Horizontal Gene Transfer is a Significant Driver of Gene Innovation in Dinoflagellates

Jennifer H. Wisecaver[1,3,*], Michael L. Brosnahan[2], and Jeremiah D. Hackett[1]

[1]Department of Ecology and Evolutionary Biology, University of Arizona

[2]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA

[3]Present address: Department of Biological Sciences, Vanderbilt University, Nashville, TN

*Corresponding author: E-mail: jen.wisecaver@vanderbilt.edu.

Accepted: November 12, 2013

Data deposition: The *Alexandrium tamarense* Group IV transcriptome assembly described in this article has been deposited at DDBJ/EMBL/GenBank under the accession GAIQ01000000.

Abstract

The dinoflagellates are an evolutionarily and ecologically important group of microbial eukaryotes. Previous work suggests that horizontal gene transfer (HGT) is an important source of gene innovation in these organisms. However, dinoflagellate genomes are notoriously large and complex, making genomic investigation of this phenomenon impractical with currently available sequencing technology. Fortunately, de novo transcriptome sequencing and assembly provides an alternative approach for investigating HGT. We sequenced the transcriptome of the dinoflagellate *Alexandrium tamarense* Group IV to investigate how HGT has contributed to gene innovation in this group. Our comprehensive *A. tamarense* Group IV gene set was compared with those of 16 other eukaryotic genomes. Ancestral gene content reconstruction of ortholog groups shows that *A. tamarense* Group IV has the largest number of gene families gained (314–1,563 depending on inference method) relative to all other organisms in the analysis (0–782). Phylogenomic analysis indicates that genes horizontally acquired from bacteria are a significant proportion of this gene influx, as are genes transferred from other eukaryotes either through HGT or endosymbiosis. The dinoflagellates also display curious cases of gene loss associated with mitochondrial metabolism including the entire Complex I of oxidative phosphorylation. Some of these missing genes have been functionally replaced by bacterial and eukaryotic xenologs. The transcriptome of *A. tamarense* Group IV lends strong support to a growing body of evidence that dinoflagellate genomes are extraordinarily impacted by HGT.

Key words: gene innovation, *Alexandrium tamarense* Group IV, phylogenetic profile, phylogenomics, de novo transcriptome assembly, mitochondrial metabolism.

Introduction

In eukaryotic evolution, the relative importance of horizontal gene transfer (HGT) compared with other sources of genetic novelty (i.e., gene duplication, modification, and de novo origination) is an unsettled topic. This is in contrast to Bacteria and Archaea, among which HGT is established as a major driver of genetic innovation (Ochman et al. 2000; Pál et al. 2005; Treangen and Rocha 2011). Although the impact of HGT on eukaryotic evolution remains poorly characterized, HGT has been implicated in the exploitation of new niches by several microbial groups, including apicomplexans (Striepen et al. 2004), ciliates (Ricard et al. 2006), diplomonads (Andersson et al. 2003), and fungi (Slot and Hibbett 2007). Thus, HGT may be a significant driver of gene innovation in at least some eukaryotic lineages (Keeling and Palmer 2008; Andersson 2009). However, poor resolution at the base of the eukaryotic tree of life as well as the dearth of next-generation sequence data from microbial eukaryotes complicates the interpretation of gene phylogenies otherwise suggestive of HGT (reviewed in Stiller 2011; see, e.g., Chan et al. 2012; Curtis et al. 2012; Deschamps and Moreira 2012).

Although challenges remain in measuring and interpreting HGT in eukaryotes, gene transfer during the evolution of mitochondria and plastids (i.e., endosymbiosis) is a well-established mechanism for gene transfer, a process referred to endosymbiotic gene transfer (EGT). The primary plastid of the Archaeplastida (i.e., red, green, and glaucophyte algae, Adl et al. 2005) arose through an endosymbiotic association between cyanobacteria and a heterotrophic, eukaryotic host (Douglas 1998). The resulting photosynthetic organelle, the plastid, maintains a reduced genome of ≤200 genes, but the majority of genes required for photosynthesis have been transferred to the nuclear genome of the algal host (Martin and Herrmann 1998). Many of these transferred genes are targeted back to the plastid, but some have been co-opted to function in novel processes and thereby increase the genetic potential of the host genome (Martin and Herrmann 1998). Plastid endosymbioses involving eukaryotic algal endosymbionts, rather than cyanobacteria, further distributed plastids, and the genes necessary to maintain them, across the eukaryotic tree of life (Archibald 2009).

Another possible mode of HGT in microbial eukaryotes is the acquisition of genetic material during prey ingestion (Doolittle 1998), which is supported by accounts of HGT in phagotrophic lineages such as ciliates (Ricard et al. 2006), euglenids (Maruyama et al. 2011), and amoebea (Eichinger et al. 2005). Further support for the prey ingestion model comes from algae (a nonmonophyletic group of photosynthetic, plastid-containing eukaryotes). Although primary plastid-containing organisms are strict autotrophs, with few exceptions, many algae with plastids derived via additional endosymbioses (e.g., euglenids and dinoflagellates) are mixotrophs that supplement photosynthesis with consumption of food particles (Stoecker 1998). In the case of these mixotrophic algae, the prey ingestion hypothesis predicts that these organisms will have genes acquired both from their plastid and from their prey (Doolittle 1998).

The dinoflagellates are protists (i.e., microbial eukaryotes) common in many aquatic environments and are ideal organisms for investigating the impact of HGT on eukaryotic evolution. Many dinoflagellate species are mixotrophs, having the ability to obtain carbon from photosynthesis as well as from ingestion of other phytoplankton and bacteria (Hackett, Anderson et al. 2004). In support of the prey ingestion model, previous work suggests that dinoflagellate nuclear genomes contain a large number of genes acquired via plastid endosymbiosis as well as genes horizontally acquired from other sources (Hackett et al. 2005, 2013; Nosenko et al. 2006; Nosenko and Bhattacharya 2007; Janouskovec et al. 2010; Minge et al. 2010; Wisecaver and Hackett 2010; Stuken et al. 2011; Chan et al. 2012; Orr et al. 2013). The placement of dinoflagellates is resolved and well supported in phylogenetic analyses, which is essential for inferring HGT based on phylogenetic incongruence between gene and species trees. Dinoflagellates are sister to the Perkinsidae, a parasitic group that includes the oyster pathogen *Perkinsus marinus* (Reece et al. 1997). Dinoflagellates and Perkinsidae together are sister to the apicomplexans, an exclusively parasitic group responsible for many human diseases including malaria and toxoplasmosis (Fast et al. 2002). The nearest major protist group to dinoflagellates and apicomplexans are the ciliates, and these three lineages together are the primary members of the superphylum Alveolata (Adl et al. 2005). Phylogenetic studies suggest that alveolates are related to stramenopiles (e.g., diatoms and giant kelp) and rhizarians (e.g., foraminifera and radiolarians) an association abbreviated as the stramenopile–alveolate–rhizaria (SAR) supergroup (Burki et al. 2007; Parfrey et al. 2010). However, despite this phylogenetic resolution as well as published cases of gene transfer, the full extent, timing, and consequences of HGT in dinoflagellates remain unknown because example genomes have not yet been sequenced.

Dinoflagellate genomes range in size from 1.5 to 185 Gbp (0.8 to over 60 times the size of the human genome) and are rife with noncoding sequence, tandem gene repeats, and other unusual features that make genome sequencing with current technology highly impractical with current assembly technology (Wisecaver and Hackett 2011). Fortunately, transcriptome sequencing is an alternative approach for questions requiring comprehensive gene discovery in nonmodel organisms with complex genomes. Here, we analyze a comprehensive de novo transcriptome assembly for the dinoflagellate, *Alexandrium tamarense* strain CCMP1598, a member of the "Group IV" clade within the *A. tamarense* species complex. This species complex comprises five such clade groups each of which likely represents a distinct cryptic species (Lilly et al. 2007). We cross-reference our *A. tamarense* Group IV gene set to transcriptomic and expressed sequences tag (EST) data from 21 additional dinoflagellate species (including transcriptome assemblies from *A. tamarense* Group I and Group III strains) to derive a final dinoflagellate unigene set that we then compare with 16 other algal and protist genomes. Using ancestral gene content reconstruction, we map gene acquisitions on the alveolate evolutionary tree and validate the results using a phylogenomic pipeline. This combined approach offers a robust, comparative exploration of the pattern of HGT in dinoflagellates relative to other eukaryotes.

