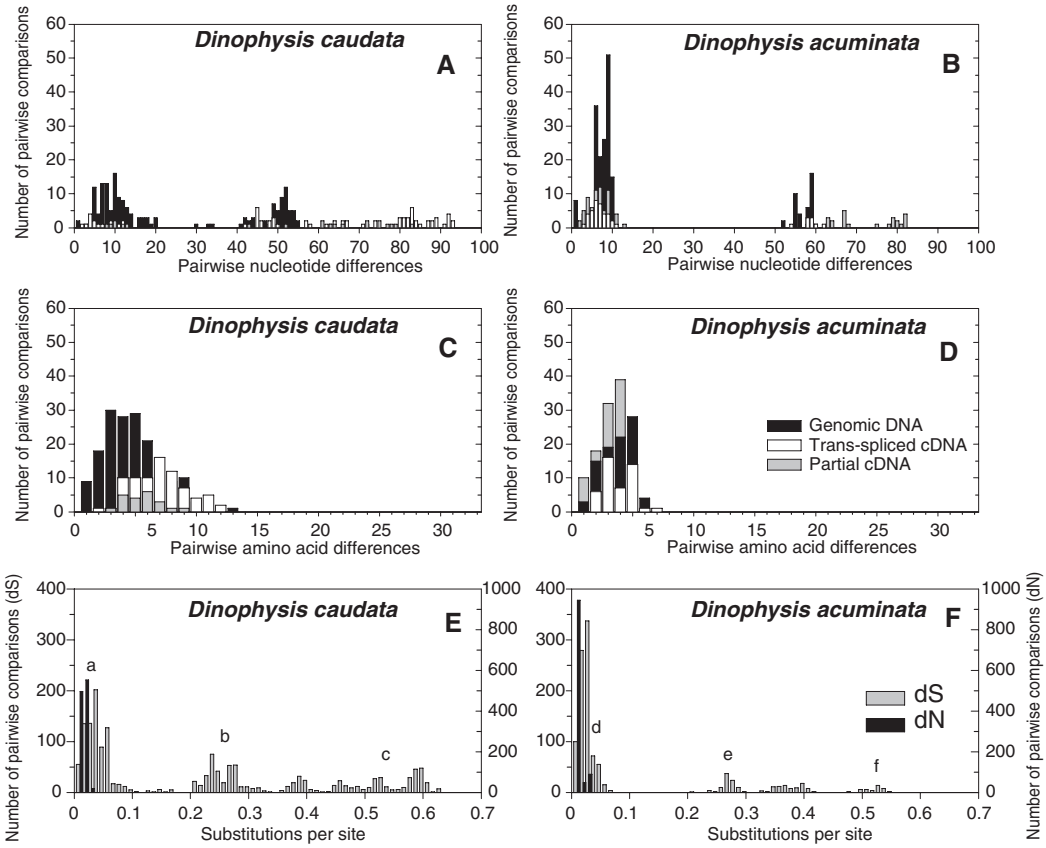

Fig. 3. Histograms showing pairwise comparisons of gene copies within the two *Dinophysis* species. These histograms show the distribution of nucleotide and amino acid differences when comparing different amplicons within the two *Dinophysis* species. Each different primer and template combination were compared including partial genomic and cDNA amplicons with gene-specific primers and full-length amplicons from cDNA. The top panels (*A* and *B*) depict nucleotide differences when comparing different gene copies. The middle panels (*C* and *D*) depict amino acid differences, and the bottom panels (*E* and *F*) depict synonymous and nonsynonymous substitution rates per site. The left (*A*, *C*, and *E*) and the right panels (*B*, *D*, and *F*) are the gene copies amplified from *D. caudata* and *D. acuminata*, respectively. The lowercase letters (a-f) in panels *E* and *F* refer to specific branch lengths in figure 4.

mon in types 1 and 2 (9/81 and 3/9, respectively) but were not found in the divergent type 3 group.

Another feature of the sequence data was differences between the species in GC content particularly in the third codon position (GC3). *Dinophysis caudata* was more GC rich (overall 60%, GC3 86%) than *D. acuminata* (59% and 81%). Dividing the sequences by type, the bushy type I sequences from *D. caudata* (60% and 85%) and *D. acuminata* (59% and 81%) were less GC biased than the type 2, similar sequences (62% and 87%), and the type 3 divergent sequences (62% and 90%) (supplemental fig. S2, Supplementary Material online). When third codon positions were excluded, the overall ML tree was similar to figure 3, although bootstrap support was lower: the two type 1 single species clades were well supported (99% bootstrap for *D. acuminata* and 76% for *D. caudata*) as was the type 2 clade (92%), and the long branch type 3 clade was present but not supported.