Materials and Methods

Cell culture and Illumina sequencing of the *A. tamarense* Group IV transcriptome has been published elsewhere (Hackett et al. 2013). Cultures of the axenic *A. tamarense* Group IV strain CCMP1598 were obtained from the National Center for Marine Algae and Microbiota (https://ncma.bigelow.org, last accessed November 26, 2013) and checked for visible signs of contamination via microscopy prior to RNA isolation. RNA-seq data were quality trimmed with the trim read module in the CLC genomics workbench (www.clcbio.com, last accessed November 26, 2013) using a quality score limit of 0.05 and removing all ambiguous nucleotides. The trimmed reads were assembled in Velvet (version 1.1.02) using the Oases extension (version 0.1.20) with tracking of short read positions enabled (Zerbino and Birney 2008). The trimmed reads were randomly subsampled using the python package HTSeq for a rarefaction curve analysis. Eight different builds were created per read subset using a range of hash lengths from 23 to 51 kmers. The final build was constructed using the largest hash length and included the seven additional builds as long sequences with the Oases conserve long option enabled (Schulz et al. 2012). This multi-hash-length assembly protocol maximized the assembly length while minimizing the number of contigs (for the rarefaction curves, see fig. 1 and supplementary fig. S1, Supplementary Material online).

Alexandrium. tamarense Group IV contigs were Blast queried against NCBI's nonredundant protein and EST databases, algal genomes from the Joint Genome Institute (http://genome.jgi.doe.gov, last accessed November 26, 2013), and additional algal 454 and Illumina transcriptome assemblies deposited in NCBI's Transcriptome Shotgun Assembly (TSA) archive (supplementary table S1, Supplementary Material online). Contigs were tagged as potential contamination if their top Blast hit was not to another dinoflagellate and removed prior to further analysis. The *A. tamarense* Group IV assembly has been deposited at DDBJ/EMBL/GenBank under the accession GAIQ01000000. Protein sequences were predicted from the nucleotide assembly using top Blast hit information and FrameDP (Gouzy et al. 2009).

Ancestral Gene Content Reconstruction

In addition to the *A. tamarense* Group IV transcriptome, 16 genomes were used in the ancestral state reconstruction analyses (Gardner et al. 2002; Abrahamsen et al. 2004; Armbrust et al. 2004; Pain et al. 2005; Aury et al. 2006; Stover et al. 2006; Brayton et al. 2007; Palenik et al. 2007; Bowler et al. 2008; Gajria et al. 2008; Worden et al. 2009; Cock et al. 2010; Gobler et al. 2011). Protein sequences were downloaded from the NCBI genome database (*Babesia bovis*, PRJNA20343; *Cryptosporidium parvum*, PRJNA15586; *Plasmodium falciparum*, PRJNA148; *Theileria annulata*, PRJNA16308; *Toxoplasma gondii*, PRJNA32719; *Paramecium tetraurelia*, PRJNA19409; and *Tetrahymena thermophila*, PRJNA16792), the JGI genome portal (*Aureococcus anophagefferens*, filtered models 3; *Emiliania huxleyi*, best proteins; *Guillardia theta*, 20101209; *Micromonas pusilla*, 20110615; *Ostreococcus lucimarinus*, filtered models 2; *O. tauri*, filtered models 2; *Phaeodactylum tricornutum*, filtered models 2; *Thalassiosira pseudonana*, filtered models 2), and Ghent University's online genome annotation server BOGAS (*Ectocarpus siliculosus*, downloaded on August 17, 2011). Genes were annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) automated annotation pipeline using the bidirectional best hit method (Moriya et al. 2007). The gene presence/absence matrix (phylogenetic profile) was analyzed in Count to reconstruct gene history using both Dollo parsimony (DP) and unweighted Wagner parsimony (WP) (Csurös 2010).

Phylogenomic Analysis

Two local databases were constructed for the purpose of this article. The protein database included NCBI's Reference Sequence (release 42) and predicted protein sequences from recently sequenced microbial eukaryotes (JGI genome portal and Ghent University's online genome annotation server BOGAS). The nucleotide database included transcript sequences from additional microbial eukaryotes from NCBI's Expressed Sequence Tag (EST) and TSA databases. Each local database was further subdivided based on major taxonomic groups (for list of groups, see supplementary table S1, Supplementary Material online), and each taxonomic group was individually queried using Blast.

The *A. tamarense* Group IV gene phylogenies were constructed using a custom phylogenetic pipeline; scripts are available from the authors upon request. Predicted amino acid sequences were first queried using BlastP and TBlastN against the local databases. For each Blast result, a hit was considered significant if the *E*-value was less than 1e^{-3} and the bit score was greater than 60. The Blast reports were parsed seven times using a range of fraction conserved (FC) thresholds (0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9). The FC score accounts for amino acid substitutions that occur frequently (given a substitution matrix) and is a more relaxed metric of sequence similarity than percent identity. If a hit passed the *E*-value, bitscore, and FC thresholds, the associated sequence was extracted from the database using a custom perl script. For matches to the nucleotide database, only the translations of the high-scoring segment pairs were included. To reduce the number of paralogs in the analysis, only the top hit per species was extracted. Extracted sequences were reordered based on global similarity to the query sequence with MAFFT using the minimum linkage clustering method and rough distance measure (number of shared 6mers) (Katoh et al. 2005). After reordering, the files were reduced to include only the top 1,000 sequences, and files with less than four sequences were eliminated. Alignments were performed with MAFFT using the auto strategy selection and the BLOSUM scoring matrix closest to the FC threshold (i.e., a lower BLOSUM matrix was used when average sequence conservation was low, and a higher BLOSUM matrix was used when average sequence conservation was high). Poorly aligned positions and sequences were removed from the alignment using REAP (Burleigh et al. 2011), and trimmed alignments were further refined by a second MAFFT alignment using the same parameters as above. Phylogenetic trees were inferred using FastTree assuming a JTT+CAT amino acid model of substitution and 1,000 resamples (Price et al. 2009, 2010; Liu et al. 2011).

Trees were filtered using a perl script that eliminated trees in which only dinoflagellates were present or dinoflagellates did not form a monophyletic group. Trees that contained only one dinoflagellate species were also removed because monophyly could not be assessed and to mitigate any potential signal from contamination. For each tree, dinoflagellate nearest neighbors were identified using Phylosort, a tool for sorting phylogenetic trees by searching for a user-specified grouping of interest (Moustafa and Bhattacharya 2008). For this analysis, a nearest neighbor association is defined as a sister relationship between a clade of dinoflagellates and another group of organisms with branch support of 0.75 or greater. Although we required dinoflagellate monophyly, other members of the neighbor group could be present elsewhere in the tree. This approach identified the most closely related sequences to our dinoflagellate clade while allowing for HGT and paralogous sequences in other lineages. To determine all possible nearest neighbors to dinoflagellates, we iterated through Phylosort using all lineages of interest (see fig. 3 for list of associations investigated) and identified all trees in which each lineage formed a neighbor association with dinoflagellates. For each iteration, the trees were rerooted with an outgroup that was automatically selected from taxa outside the relationship of interest. For each tree, the phylogenetic nearest neighbor to the dinoflagellate clade (*A. tamarnese* Group IV plus at least one additional dinoflagellate species) was determined, and the results from multiple pipeline iterations (using different FC thresholds) were combined to derive a consensus dinoflagellate sister association for each contig (supplementary table S2, Supplementary Material online). Sequences for which no consensus could be determined (i.e., different pipeline iterations yielded different organisms grouping sister to dinoflagellates) were excluded from the analyses.

When manual curation of taxa or alignments was required, additional trees were built using RAxML rapid bootstrapping (100 replications) and maximum-likelihood (ML) search assuming a WAG amino acid model of substitution and Γ site heterogeneity model (Stamatakis 2006). RAxML trees were built for NDH-2, MQO, PNO, fumarase class I and II, FAD-DHAP, NADPH-IDH-II, and IlvE. Tests of monophyly were performed in RAxML using the log-likelihood Shimodaira-Hasegawa (SH) test between the unconstrained best tree and the best tree given a constrained topology (Shimodaira and Hasegawa 1999).

Results and Discussion

Transcriptome sequencing of *A. tamarense* Group IV produced 10.6 Gbp of raw read length that assembled into 142,638 contigs. The Oases de novo transcriptome assembler further clustered the contigs into 101,118 groups, hereafter referred to as Oases loci. This feature of the Oases assembler was originally developed to cluster alternatively spliced transcript isoforms into distinct loci. In dinoflagellates, however, it is likely that these Oases loci correspond instead to large tandem gene arrays, which may contain tens to thousands of highly similar gene copies (Liu and Hastings 2006; Bachvaroff and Place 2008; Hou and Lin 2009). Random

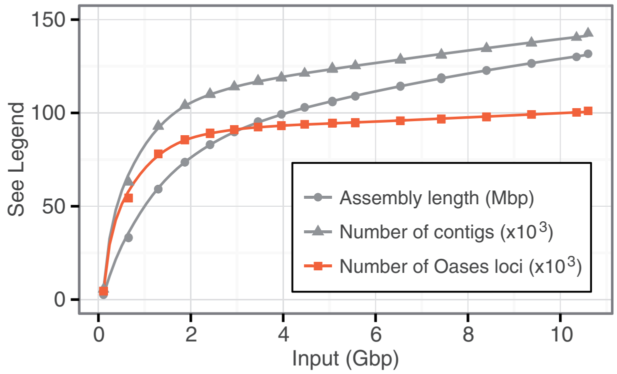

Fig. 1.—Rarefaction curve showing the effect of adding raw sequence input (in billions of basepairs) on the *Alexandrium tamarense* Group IV de novo transcriptome assembly. Assembly length was measured in millions of basepairs. Number of sequences is illustrated both in terms of number of contigs in the full assembly and number of Oases loci.

subsampling of the raw data and rarefaction curve analyses suggested that, although the overall assembly length and contig count continued to grow with additional input sequence, the number of Oases loci did not increase dramatically (fig. 1). From these analyses, we concluded that our sequencing effort was sufficient to uncover most, if not all, gene families from *A. tamarense* Group IV, expressed under the culture conditions.

The longest contig from each Oases locus in the *A. tamarense* Group IV assembly was cross-referenced against a custom database (supplementary table S1, Supplementary Material online) that included publicly available dinoflagellate transcriptomes and ESTs. Only the 85,163 contigs with a top Blast hit to another dinoflagellate were retained for downstream comparative analyses. The purpose of this contig-filtering step was to remove potential contamination, which can be difficult to identify in de novo transcriptome assemblies that lack genomic context. A second consequence of this filtering step is that any *A. tamarense* Group IV-specific genes were not included in our downstream comparative analysis, potentially underestimating the number of genes present in this organism. Because the focus of our analysis is gene change shared among dinoflagellate species, we consider this a reasonable, conservative tradeoff. A final caveat to this filtering step is that for many *A. tamarense* Group IV contigs, the only other dinoflagellate homologs found were from the sister species *A. tamarense* Group I and Group III. In some cases, this may reflect *Alexandrium*-specific inheritance (i.e., not shared broadly across dinoflagellates). In alternative cases, it may reflect the high depth of sequencing used for these transcriptomes; when this analysis was begun, no other dinoflagellate transcriptomes had been sequenced to similar depth.

The total number of genes in *A. tamarense* Group IV is unknown, but flow cytometry experiments have estimated the haploid genome to be approximately 100 Gbp (LaJeunesse et al. 2005). Current estimates of gene number in dinoflagellates are highly variable, complicated by widespread gene duplication and the absence of genome sequences (Bachvaroff and Place 2008). One massively parallel signature sequencing experiment identified 40,029 unique 21 bp transcript tags in this species, representing a lower-bound estimate of gene number (Erdner and Anderson 2006; Moustafa et al. 2010). A regression analysis based on the relationship between gene content and genome size in sequenced organisms projected 87,688 protein-coding genes in *Prorocentrum micans*, a dinoflagellate whose genome is of comparable size (Hou and Lin 2009). Based on these estimates, the number of Oases loci in our assembly (101,118), although high relative to gene content estimates for other organisms, is not unreasonable for a dinoflagellate transcriptome.

Reconstruction of Ancestral Gene Content Using Parsimony

The *A. tamarense* Group IV transcriptome was compared with the gene sets of 16 fully sequenced eukaryotes. Five apicomplexans (*B. bovis*, *C. parvum*, *Pl. falciparum*, *T. annulata*, and *To. gondii*) and two ciliates (*Pa. tetraurelia* and *Te. thermophila*) were selected as representatives of the superphylum Alveolata. Stramenopile species included two diatoms (*Ph. tricornutum* and *Th. pseudonana*), a pelagophyte (*Au. anophagefferens)*, and a filamentous brown alga (*Ec. siliculosus*). Other species included in the analysis included a haptophyte (*E. huxleyi*), a cryptophyte (*G. theta*), and three prasinophyte green algae (*M. pusilla*, *O. lucimarinus*, and *O. tauri*). In total, 5,347 *A. tamarense* Group IV contigs were assigned to 2,959 KEGG ortholog (KO) groups (for distribution of biological pathways, see supplementary fig. S3, Supplementary Material online). *A. tamarense* Group IV had the largest number of KO annotations, but comparable numbers were obtained for all genomes analyzed (supplementary table S3, Supplementary Material online).

The taxa selected for the ancestral gene content reconstruction emphasized the difference in gene content between photosynthetic dinoflagellates, including *A. tamarense* Group IV, and their heterotrophic sister groups (i.e., apicomplexans and ciliates). The nine algal genomes included in the analysis allowed the comparison of three hypothetical sequence sets in dinoflagellates. Sequences shared between *A. tamarense* Group IV and other alveolates were most likely retained from a common ancestor and likely demonstrate vertical inheritance. Sequences shared among algae, including *A. tamarense* Group IV, were potentially acquired by dinoflagellates during plastid endosymbiosis via EGT. However, the exact number and timing of endosymbioses in the history of dinoflagellate plastids remains unknown (for hypotheses, see Archibald 2009), therefore, it is also possible that genes shared between *A. tamarense* Group IV and other algae were acquired via HGT from algal prey or even retained from a photosynthetic ancestor. Finally, sequences found in *A. tamarense* Group IV and missing in the other 16 genomes in the analysis may represent potential instances of HGT in dinoflagellates, which can be further evaluated with phylogenetics.

DP and unweighted WP were used to reconstruct the gene content (i.e., the presence or absence of each KO) for all ancestors in the phylogenetic tree (fig. 2). Gene gain, in our ancestral state reconstruction analysis, is the shift from KO absence at the parent node to presence at the node of interest, and gene loss is the opposite transition. Because our comparisons were restricted to genes with KEGG annotations, de novo gene origination is an unlikely mechanism for gene gain in this analysis. Gene transfer, either horizontally or endosymbiotically, is one possible mechanism for gene gain; however, the results of ancestral gene content reconstruction are highly dependent on 1) the taxa included in the analysis and 2) the selected method for ancestral state reconstruction. DP is a conservative estimator of gene gain at terminal branches, because a gene may be lost multiple times but gained only once. This approach pushes any gain back in time to the last common ancestor of any taxa sharing the KO. Conversely, a WP approach, in which gene gain and gene loss are weighted the same, is the more relaxed method for estimating gene gain at the leaves of the tree. For this analysis, the two parsimony methods represent realistic bounds for estimating the number of gene changes on a branch.

Regardless of the method used, the branch leading to *A. tamarense* Group IV had the largest number of genes gained (1,563 WP; 314 DP) by a large margin, and the branch was unique in its asymmetry between gene gain and loss regardless of parsimony method (fig. 2). These genes gained on the dinoflagellate branch under DP are necessarily missing from all other species in the analysis and are potentially acquired through HGT. Of the genes gained along the dinoflagellate branch using WP, 955 were independently lost in apicomplexans and ciliates in the DP analysis. Put another way, these 955 genes are shared between *A. tamarense* Group IV and one or more algal genomes in the analysis but are absent in the alveolate relatives of dinoflagellates. Predictably, many of these genes are algal in nature, including genes involved in photosynthesis (ko00195), the urea cycle (ko00330), the plant-like shikimate pathway (ko00400), and histidine biosynthesis (ko00340). In general, however, the KOs unique to *A. tamarense* Group IV as well as those shared between the dinoflagellate and other algae are involved in many biological processes rather than a few specific pathways (supplementary fig. S3, Supplementary Material online). *Alexandrium tamarense* Group IV also appears to have both alveolate and algal copies of many genes in pathways such as fatty acid biosynthesis and fatty acid metabolism (ko0061 and ko00071). These KOs are ideal candidates for investigation of functional divergence