Phylogenetic Analyses of Dinoflagellate Actin Genes

Maximum likelihood trees were made using a nucleotide alignment from the 142 actin gene copies from this study and an additional 135 dinoflagellate sequences from NCBI (fig. 5). Except for the two closely related *Dinophysis* species, the sequences from each species formed monophyletic groups with bootstrap support $\geq$70%. The type 1 conserved, tightly clustered sequences with short branch lengths from *D. acuminata* and *D. caudata* again formed two distinct well-supported monophyletic groups. However, the type 2 and 3 sequences from the two species showed interspecific clustering of putatively orthologous sequences, as was seen in the tree containing only *Dinophysis* species (fig. 3). Trees made from the translated alignment also contained intermingled sequences from the two *Dinophysis* species, and the two fucoxanthin-containing species, *K. veneficum* and *K. brevis*, were nested within the *Dinophysis* clade but not with strong bootstrap support (supplementary fig. S2, Supplementary Material online). The remaining species were recovered as monophyletic clades.

Two more distantly related species, *D. caudata* and *A. carterae*, were selected for comparison of raw pairwise nucleotide and amino acid differences. These distributions showed that nucleotide differences between species were larger than within species, but that amino acid differences between species clearly overlapped with the variation

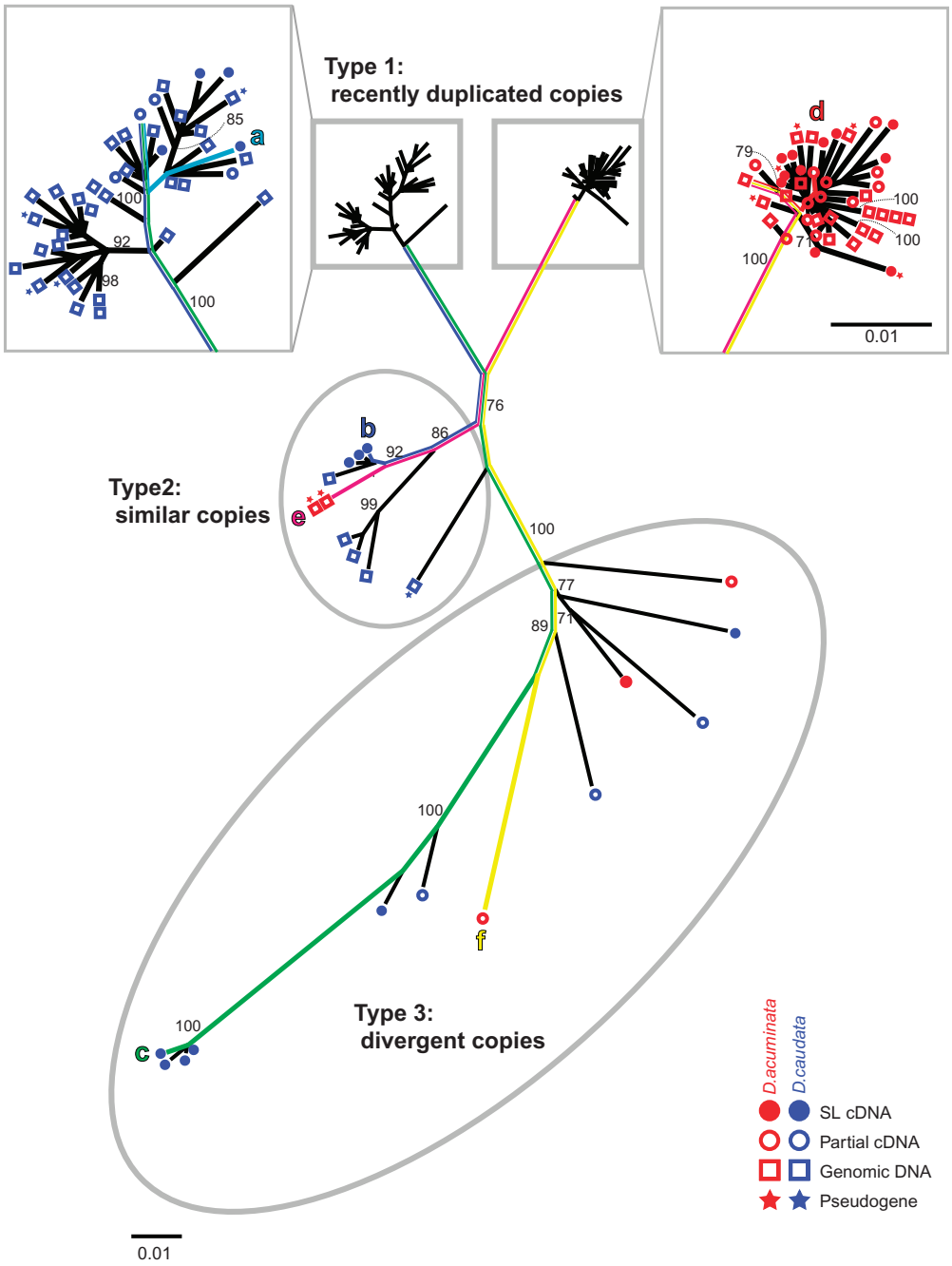

FIG. 4. Unrooted maximum likelihood tree of two *Dinophysis* species using a nucleotide alignment. The most likely unrooted tree using 103 actin gene copies from the two *Dinophysis* species is shown with selected bootstrap values (>70%) shown. The tree can be divided into three categories based on comparisons between the two species. Most copies form tight clusters of similar sequences (top) that vary by eight to ten nucleotide differences (type 1). In the middle are the type 2 sequences that are most similar in between-species comparisons. At the bottom are sequences that are dramatically divergent in two species comparisons or the type 3 divergent copies. The different types of amplicons and pseudogenes are shown with symbols, blue and red colors refer to *D. caudata* and *D. acuminata*, respectively (see supplementary table S1, Supplementary Material online, for more details). The lowercase letters on colored branches refer to specific branches corresponding to comparisons within species shown in figure 3.

within a species (fig. 5, inset). The average pairwise distance for *A. carterae* and *D. caudata* was nine amino acids but within-species distributions centered around three or five amino acids. Branch lengths in the amino acid tree were quite short, as reflected by the overlapping distributions in raw pairwise distances.