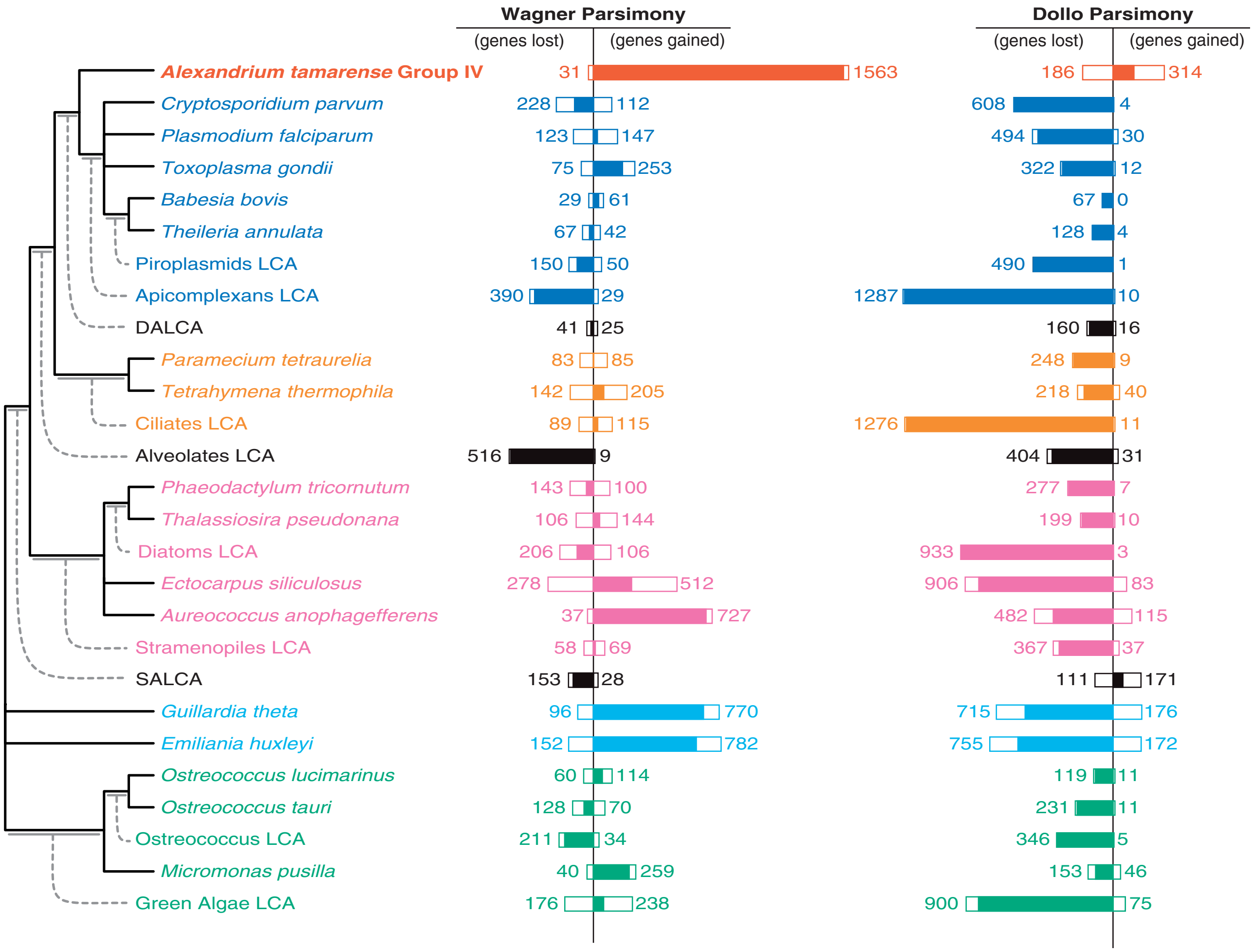

Fig. 2.—Ancestral gene content reconstruction of KOs. For both WP and DP analyses, bars and numbers to the right of the vertical axis represent the number of gene families gained at a branch relative to its parent. Bars and numbers to the left of the axis represent the number of gene families lost. The solid bar represents the net effect, either gain or loss of gene families, at each branch. LCA, last common ancestor; DALCA, dinoflagellate, apicomplexan LCA; SALCA, stramenopile, alveolate LCA.

(e.g., subfunctionalization and neofunctionalization) of redundant genes in eukaryotes impacted by endosymbiosis and HGT. Color-coded maps of KEGG pathways discussed are provided in supplementary figure S4, Supplementary Material online.

Validating Ancestral Gene Content Reconstruction with Phylogenomics

In our analysis, the use of phylogenetic profiles for ancestral gene content reconstruction can predict cases of HGT in dinoflagellates only if the transfer resulted in the acquisition of the first copy of a given KO. Gene replacement or acquisition of additional paralogs via HGT cannot be detected using this method and can lead to underestimation of gene transfer. Conversely, KOs shared between dinoflagellates and other alveolates not included in the reconstruction analysis (e.g., gregarines or chromerids) would appear as unique gains in *A. tamarense* Group IV, potentially overrepresenting the number of true gene transfers in dinoflagellates. Phylogenetic incongruence between gene and species trees is another common method for detecting HGT and is not limited to the detection of novel gene families. Another strength of phylogenetic analysis is the ability to bring in additional sequences from many more taxa that lack fully sequenced genomes.

We created an automated phylogenomic pipeline to build gene trees for KEGG-annotated contigs from *A. tamarense* Group IV. The pipeline was unique in that query contigs were run through the pipeline several times while altering the threshold for what was considered an acceptable match for extracting full-length homologs for phylogenetic analysis. This approach uses the consensus between pipeline iterations, rather than a conservative *E*-value cutoff, to infer reliable phylogenetic relationships to query sequences (supplementary table S2, Supplementary Material online). Of the 5,347 KEGG-annotated *A. tamarense* Group IV contigs run through our phylogenetic pipeline, 876 (16.4%) had an identifiable clade that grouped sister to dinoflagellates across the majority of pipeline iterations (table 1). In agreement with the species tree, the most common sister group to dinoflagellates was the perkinsid *P. marinus*, which was recovered in 266 contigs (30% of trees built). Stramenopiles and bacteria were also common sister groups to dinoflagellates, recovered in 103 and 92 contigs, respectively.

To assess the congruence between the presence–absence and phylogenetic analyses, we tested three KO annotated gene sets in *A. tamarense* Group IV: 1) KOs present in all species included in the presence–absence profile (hereafter referred to as the ubiquitous set); 2) KOs shared between dinoflagellates and other algal species (hereafter referred to as the algal set); and 3) KOs present in dinoflagellates and absent from all other genomes in the presence–absence profile (hereafter referred to as the dinoflagellate set). The ubiquitous set consisted of 393 *A. tamarense* contigs and, not surprisingly, contained many essential genes, including components of the ribosome as well as DNA and RNA polymerase. The phylogenomic pipeline constructed gene trees for 173 of the 393 contigs in this set, and the majority (101, 58%) had dinoflagellates grouping sister to *P. marinus* in agreement with our expectation that these sequences would be vertically inherited (fig. 3). A further 22 contigs in the ubiquitous set showed dinoflagellates grouping with other alveolates, also consistent with vertical inheritance. The algal set consisted of 955 contigs that we predicted would be comprised of sequences that were either vertically inherited from an algal ancestor (and independently lost in ciliates and apicomplexans) or acquired through plastid EGT or HGT from algal prey. For this set, we constructed gene trees for 247 contigs, many of which (112, 45%) were consistent with our prediction and showed dinoflagellates grouping sister to algae, most often stramenopiles (fig. 3). However, some contigs in this set did not match the prediction, including 16% that showed dinoflagellates grouping with bacteria and another 15% that showed dinoflagellates grouping basally to a diverse clade of eukaryotes. One possible explanation for these associations is that the genes were acquired independently in photosynthetic dinoflagellates and other algae. Methodological artifacts including taxon sampling (e.g., Rokas et al. 2003), long-branch attraction (e.g., Brinkmann et al. 2005), and differential gene loss (e.g., Qiu et al. 2012) could also be responsible for atypical phylogenetic associations.

Finally, the dinoflagellate set included 314 contigs that we predicted were likely examples of HGT because they were not present in other species in the presence–absence analysis. The phylogenetic analysis (66 contigs with trees) supported this conclusion in most cases (fig. 3). Over 37% (26 contigs) showed dinoflagellates grouping with bacteria compared with just 1% and 16% showing the same association in the ubiquitous and algal sets, respectively. Similarly, another 17% (11 contigs) showed dinoflagellates grouping with opisthokonts (e.g., fungi and choanoflagellates), eukaryotes that are very distantly related to dinoflagellates (Parfrey et al. 2010). Taken together, the results from the phylogenetic pipeline analysis support the ancestral gene-state reconstruction analysis and demonstrate extensive HGT in dinoflagellates.

Table 1

Summary of Results from Phylogenomic Pipeline Showing Number of KEGG-Annotated Gene Trees Supporting Different Dinoflagellate Nearest-Neighbor Associations

Clade Sister to Dinoflagellates	No. Gene Trees			
	Ubiquitous[a]	Algal[b]	Dinoflagellate[c]	Total
Perkinsus	101	20	5	266
Chromera	0	4	0	5
Apicomplexans	12	6	2	58
DAC[d]	8	0	0	29
Ciliates	2	3	1	17
Alveolates	0	0	0	2
Stramenopiles	8	52	3	103
Rhizaria	1	7	3	19
SAR[e]	1	1	0	3
Plantae	3	22	4	48
Haptophytes	1	34	2	53
Cryptophytes	4	4	0	14
Excovates	1	4	1	17
Amoebozoa	1	1	0	4
Opistokonts	4	12	11	37
Eukaryotes	24	37	8	109
Bacteria	2	40	26	92
All	173	247	66	876

Note.—Ubiquitous, algal, and dinoflagellate gene sets based on KEGG ancestral gene content reconstruction (see fig. 2 for list of species in KEGG analysis). These putative sets are compared with the phylogenomic results, which included sequences from many more organisms.