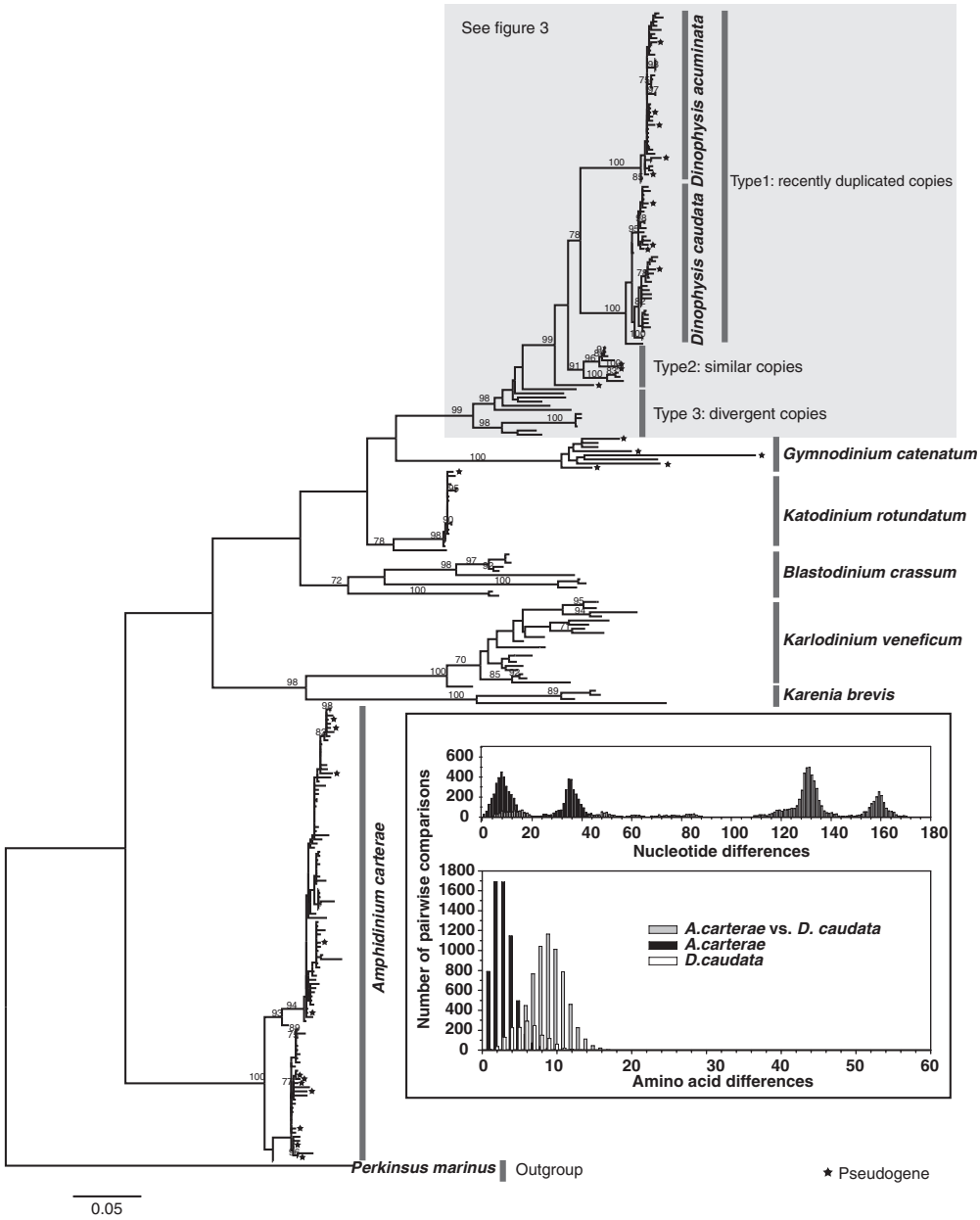

Fig. 5. Maximum likelihood nucleotide tree of actin genes in eight dinoflagellate species. The most likely tree using 276 sequences from eight species with a 774 nucleotide alignment is shown. The three types of *Dinophysis* sequences are bracketed. Inset into the figure are histograms showing raw pairwise nucleotide and amino acid differences within and between *Amphidinium carterae* and *Dinophysis caudata*.

Gene Copies from Other Species

When sequences from the six other species were examined, similar trends of within-species nucleotide changes were seen with synonymous differences exceeding nonsynonymous changes (table 1). Here, we used the d_S (synonymous substitution rate) and d_N (nonsynonymous substitution rate) measures calculated in the context of a phylogenetic tree. Overall, the mean values of d_S between copies in a species were much greater than those of d_N with values ranging from 0.034 to 0.532 (fig 3*E* and *F*). The d_N values were 1 order of magnitude lower than d_S and ranged from 0.004 to 0.048. *Gymnodinium catenatum*, *B. crassum*, and *A. cartarae* genomic amplicons had sequences with frameshifts that we interpret as pseudogenes (five of eight sequences for *G. catenatum*) and showed relatively high d_S values (fig. 5 and table 1). The corresponding d_N values for these species were also large. In contrast, the highly conserved *Kat. rotundatum* sequences had the lowest d_S value, but the d_S value in this species was still an order of magnitude greater than the d_N value.

The pseudogene sequences of *G. catenatum* were found on long branches and contained substantial amino acid substitutions (fig. 5 and supplementary fig. S2, Supplementary Material online). *Blastodinium crassum* was also highly divergent and appeared to accumulate amino acid substitution changes. In contrast, the highly diverged nucleotide sequences of *Kar. veneficum* were mostly

identical after translation into amino acids with only very few amino acid differences.

Discussion

Dinoflagellate Nuclear Genomics

The nuclear genome of dinoflagellates remains among the most poorly understood of any other major eukaryote lineage. With genome sizes ranging around 100 pg, direct complete genomic sequencing and assembly would challenge even the most modern sequencing methods and deepest pockets. However, Expressed Sequence Tag (EST) sequencing from dinoflagellates has been a popular tool to bypass the daunting genome (Bachvaroff et al. 2004; Hackett et al. 2004; Lidie et al. 2005; Patron et al. 2005, 2006), and at least one directed study has compared cDNA sequences with their genomic counterparts (Bachvaroff and Place 2008). From these sequencing efforts, a rough model can be described for dinoflagellate genomes. Many, but not all genes were found to have variation between gene copies, often with most differences found in synonymous sites (Reichman et al. 2003; Zhang et al. 2006; Bachvaroff and Place 2008; Bachvaroff et al. 2009). Here, we use a targeted approach to examine the evolution of a representative dinoflagellate gene family. Two closely related but morphologically distinct species were selected, and a conserved multicopy gene family, actin, was used to understand the mode and tempo of gene amplification in dinoflagellate nuclear genomes.

Gene Duplication in Dinoflagellates Compared with Related Genomes

Sequencing multiple versions of actin from different dinoflagellate species clearly demonstrates the abundance and diversity of copies in the genome, especially when compared with genomes from other alveolates. In genomic study of the related apicomplexa, only a few gene families were multicopy (Brayton et al. 2007). In the third alveolate lineage, the ciliates, three rounds of whole genome duplication have been described in the ciliate *Paramecium tetraurelia* (Aury et al. 2006), and gene duplication and diversification in the ciliate somatic genome appear to have evolved independently in at least two ciliate lineages (Robinson and Katz 2007). By contrast, the dinoflagellate *Dinophysis* genome is large and has multiple actin gene copies, but dinoflagellates do not appear to maintain somatic and germinal nuclei, as do ciliates.