[a] Ubiquitous indicates genes present in all genomes included in the ancestral gene content reconstruction.

[b] Algal includes genes shared between *Alexandrium tamarense* Group IV and other algal species in the ancestral gene content reconstruction.

[c] Dinoflagellate indicates genes present in *A. tamarense* Group IV and absent from all other genomes in the ancestral gene content reconstruction.

[d] Dinoflagellate–apicomplexan clade. In these gene trees, dinoflagellate sequences were sister to a complex clade consisting of two or more of the following groups/species: apicomplexans, *Chromera, Perkinsus.*

[e] Stramenopile, alveolate, rhizaria supergroup. In these gene trees, dinoflagellate sequences were sister to a complex clade consisting of two or more of the following groups: stramenopiles, alveolates (other than dinoflagellates), rhizarians.

Dinoflagellate Mitochondrial Metabolism

Despite being sister lineages, apicomplexans and dinoflagellates appear to have, at least superficially, vastly different ecologies. The majority of apicomplexans are obligate, intracellular, animal parasites, whereas many dinoflagellates (including the *A. tamarense* species) are photosynthetic, marine, and free-living. However, several apicomplexan genomic characteristics thought to be derived and the result of a

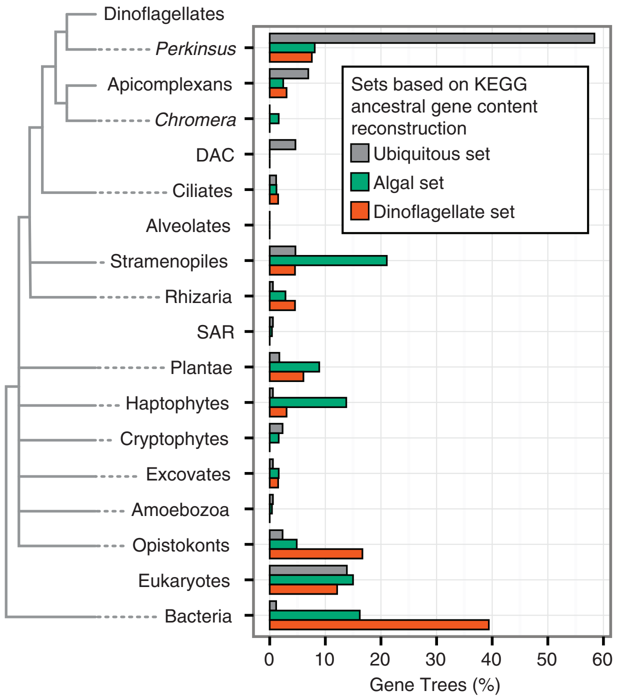

Fig. 3.—Phylogenomic bargraph showing the distribution of gene trees supporting diverse phylogenetic associations between dinoflagellates and other groups of organisms. Species tree is provided for reference, left. Bars represent putative categories based on the phylogenetic profile of the 17 species used in the KEGG ancestral gene content reconstruction (for list of species see fig. 2). The ubiquitous set (gray bars) represents genes found in all 17 species. The algal set (green bars) represents genes found in *Alexandrium tamarense* Group IV and the other five algal species. The dinoflagellate set (red bars) represents genes present in *A. tamarense* Group IV and absent from the other 16 genomes. These putative categories were compared with the phylogenomic results, which included sequences from many more organisms. DAC, dinoflagellate, apicomplexan clade.

parasitic lifestyle are shared with dinoflagellates, suggesting that these features evolved in their common ancestor (Danne et al. 2012). Notably, apicomplexans and dinoflagellates share extremely reduced mitochondrial genomes encoding only three protein-coding genes: cob, cox1, and cox3 (Waller and Jackson 2009), indicating extreme mitochondrial genome reduction in the Dinoflagellate-Apicomplexan Last Common Ancestor (hereafter referred to as DALCA). Despite the loss of most mitochondria-encoded genes, free-living dinoflagellates are thought to retain typical mitochondrial metabolic pathways including oxidative phosphorylation and the tricarboxylic acid (TCA) cycle (Danne et al. 2012). It is possible that genes necessary for these mitochondrial processes were transferred to the dinoflagellate nuclear genome during mitochondrial genome reduction. A similar pattern of gene transfer from organelle to nuclear genome occurred in dinoflagellates during plastid genome reduction (Bachvaroff et al. 2004; Hackett, Yoon et al. 2004). Alternatively, mitochondrial metabolism could have been functionally rescued by nuclear-encoded genes, either through repurposing of native genes or acquisition of novel genes via HGT. An analysis of dinoflagellate nuclear-encoded genes involved in mitochrondrial metabolism has been hindered by the lack of comprehensive gene surveys for these organisms. In our gene content reconstruction analysis of nuclear-encoded genes, DALCA had a small but significant number of genes lost (41 WP; 160 DP). Consistent with shared mitochondrial genome reduction, the majority of these gene losses in DALCA were related to mitochondrial metabolism. What follows is a discussion of mitochondria-related gene loss in dinoflagellates and apicomplexans and a phylogenetic analysis of dinoflagellate nuclear-encoded genes that may functionally replace those lost in DALCA (table 2).

NADH Dehydrogenase

The most notable gene change in DALCA was the loss of all subunits of NADH dehydrogenase (Complex I of oxidative phosphorylation). Genes present in the alveolate ancestor and lost in DALCA include those that encode the hydrophilic subunits of Complex I, NDUFS1, NDUFS4, NDUFS6, NDUFS8, NDUFV1, NDUFV2, NDUFA2, NDUFA5, NDUFA9, and NDUFAB1. In fact, no *A. tamarense* Group IV contigs mapped to Complex I in the KEGG oxidative phosphorylation pathway (ko00190, see supplementary fig. S4G, Supplementary Material online), and a search of all apicomplexan and dinoflagellate sequences in NCBI also yielded zero matches to Complex I. Additionally, no Complex I genes were found in the recently sequenced transcriptome of the dinoflagellate, *Hematodinium* sp. (Danne et al. 2012). In typical mitochondria, the hydrophobic arm of Complex I is mitochondria encoded, whereas the hydrophilic subunits listed above are located in the nucleus (Kerscher et al. 2008). Presumably, mitochondrial genome reduction—including the loss of the hydrophobic subunits of Complex I—occurred in DALCA and was associated with the loss of the nuclear-encoded subunits as well.

In place of Complex I, five copies of an alternative NADH dehydrogenase, NDH-2, were identified in the *A. tamarense* Group IV transcriptome (table 3). NDH-2 consists of a single protein and is found across the tree of life, most often acting as an accessory enzyme to Complex I (Kerscher et al. 2008). When Complex I is present, the specific purpose of NDH-2 is unknown, but suggested functions include elimination of excess reducing equivalents and the prevention of reactive oxygen species formation (Moller 2001; Raghavendra and Padmasree 2003). In the apicomplexan *To. gondii*, NDH-2 is internally oriented, consistent with it oxidizing matrix NADH and acting as a functional replacement of Complex I (Lin et al. 2011). The five dinoflagellate isoenzymes identified in

Table 2

Summary of Discussed Mitochondrial Proteins in Dinoflagellates

Mitochondrial Protein/Protein Complex	S	C	A	P	D	Origin of Dinoflagellate Gene Copy[a]
NADH dehydrogenase						
Complex I	+	+	–	–	–	Lost in dinoflagellate–apicomplexan ancestor
NADH-2	+	+	+	+	+	Not resolved; see table 3
G3P DH FAD-dep	+	+	+	+	+	Not resolved; SH test supports nonvertical inheritance
MQO	–	–	+	+	+	Not resolved; SH test supports nonvertical inheritance
Pyruvate dehydrogenase						
PDHA, PDHB, DLAT	+	+	–	–	–	Lost in dinoflagellate–apicomplexan ancestor
AceE	–	–	–	–	+	Horizontally acquired from bacteria
PNO	+[b]	–	+[c]	+	+	Horizontally acquired in dinoflagellate–apicomplexan ancestor from unknown donor
TCA cycle						
NAD-IDH	+	+	–	–	–	Lost in dinoflagellate–apicomplexan ancestor
NADP-IDH-I	+	+	+	+	+	Vertically inherited
NADP-IDH-II	+	–	–	–	+	Not resolved; HGT from bacteria or H/EGT from algae
Fumarase-I	+	–	+	+	+	Not resolved; SH test supports vertical inheritance
Fumarase-II	+	+	–	–	+	Not resolved; SH test supports nonvertical inheritance

NOTE.—Presence (+) or absence (–) of genes is shown for each lineage: S, Stramenopiles; C, Ciliates; A, Apicomplexans; P, *Perkinsus marinus*; and D, Dinoflagellates.