The Actin Gene Complement in Other Genomes

Actin is often a multicopy gene. However, comparison of the multicopy actin genes in dinoflagellates with the actin gene complement in completely sequenced genomes suggests distinct differences in the dinoflagellate duplication pattern. In multicellular animals, the actin gene family phylogeny corresponds well with distinct functional categories into, for example, cytosolic, smooth, or cardiac muscle actins (OOta and Saitou 1999). More broadly, many eukaryotes have a complement of more divergent actin-related proteins (arp1–11), each with distinct functions (Muller et al. 2005).

As an example of actin gene diversity, we can look to a recent study of the actin gene family from the complete genome of the "slime mold" *Dictyostelium discoideum*. The total complement of actin and arp genes in *D. discoideum* is quite large, with 41 different genes, of which eight have been identified as arp genes (Joseph et al. 2008). However, of the 29 canonical actins from *D. discoideum*, only 21 are contiguously alignable without gaps, and 17 encode identical proteins. The remaining "orphans" are divergent actin genes relative to canonical actins but not clearly related to an arp. Expression is overwhelmingly drawn from the pool of identical amino acid copies, and four of the additional actin copies may not be expressed (Joseph et al. 2008). Thus, in *D. discoideum*, both expression and conservation of amino acid sequence act together to restrict actin variation in the proteome.

In dinoflagellates, by contrast, almost all canonical actin protein sequences vary subtly from copy to copy, and there seems to be no correlation of conservation and expression. In the species that have been examined, (here the two *Dinophysis* species and previously in *A. carterae*; Bachvaroff and Place 2008), expression from the pool of total genomic actin gene copies included variant gene copies, sometimes even apparent pseudogenes. However, the amino acid variation has a limit: the sequences can be readily aligned without gaps, and the distribution of amino acid differences was clearly restricted, suggesting that we are only examining canonical actins. Clearly, deeper proteomic and subcellular localization methods would be required to prove similarity or difference in function, but the apparent continuous distribution of amino acid differences suggests that we are studying a single population of copies without distinct functional differences. The rarefaction analysis suggests that the sequences presented here are likely subsamples of the overall set of copies, so it is likely that many more actin gene copies are present in some dinoflagellates than *D. discoideum*. Clearly, deep EST sequencing and more difficult contiguous genome sequencing would be required to fully sample both conserved and divergent actin gene copies in dinoflagellates as well as to sample the actin-related proteins.

To investigate the possibility that some unknown amplification, sequencing, or cloning artifact is at least partially responsible for our observations of actin gene diversity, we did rarefaction curves across a broad range of sequence differences from 0 to 7 base differences. Sequencing or amplification error estimates would be on average less than 0.1%, or up to 2 bases in pairwise comparisons, and most sequences differed much more. The rarefaction curves showed substantial diversity between sequences and did not collapse until sequences with 4 or more base differences were assumed to be identical (fig. 2). Furthermore, the sequence differences were mostly synonymous; clearly, sequencing or amplification error would not lead to such a pattern. The degree of variation differed from taxon to taxon, with less variation observed in *Kat. rotundatum* and more in *B. crassum*. Overall, our sampling method,

although not unbiased, appears to represent the underlying features of the different genomes we sampled. Chimerism, or PCR recombination, is another possible source of error, but in this case, we would expect to see sequences intermediate between the two parent sequences in phylogenetic trees (Bachvaroff and Place 2008; Lahr and Katz 2009). Thus, chimerism would not explain the observed divergent sequences in our trees.

Other Gene Families in Dinoflagellates

The pattern of actin gene family evolution probably applies to other dinoflagellate genes. Here, we used the actin gene as a typical highly expressed gene or as an example for describing dinoflagellate genome evolution. In the present study, we have shown a diverse multicopy actin gene family in five additional dinoflagellate species, and this pattern may also be true for other highly expressed dinoflagellate genes (Bachvaroff and Place 2008; Moustafa et al. 2010). The pattern we describe here of multicopy divergent gene families seems to apply well to a class of dinoflagellate genes including the peridinin chlorophyll protein (PCP) gene in *Symbiodinium* sp. (Reichman et al. 2003), the proliferating cell nuclear antigen gene in *Pfiesteria piscicida* (Zhang et al. 2006), and perhaps other gene families in dinoflagellates including rubisco (Rowan et al. 1996; Zhang and Lin 2003; Moustafa et al. 2010). Many plastid-associated genes such as *psbO* and light harvesting complex (LHC) genes were also found in multicopy variant families (Bachvaroff et al. 2004). Interestingly, data from the syndinian dinoflagellate *Amoebophrya* sp. did not find multiple divergent gene copies for a small set of genes (Bachvaroff et al. 2009). Thus, the dinoflagellate gene family pattern described here may be restricted to the dinokaryotes or those dinoflagellates with full-fledged dinokaryotic chromosome organization. In this context, *Oxyrrhis marinus*, a free-living heterotrophic dinoflagellate that does not have a clear affinity either with dinokaryotic or syndinian dinoflagellates (Saldarriaga et al. 2003) also has multiple actin copies (Sano and Kato 2009).

Many of these multicopy genes, including actin in *A. carterae*, and *Kar. veneficum*, were found in tandem arrays of slightly varying copies, a feature we were unable to confirm using PCR strategies in these two *Dinophysis* species (data not shown). Another example would be the PCP gene in *Lingulodinium polyedra* that seems to form a homogenous repeat array of thousands of identical copies (Le et al. 1997). Repeats of individual protein units within a single transcript appear to also be common, as was found for rubisco (Rowan et al. 1996; Zhang and Lin 2003), LHC (Hiller et al. 1995), and luciferase genes (Li et al. 1997).

Mode and Tempo of Gene Family Evolution

Comparisons of the two closely related *Dinophysis* species illuminate the dynamics of dinoflagellate gene family evolution. Actin gene family evolution in dinoflagellates likely represents a pattern typical of some other dinoflagellate genes, and comparisons of two closely related species should expose the dynamics of gene family evolution more clearly than studies of a single species could. The two species allow the different gene copies to be calibrated in time relative to a speciation event. In this context, we selected two clearly distinct but closely related species based on cell form. Our reference frame is relative to the unknown divergence time of the two species, not units of absolute time.

Strikingly, pairwise distances among sequences do not show a monotonic increase in distance, but rather segregate into three distinct distance types, suggesting episodic diversification of sequences. The different actin gene copies seem to be drawn from three different populations based on between-species divergence. These three types of sequences appear to have diverged from each other at different time points during and after speciation. First, the most commonly found sequences were recently duplicated after speciation and seem to have diverged consistently in both species (termed here "type 1"). Variation within the pool of recently duplicated copies is quite similar in both species and was centered about eight to ten nucleotides in within-species comparisons. The majority of sequences from both species were sampled from this pool suggesting recent parallel amplification in both species of a specific gene copy.

A second category of sequences, termed here "type 2" included those sequences most similar in comparisons between the species. Although these copies could potentially be the result of gene transfer between species either via lateral transfer or gene flow between species, their similarity likely represents the small divergence between these gene copies. More likely these copies were shared between the two species right up until species divergence.

Finally, there is a set of very divergent sequences, termed here "type 3." These copies seem to represent a different class of actin gene copies, likely present in the most recent common ancestor of the two species, and either subject to faster relative mutation rate or longer divergence time than the other gene copies. These sequences do not have any clear functional difference from canonical actins because most differences are only present at the nucleotide level. Some of the divergent cDNA sequences from *D. caudata* do appear to form the tail of the distributions on when comparing amino differences (fig. 3C), suggesting potential phenotypic differences.

The highly conserved nature of the actin amino acid sequence helps to confirm that orthologous sequences, or inparalogs, are being compared, as was demonstrated in the phylogenetic trees. Thus, most of the variation is neutral or near neutral, in terms both of selection and of phenotype.

Models of Gene Family Evolution in Dinoflagellates

Taken together, the three actin gene types indicated incomplete lineage sorting during the divergence of the two *Dinophysis* species and inconsistent divergence time or rate of change between gene copies. From the presence of three catergories of actin sequences, we infer that multiple rounds of gene duplication events have occurred and that the most recent common ancestor likely contained

multiple paralogous actin gene copies. This is a special case of incomplete lineage sorting, which has been described as parologous lineage sorting (Rooney and Ward 2008). The data suggest discrete saltatory gene amplification events rather than a continuous production of new gene copies, although deeper sampling might produce a more continuous distribution of differences. Also apparent from our data is parallel selection in both species of a particular set of gene copies for duplication. We conclude that the most recent common ancestor of these two species likely contained a diverse suite of actin gene copies and that gene family expansion and contraction have continued during and after the two species become isolated.

Overall, the number and diversity of gene copies, the accumulation of neutral mutations, the presence of many pseudogenes, and the pattern of incomplete lineage sorting fit well with the birth and death model of gene family evolution, especially when compared with the concerted evolution model (Nei and Rooney 2005). The data suggest that new gene copies are produced over time and that divergence between the gene copies seems to happen independently. Incomplete lineage sorting is considered to be a strong indicator of birth and death evolution (Rooney 2004). However, we have to keep in mind that in dinoflagellates, the evolution of gene families seems to apply broadly to more genes, with larger numbers of copies than seen in other eukaryotic gene families although apparently not as a byproduct of whole genome duplication. Also the apparent concerted evolution of synonymous sites in actin gene copies in the dinoflagellate *A. carterae*, and the identity of the PCP gene family in *L. polyedra* argue for differences in specific lineages and genes. This suggests that a mixed model with some features of concerted evolution, or a birth and death model with strong purifying selection coupled with rapid gene turnover may best explain these data (Rooney and Ward 2005). Certainly, deeper study of dinoflagellate populations including several genes from multiple closely related species would provide deeper insight.