[a] See supplementary figure S5, Supplementary Material online, for phylogenies.

[b] Within apicomplexans, PNO has only been described in *Cryptosporidium hominis* (Rotte et al. 2001).

[c] Within stramenopiles, PNO has only been described in the parasite, *Blastocystis hominis*. (Lantsman et al. 2008).

Table 3

Alexandrium tamarense Group IV Alternative NADH Dehydrogenases

Contig	First GXGXXG[a]	Second GXGXXG[a]	EF-Hand Motif[b]	Phylogenetic Relationship[c]
Locus 23902	GSGWAA	GGGPTG	No	Algae
Locus 31037	GSGWGA	GGGPTG	No	Not resolved
Locus 74153	GSGWGA	GGGPTG	Yes	Locus 7449, Viridiplantae
Locus 7449	GSGWGC	GGGPTG	Yes	Locus 74153, Viridiplantae
Locus 87599	GSGWGS	GGGPTG	No	Not resolved

[a] Transcript amino acids corresponding to the two GX(X)GXXG motifs characteristic of alternative NADH dehydrogenases.

[b] Presence or absence of the Ca^{2+}-binding EF hand motif indicative of NDH-2 group B.

[c] Sister clade to the dinoflagellate NADH dehydrogenase(s). See supplementary figure S5*A*, Supplementary Material online, for phylogeny.

A. tamarense Group IV all possess the two GX(X)GXXG motifs characteristic of alternative NADH dehydrogenases, and two copies contain a Ca^{2+}-binding EF hand motif present in NDH-2 group B (Melo and Bandeiras 2004).

Phylogenetic analyses were inconclusive but suggest that at least one of the NDH-2s in *A. tamarense* Group IV is not related to the alternative NDH dehydrogenases of apicomplexans. None of the dinoflagellate sequences group with apicomplexans in ML analyses, although bootstrap support values are low for some nodes (supplementary fig. S5*A*, Supplementary Material online). Instead, dinoflagellate NDH-2s group with plastid-containing stramenopiles, Viridiplantae, and other algae. However, log-likelihood SH tests between the ML best tree and trees in which apicomplexans and dinoflagellate sequences were constrained demonstrated that the ML best tree was not significantly better than most constrained trees. Only for the characteristically algal copy (Locus 23902)—which groups with plastid-containing stramenopiles, haptophytes, rhodophytes, and green algae—could the analysis of monophyly reject the dinoflagellate–apicomplexan constrained tree.

An additional candidate pathway for channeling electrons into oxidative phosphorylation is the FAD-dependent glycerol-3-phosphate dehydrogenase system. Two copies of this dehydrogenase were found in the *A. tamarense* Group IV transcriptome, but neither group with other alveolates in phylogenetic analyses (supplementary fig. S5*B*, Supplementary Material online). Instead, one dinoflagellate copy groups with stramenopiles and rhizaria. The second copy of FAD-dependent glycerol-3-phosphate dehydrogenase in *A. tamarense* Group IV is of uncertain evolutionary origin, but SH tests of monophyly reject either copy grouping with alveolates. Intriguingly, no version of this enzyme could be found in the recently sequenced transcriptome of *Hematodinium* sp., a member of the early diverging and parasitic group of dinoflagellates, the Syndiniales (Danne et al.

2012). Until a genome sequence is available, it is unclear whether FAD-dependent glycerol-3-phosphate dehydrogenase is truly missing in *Hematodinium*. However, its absence raises the possibility that the alveolate copy of this gene was lost in dinoflagellates following the divergence with perkinsids, and later the function was reacquired in some lineages via HGT.

Another potential entry point for electrons into oxidative phosphorylation is via malate:quinone oxidoreductase (MQO), which reduces FAD and passes electrons onto ubiquinone in the transport chain. First described in bacteria (Kather et al. 2000), MQO has also been found in apicomplexans (presumably acquired via HGT, Gardner et al. 2002). MQO has already been described in dinoflagellates and was recovered in our *A. tamarense* Group IV transcriptome, but the dinoflagellate sequences do not group with apicomplexans in phylogenetic trees as would be expected if the gene was horizontally acquired in DALCA (Nosenko and Bhattacharya 2007). Instead, the dinoflagellate MQOs group with haptophytes, cryptophytes, and actinobacteria albiet with only weak to moderate support for the specific branching pattern (supplementary fig. S5C, Supplementary Material online). In contrast, the *P. marinus* MQO groups with apicomplexans, suggesting that the apicomplexan and *P. marinus* version of the gene was acquired in DALCA, and that dinoflagellates at some point lost this ancestral copy and replaced it with an algal version of the gene following the split from the Perkinsidae.

Pyruvate Dehydrogenase

In addition to the loss of Complex I, our ancestral gene content reconstruction suggests DALCA also lost the canonical eukaryotic enzymes involved in pyruvate oxidation (PDHA, PDHB, and DLAT). Pyruvate dehydrogenase (PDH) is responsible for the conversion of pyruvate to acetyl-CoA and is required for the TCA cycle. Despite missing the canonic PDH enzymes, labeling studies of *P. marinus* suggest that pyruvate metabolism is functional in this organism (Danne et al. 2012). The *A. tamarense* Group IV transcriptome revealed two other enzymes capable of catalyzing this reaction. The first is a bacterial copy of pyruvate dehydrogenase (aceE, K00163). Our phylogenetic analysis shows that dinoflagellate aceE forms a monophyletic group sister to actinobacteria (supplementary fig. S5*D*, Supplementary Material online).

In addition to aceE, the *A. tamarense* Group IV transcriptome contains an O_2-sensitive pyruvate:NADP+oxidoreductase (PNO) that also catalyzes pyruvate oxidation. PNO is a fusion protein containing an N-terminal pyruvate:ferredoxin oxidoreductase domain, found in amitochondriate protists, and a C-terminal NADPH-cytochrome P450 reductase (CYPOR) domain that is ubiquitous in eukaryotes (Rotte et al. 2001). PNO has been described in the apicomplexan *Cryptosporidium hominis* (Rotte et al. 2001; Xu et al. 2004), and our analysis also identified copies of PNO in the perkinsid *P. marinus.* Although the phylogenetic trees of the individual domains are poorly resolved, SH tests could not reject the monophyly of *C. hominis*, *P. marinus*, and dinoflagellate PNO (supplementary fig. S5*E* and *F*, Supplementary Material online) and suggests that this gene had been acquired by DALCA and subsequently lost in other apicomplexans.

TCA Cycle

Another gene lost in DALCA related to mitochondrial metabolism was the NAD(H)-dependent form of isocitrate dehydrogenase (NAD-IDH), an enzyme involved in the forward reaction of the TCA cycle. The absence of NAD-IDH in apicomplexans has led to speculation that the forward cycle may not be working in these organisms (Olszewski and Llinás 2011). However, the lack of this enzyme in dinoflagellates and *Perkinsus*, despite possessing functional TCA cycles, has forced a reevaluation of the timing and consequences of gene loss in apicomplexans (Danne et al. 2012). One possible alternative for NAD-IDH is the NADP(H)-dependent version of the enzyme (NADP-IDH), which is typically involved in the reverse reaction of the TCA cycle. Two versions of the NADP(H)-dependent IDH were identified in dinoflagellates: a eukaryotic dimer (NADP-IDH-I) and a less common monomer (NADP-IDH-II) form of the enzyme. Dinoflagellate NADP-IDH-I groups with *P. marinus* and apicomplexans in phylogenetic trees consistent with it being vertically inherited (supplementay fig. S5*G*, Supplementary Material online). In contrast, NADP-IDH-II is only found in photosynthetic eukaryotes and bacteria, a pattern indicative of EGT (Nosenko and Bhattacharya 2007), although the source of the dinoflagellate copy cannot be determined in our analysis (supplementary fig. S5*H*, Supplementary Material online). The reason for the acquisition of NADP-IDH-II in these algae is unknown, but the enzyme is involved in adaptation to cold temperatures in some marine bacteria (Suzuki et al. 1995).