In some sense, these multicopy gene families can operate as a population within an individual. Each individual can contain a specific and dynamic population of gene copies, creating an additional hierarchy to consider in the context of standard population genetics models.

Implications for Phylogeny

Amino acid–based phylogenetic trees using multiple actin gene copies showed that actin may not be useful for estimating phylogeny of closely related dinoflagellate species. Because most variation between actin copies within a species was synonymous, translation into amino acids would eliminate most within-species variability. However, in the case of actin, amino acid variation between distantly related species overlaps with variation within a species (fig. 5), and the two closely related *Dinophysis* species were intermingled in amino acid trees (supplementary fig. S2, Supplementary Material online). Thus, actin may be too conserved to be useful for phylogeny when comparing closely related species within the dinokaryotes. In general, phylogenies using protein-coding genes in dinoflagellates will require estimates of within- and between-species variation for that gene.

Supplementary Material

Supplementary figures S1 and S2 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to thank Dr D. Wayne Coats of the Smithsonian Environmental Research Center for mentoring and support. The authors thank G. Concepcion, G. Mendez, W. Macturk, and R. Timme of the University of Maryland and two anonymous reviewers for critical reviews of the manuscript. This project was funded in part by a National Science Foundation grant from the Assembling the Tree of Life project (#EF-0629624). This work was supported by a Smithsonian Postdoctoral Fellowship grant to S.K.

References

Aury JM, Jaillon O, Duret L, et al. (43 co-authors). 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444:171–178.

Bachvaroff TR, Concepcion GT, Rogers CR, Delwiche CF. 2004. Dinoflagellate EST data indicate massive transfer of chloroplast genes to the nucleus. *Protist* 55:65–78.

Bachvaroff TR, Place AR. 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS One*. 3:e2929.

Bachvaroff TR, Place AR, Coats DW. 2009. Expressed sequence tags from *Amoebophrya* sp. Infecting *Karlodinium veneficum*: comparing host and parasite sequences. *J Eukaryot Microbiol*. 56:531–541.

Brayton KA, Lau AO, Herndon DR, et al. (28 co-authors). 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog*. 3:1401–1413.

Coats DW, Bachvaroff TR, Handy SM, Kim S, Garate-Lizarrage I, Delwiche CF. 2008. Prevalence and phylogeny of parasitic dinoflagellates (Genus *Blastodinium*) infecting copepods in the Gulf of California. *Oceanides* 23:63–77.

Delwiche CF. 1999. Tracing the thread of plastid diversity through the tapestry of life. *Am Nat*. 154:S164–S177.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. 19:11–15.

Eun HM. 1996. Enzymology primer for recombinant DNA technology. San Diego (CA): Academic Press.

Galuzzi L, Penna A, Bertozzini E, Vila M, Garces E, Magnani M. 2004. Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a dinoflagellate). *Appl Environ Microbiol*. 70:1199–1206.

Hackett J, Yoon H, Soares M, Bonaldo M, Casavant T, Scheetz T, Nosenko T, Bhattacharya D. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Curr Biol*. 14:213–218.

Handy SM, Bachvaroff TR, Timme RE, Coats DW, Kim S, Delwiche CF. 2009. Phylogeny of four dinophysiacean genera (Dinophyceae, Dinophysiales) based on rDNA sequences from single cells and environmental samples. *J Phycol*. 45:1163–1174.

Hiller RG, Wrench PM, Sharples FP. 1995. The light harvesting chlorophyll a-c-binding protein of dinoflagellates: a putative polyprotein. *FEBS lett*. 363:175–178.

Huber T, Faulkner G, Hugenholtz P. 2004. Bellerophon: a progarm to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20:2317–2319.

Joseph JM, Fey P, Ramalingam N, Liu X, Rohlfs M, Noegel AA, Müller-Taubenberger A, Glöckner G, Schleicher M. 2008. The actinome of Dictyostelium discoideum in comparison to actins and actin-related proteins from other organisms. *PLoS One*. 3:e2654.

Lahr DJ, Katz LA. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47:857–866.

Le QH, Markovic P, Hastings JW, Jovine RV, Morse D. 1997. Structure and organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*. *Mol gen genet*. 255:595–604.

Li L, Hong R, Hastings J. 1997. Three functional luciferase domains in a single polypeptide chain. *Proc Natl Acad Sci*. 94:8954–8958.

Lidie KB, Ryan JC, Barbier M, Van Dolah FM. 2005. Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of a DNA microarray. *Mar Biotechnol*. 7:481–493.