A similar pattern of functional overlap in ancestral and horizontally acquired genes has occurred in the TCA enzyme fumarase. Dinoflagellates have both class I and class II versions of this enzyme. Class I is also present in *P. marinus* and apicomplexans, and—although the ML best tree does not recover alveolate monophyly—the SH analysis suggests that the alveolate fumarase-I are all related (supplementary fig. S5*I*, Supplementary Material online). In contrast, fumarase-II is present in ciliates, but the SH test rejected dinoflagellate–ciliate monophyly. Instead dinoflagellate fumarase-II groups with bacteria and green algae suggesting that the gene was acquired via HGT rather than inherited from an alveolate common ancestor (supplementary fig. S5*J*, Supplementary Material online).

Genes of Bacterial Origin in Dinoflagellates

Dinoflagellates grouped with bacteria in gene phylogenies for 92 KEGG-annotated *A. tamarense* Group IV contigs (10.5%

of phylogenetically informative trees with KEGG annotations). There was no predominant taxonomic signal to suggest that the genes were acquired from one or a few bacterial donors (supplementary fig. S2, Supplementary Material online). Similarly, rather than demonstrating gain of individual pathways, the distribution of KOs suggests that horizontally acquired genes are involved in a variety of metabolic processes (supplementary fig. S3, Supplementary Material online). Metabolism, particularly carbohydrate and amino acid metabolism, made up a larger percentage of KOs in the bacteria-like subset compared with the total set. This is in general agreement with the complexity hypothesis that predicts that genes involved in information processing (e.g., translation and transcription) as well as genes with high connectivity (e.g., ribosomal proteins) are less likely to be horizontally transferred (Jain et al. 1999; Cohen et al. 2011).

Many bacteria-like sequences in dinoflagellates appear to overlap or replace eukaryotic functional analogs. Dinoflagellates posses a bacterial valine:pyruvate aminotransferase (avtA) responsible for valine biosynthesis (supplementary fig. S5*K*, Supplementary Material online) (Marienhagen and Eggeling 2008). Intriguingly, *A. tamarense* Group IV also retains the eukaryotic functional equivalent, a branched-chain amino acid aminotransferase (ilvE). Ciliates and *P. marinus* also contain ilvE, but the gene has been lost in apicomplexans. The phylogenetic tree of ilvE offers poor resolution, but the test for alveolate monophyly suggests that the gene has been vertically inherited in dinoflagellates (supplementary fig. S5*L*, Supplementary Material online). The driving force behind the acquisition of avtA in addition to ilvE is unknown, but this is not the only case of functional overlap between vertically and horizontally acquired genes in dinoflagellates. The presence of bacterial histone-like proteins in addition to canonical histone proteins and the replacement of eukaryotic RuBisCO with bacterial form II RuBisCO are two well-known, published examples of this phenomenon (Morse et al. 1995; Hackett et al. 2005; Janouskovec et al. 2010; Lin et al. 2010). Our phylogenetic analysis of mitochondria-related genes provide additional instances: in pyruvate metabolism, dinoflagellates possess PNO and aceE in place of eukaryotic PDH; and in the TCA cycle, these organisms have a putatively horizontally acquired fumarase-II in addition to the vertically retained fumarase-I.

Perhaps, the best example of functional overlap between vertically and horizontally acquired genes in dinoflagellates is the molecular chaperone grpE. Three copies of this protein were identified in the *A. tamarense* Group IV transcriptome, each one with a different evolutionary history. Phylogenetic analysis of the first copy shows dinoflagellates grouping sister to *P. marinus* and other alveolates, indicative of vertical inheritance (supplementary fig. S5*M*, Supplementary Material online). The second dinoflagellate grpE groups with rhodophytes and "chromalveolate" algae (supplementary fig. S5*N*, Supplementary Material online). This topology is suggestive of EGT via plastid endosymbiosis (Li et al. 2006; Baurain et al. 2010). The final dinoflagellate grpE groups with bacteria (supplementary fig. S5*O*, Supplementary Material online). It is not uncommon for organisms to maintain multiple copies of this protein (Hu et al. 2012), but dinoflagellates seemingly maintain three copies from three different sources.

Bacterial-like genes in dinoflagellates also serve novel functions. One such example is phosphatidylserine synthase, an enzyme important for programmed cell death and apoptosis (Blankenberg et al. 1998). Phylogenetic analysis shows the dinoflagellate phosphatidylserine synthase sequences grouping with proteobacterial genes (supplementary fig. S5*P*, Supplementary Material online). Citrate lyase beta subunit (CitE) shows the same phylogenetic pattern (supplementary fig. S5*Q*, Supplementary Material online). Citrate lyase, the first enzyme of the reverse TCA cycle in bacteria, is a three-subunit enzyme, but only CitE has been identified in dinoflagellates. Although it is possible that the other subunits are missing from current surveys of dinoflagellate genes, it is more likely—given the intensity of EST and transcriptome sequencing—that the gene has undergone neofunctionalization and now acts independently in these organisms.

Conclusions

There is a growing awareness that microbial eukaryotes are susceptible to HGT, but perhaps none more so than dinoflagellates. Our analysis of ancestral gene content using KOs indicates that *A. tamarense* Group IV has a large number of genes likely absent in the dinoflagellate–apicomplexan last-common ancestor. The amount of genes gained in *A. tamarense* Group IV is larger than in any of the other 16 species analyzed. As more dinoflagellate sequences become available, we are likely to better resolve the timing of gene losses and acquisitions within the dinoflagellate lineage more broadly, but the *A. tamarense* Group IV transcriptome presents a valuable starting point for exploring the impact of HGT on dinoflagellates as a whole. Phylogenomic analysis supports the interpretation of these genes being acquired via HGT from various sources including bacteria. Although additional analyses are needed to confirm the in silico functional annotations of *A. tamarense* Group IV transcripts, transferred genes appear to be involved in a diverse array of processes, such as carbohydrate, amino acid, and energy metabolism. Mitochondrial metabolism has been particularly impacted by HGT, perhaps driven by mitochondrial genome reduction in the common ancestor of apicomplexans and dinoflagellates. In many cases, horizontally acquired genes functionally replaced missing eukaryotic enzymes (e.g., NADH dehydrogenase and pyruvate dehydrogenase). Other transferred genes coexist with vertically inherited copies (e.g., isocitrate dehydrogenase, fumarase, valine aminotransferase, and molecular chaperones), and still others appear to confer novel functions (e.g., citrate lyase and phosphatidylserine synthase).

Supplementary Material

Supplementary figures S1–S5 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Michael Sanderson, Betsy Arnold, Michael Nachman, Don Anderson, and Jason Slot for providing helpful feedback on the manuscript. They also thank Susan Miller, William Driscoll, and Bonnie Hurwitz for computational and statistical advice. J.H.W. was supported by the NSF IGERT Program in Comparative Genomics at the University of Arizona (grant number DGE-0654435). This work was supported by grants from the National Science Foundation (grant numbers OCE-0723498, EF-0732440) and funding provided by the BIO5 Institute at the University of Arizona to J.D.H.

Literature Cited

Abrahamsen MS, et al. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304:441–445.

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 52: 399–451.

Andersson JO. 2009. Gene transfer and diversification of microbial eukaryotes. Annu Rev Microbiol. 63:177–193.

Andersson JO, Sjogren A, Davis L, Embley TM, Roger A. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. Curr Biol. 13:94–104.

Archibald JM. 2009. The puzzle of plastid evolution. Curr Biol. 19: R81–R88.

Armbrust EV, et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86.

Aury J, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178.

Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. 2004. Dinoflagellate expressed sequence tag data indicate massive transfer of chloroplast genes to the nuclear genome. Protist 155: 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.

Baurain D, Brinkmann H, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes and stramenopiles. Mol Biol Evol. 27:1698–1709.

Blankenberg FG, et al. 1998. *In vivo* detection and imaging of phosphatidylserine expression during programmed cell death. Proc Natl Acad Sci U S A. 95:6349–6354.

Bowler C, et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature 456:239–244.

Brayton KA, et al. 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 3: 1401–1413.

Brinkmann H, Van der Giezen M, Zhou Y, De Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 54:743–757.

Burki F, Shalchian-Tabrizi K, Minge M. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2:e790.

Burleigh J, Bansal M, Eulenstein O. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst Biol. 60: 117–125.

Chan CX, et al. 2012. Analysis of *Alexandrium tamarense* (Dinophyceae) genes reveals the complex evolutionary history of a microbial eukaryote. J Phycol. 48:1130–1142.

Cock JM, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465:617–621.

Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28:1481–1489.

Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912.

Curtis BA, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature 492:59–65.

Danne JC, Gornik SG, MacRae JI, McConville MJ, Waller RF. 2012. Alveolate mitochondrial metabolic evolution: dinoflagellates force reassessment of the role of parasitism as a driver of change in apicomplexans. Mol Biol Evol. 30:123–139.