Lidie KB, Van Dolah FM. 2007. Spliced leader RNA-mediated *trans*-splicing in a dinoflagellate, *Karenia brevis*. *J Eukaryot Microbiol*. 54:427–435.

Lucas I, Vesk A. 1990. The fine structure of two photosynthetic species of *Dinophysis* (Dinophysiales, Dinophyceae). *J Phycol*. 26:345–357.

Maddison WP, Maddison PR. 2002. MacClade version 4: analysis of phylogeny and character evolution. Sunderland (MA): Sinauer Associates.

Morgan DR, Soltis DE. 1995. Phylogenetic relationships among members of Saxifragaceae *sensu lato* based on *rbcL* sequence data. *Ann Mo Bot Gard*. 82:208–234.

Moustafa A, Evans AN, Kulis DM, DL Erdner HJD, Anderson DM, Bhattacharya D. 2010. Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One*. 5:e9688.

Muller J, Oma Y, Vallar L, Friederich E, Poch O, Winsor B. 2005. Sequence and comparative genomic analysis of actin-related proteins. *Mol Biol Cell*. 16:5736–5748.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3:418–426.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 39:121–152.

OOta S, Saitou N. 1999. Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees. *Mol Biol Evol*. 16:856–867.

Park MG, Kim S, Kim HS, Myung G, Kang YG, Yih W. 2006. First successful culture of the marine dinoflagellate *Dinophysis acuminata*. *Aquat Microb Ecol*. 45:101–106.

Patron NJ, Waller RF, Keeling PJ. 2005. Complex protein targeting to dinoflagellate plastids. *J Mol Biol*. 348:1015–1024.

Patron NJ, Waller RF, Keeling PJ. 2006. A tertiary plastid uses genes from two endosymbionts. *J Mol Biol*. 357:1373–1382.

Rae P. 1976. Hydroxymethyluracil in eukaryote DNA: a natural feature of the pyrrophyta (dinoflagellates). *Science* 194:1062–1064.

Reichman J, Wilcox T, Vize P. 2003. PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low level of concerted evolution, isoform diversity and spectral tuning of chromophores. *Mol Biol Evol*. 20:2143–2154.

Robinson T, Katz LA. 2007. Non-mendelian inheritance of paralogs of 2 cytoskeletal genes in the ciliate *Chilodonella uncinata*. *Mol Biol Evol*. 24:2495–2503.

Rooney AP. 2004. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans. *Mol Biol Evol*. 21:1704–1711.

Rooney AP, Ward TJ. 2005. Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *Proc Natl Acad Sci U S A*. 102: 5084–5089.

Rooney AP, Ward TJ. 2008. Birth-and-death evolution of the internalin multigene family in *Listeria*. *Gene* 427:124–128.

Rowan R, Whitney SM, Fowler A, Yellowlees D. 1996. Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. *Plant Cell*. 8:539–553.

Saldarriaga JF, McEwan ML, Fast NM, Taylor FJR, Keeling PJ. 2003. Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int J Syst Evol Microbiol*. 53:355–365.

Sano J, Kato K. 2009. Localization and copoy number of the protein-coding genes actin, α-tubulin, and hsp90 in the nucleus of a primitive dinoflagellate, *Oxyrrhis marina*. *Zool Sci*. 26:745–753.

Schnepf E, Elbrächter M. 1988. Cryptophycean-like double membrane-bound chloroplast in the dinofalgellate, *Dinophysis* Ehrenb.: evolutionary, phyllogenetic and toxicological implications. *Botanica Acta*. 101:196–203.

Schnepf E, Elbrächter M. 1999. Dinophyte chloroplasts and phylogeny—a review. *Grana* 38:81–97.

Spector DL. 1984. Dinoflagellate nuclei. In: Spector DL, editor. Dinoflagellates. London: Academic Press. p. 107–147.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Takishita K, Koike K, Maruyama T, Ogata T. 2002. Molecular evidence for plastid robbery (Kleptoplastidy) in Dinophysis, a dinoflagellate causing diarrhetic shellfish poisoning. *Protist* 153(3):293–302.

Yang Z, Nelsen R, Goldman N, Pedersen AK. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.

Zhang H, Hou Y, Lin S. 2006. Isolation and characterization of proliferating cell nuclear antigen from the dinoflagellate *Pfeisteria piscicida*. *J Eukaryot Microbiol*. 53:142–150.

Zhang J, Hou Y, Miranda L, Campbell D, Sturm N, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci*. 104:4618–4623.

Zhang H, Lin S. 2003. Complex gene structure of the form II Rubisco in the dinoflagellate *Prorocentrum minimum* (Dinophyceae). *J Phycol*. 39:1160–1171.