Deschamps P, Moreira D. 2012. Reevaluating the green contribution to diatom genomes. Genome Biol Evol. 4:683–688.

Doolittle WF. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14:307–311.

Douglas SE. 1998. Plastid evolution: origins, diversity, trends. Curr Opin Genet Dev. 8:655–661.

Eichinger L, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43–57.

Erdner DL, Anderson DM. 2006. Global transcriptional profiling of the toxic dinoflagellate *Alexandrium fundyense* using massively parallel signature sequencing. BMC Genomics 7:88.

Fast N, Xue L, Bingham S, Keeling PJ. 2002. Re-examining alveolate evolution using multiple protein molecular phylogenies. J Eukaryot Microbiol. 49:30–37.

Gajria B, et al. 2008. ToxoDB: an integrated *Toxoplasma gondii* database resource. Nucleic Acids Res. 36:D553–D556.

Gardner M, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511.

Gobler CJ, et al. 2011. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. Proc Natl Acad Sci U S A. 108: 4352–4357.

Gouzy J, Carrere S, Schiex T. 2009. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25:670–671.

Hackett JD, Anderson DM, et al. 2004. Dinoflagellates: a remarkable evolutionary experiment. Am J Bot. 91:1523–1534.

Hackett JD, Yoon HS, et al. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. Curr Biol. 14:213–218.

Hackett JD, et al. 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics 6:80.

Hackett JD, et al. 2013. Evolution of saxitoxin synthesis in cyanobacteria and dinoflagellates. Mol Biol Evol. 30:70–78.

Hou Y, Lin S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS One 4:e6978.

Hu C, Lin S-Y, Chi W-T, Charng Y-Y. 2012. Recent gene duplication and subfunctionalization produced a mitochondrial GrpE, the nucleotide exchange factor of the Hsp70 complex, specialized in thermotolerance to chronic heat stress in *Arabidopsis*. Plant Physiol. 158:747–758.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96: 3801–3806.

Janouskovec J, Horak A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci U S A. 107:10949–10954.

Kather B, Stingl K, van der Rest ME, Altendorf K, Molenaar D. 2000. Another unusual type of citric acid cycle enzyme in *Helicobacter pylori*: the malate:quinone oxidoreductase. J Bacteriol. 182: 3204–3209.

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, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9:605–618.

Kerscher S, Dröse S, Zickermann V, Brandt U. 2008. The three families of respiratory NADH dehydrogenases. Results Probl Cell Differ. 45: 185–222.

LaJeunesse T, Lambert G, Andersen R, Coffroth M, Galbraith D. 2005. *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates. J Phycol. 41:880–886.

Lantsman Y, Tan KSW, Morada M, Yarlett N. 2008. Biochemical characterization of a mitochondrial-like organelle from *Blastocystis* sp. subtype 7. Microbiology 154:2757–2766.

Li S, Nosenko T, Hackett JD, Bhattacharya D. 2006. Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. Mol Biol Evol. 23:663–674.

Lilly EL, Halanych KM, Anderson DM. 2007. Species boundaries and global biogeography of the *Alexandrium tamarense* complex (Dinophyceae). J Phycol. 43:1329–1338.

Lin S, Zhang H, Zhuang Y. 2010. Spliced leader–based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. Proc Natl Acad Sci U S A. 107:20033–20038.

Lin SS, Gross U, Bohne W. 2011. Two internal type II NADH dehydrogenases of *Toxoplasma gondii* are both required for optimal tachyzoite growth. Mol Microbiol. 82:209–221.

Liu K, Linder CR, Warnow T. 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One 6:e27731.

Liu L, Hastings J. 2006. Novel and rapidly diverging intergenic sequences between tandem repeats of the luciferase genes in seven dinoflagellate species. J Phycol. 42:96–103.

Marienhagen J, Eggeling L. 2008. Metabolic function of *Corynebacterium glutamicum* aminotransferases AlaT and AvtA and impact on L-valine production. Appl Environ Microbiol. 74:7457–7462.

Martin W, Herrmann R. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118:9–17.

Maruyama S, Suzaki T, Weber APM, Archibald JM, Nozaki H. 2011. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evol Biol. 11:105.

Melo A, Bandeiras T. 2004. New insights into type II NAD(P)H:quinone oxidoreductases. Microbiol Mol Biol Rev. 68:603–616.

Minge MA, et al. 2010. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. BMC Evol Biol. 10:191.

Moller IM. 2001. Plant mitochondria and oxidative stress: electron transport, NADPH turnover, and metabolism of reactive oxygen species. Annu Rev Plant Physiol Plant Mol Biol. 52:561–591.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.

Morse D, Salois P, Markovic P, Hastings J. 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. Science 268:1622–1624.

Moustafa A, Bhattacharya D. 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. BMC Evol Biol. 8:6.

Moustafa A, et al. 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.

Nosenko T, Bhattacharya D. 2007. Horizontal gene transfer in chromalveolates. BMC Evol Biol. 7:173.

Nosenko T, et al. 2006. Chimeric plastid proteome in the Florida "red tide" dinoflagellate *Karenia brevis*. Mol Biol Evol. 23:2026–2038.

Ochman H, Lawrence J, Groisman E. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Olszewski KL, Llinás M. 2011. Central carbon metabolism of *Plasmodium* parasites. Mol Biochem Parasitol. 175:95–103.

Orr RJS, Stuken A, Murray SA, Jakobsen KS. 2013. Evolutionary acquisition and loss of saxitoxin biosynthesis in dinoflagellates: the second "core" gene, sxtG. Appl Environ Microbiol. 79:2128–2136.

Pain A, et al. 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science 309:131–133.

Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 37:1372–1375.

Palenik B, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 104:7705–7710.

Parfrey LW, et al. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol. 59:518–533.

Price MN, Dehal PS, Arkin AP. 2009. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26:1641–1650.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Qiu H, Yang EC, Bhattacharya D, Yoon HS. 2012. Ancient gene paralogy may mislead inference of plastid phylogeny. Mol Biol Evol. 29: 3333–3343.

Raghavendra AS, Padmasree K. 2003. Beneficial interactions of mitochondrial metabolism with photosynthetic carbon assimilation. Trends Plant Sci. 8:546–553.

Reece K, Siddall M, Burreson E, Graves J. 1997. Phylogenetic analysis of *Perkinsus* based on actin gene sequences. J Parasitol. 83:417–423.

Ricard G, et al. 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. BMC Genomics 7:22.

Rokas A, King N, Finnerty J, Carroll SB. 2003. Conflicting phylogenetic signals at the base of the metazoan tree. Evol Dev. 5:346–359.

Rotte C, Stejskal F, Zhu G, Keithly JS, Martin W. 2001. Pyruvate:NADP+ oxidoreductase from the mitochondrion of *Euglena gracilis* and from the apicomplexan *Cryptosporidium parvum*: a biochemical relic linking pyruvate metabolism in mitochondriate and amitochondriate protists. Mol Biol Evol. 18:710–720.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

Slot JC, Hibbett DS. 2007. Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. PLoS One 2:e1097.

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

Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol Biol. 11:259.

Stoecker D. 1998. Conceptual models of mixotrophy in planktonic protists and some ecological and evolutionary implications. Eur J Protistol. 34: 281–290.

Stover NA, et al. 2006. *Tetrahymena* Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. Nucleic Acids Res. 34:D500–D503.

Striepen B, et al. 2004. Gene transfer in the evolution of parasite nucleotide biosynthesis. Proc Natl Acad Sci U S A. 101:3154–3159.

Stuken A, et al. 2011. Discovery of nuclear-encoded genes for the neurotoxin saxitoxin in dinoflagellates. PLoS One 6:e20096.

Suzuki M, Sahara T, Tsuruha J, Takada Y, Fukunaga N. 1995. Differential expression in *Escherichia coli* of the *Vibrio* sp. strain ABE-1 icdI and icdII genes encoding structurally different isocitrate dehydrogenase isozymes. J Bacteriol. 177:2138–2142.

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7: e1001284.

Waller RF, Jackson CJ. 2009. Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. Bioessays 31: 237–245.

Wisecaver JH, Hackett JD. 2010. Transcriptome analysis reveals nuclear-encoded proteins for the maintenance of temporary plastids in the dinoflagellate *Dinophysis acuminata*. BMC Genomics 11:366.

Wisecaver JH, Hackett JD. 2011. Dinoflagellate genome evolution. Annu Rev Microbiol. 65:369–387.

Worden AZ, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324:268–272.

Xu P, et al. 2004. The genome of *Cryptosporidium hominis*. Nature 431: 1107–1112.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Associate editor: Martin Embley