Algae
Volume 21(1): 1-10, 2006

[Review]

Phylogenomics and its Growing Impact on Algal Phylogeny and Evolution

Adrian Reyes-Prieto, Hwan Su Yoon and Debashish Bhattacharya*

University of Iowa, Department of Biological Sciences and the Roy J. Carver Center for Comparative Genomics, 446 Biology Building, Iowa City, IA 52242, USA

Genomic data is accumulating in public database at an unprecedented rate. Although presently dominated by the sequences of metazoan, plant, parasitic, and picoeukaryotic taxa, both expressed sequence tag (EST) and complete genomes of free-living algae are also slowly appearing. This wealth of information offers the opportunity to clarify many long-standing issues in algal and plant evolution such as the contribution of the plastid endosymbiont to nuclear genome evolution using the tools of comparative genomics and multi-gene phylogenetics. A particularly powerful approach for the automated analysis of genome data from multiple taxa is termed phylogenomics. Phylogenomics is the convergence of genomics science (the study of the function and structure of genes and genomes) and molecular phylogenetics (the study of the hierarchical evolutionary relationships among organisms, their genes and genomes). The use of phylogenetics to drive comparative genome analyses has facilitated the reconstruction of the evolutionary history of genes, gene families, and organisms. Here we survey the available genome data, introduce phylogenomic pipelines, and review some initial results of phylogenomic analyses of algal genome data.

Key Words: algal evolution, endosymbiosis, genome database, genomics, phylogenomics

INTRODUCTION

The most recent version of NCBI (National Center for Biotechnology Information), November 2005, that contains information from several international databases (GenBank, European Molecular Biology Laboratory, and DNA DataBank of Japan) exceeds 100 gigabases in size. This database is predicted to grow exponentially in the upcoming years [www.ncbi.nlm.nih. gov/Genbank/index.html]. Nearly 50% of the total sequences available in NCBI come from more than 300 complete genomes that have been deposited since 2000 (Benson *et al.* 2006). During 2005 and early 2006, more than 80 genomes were sequenced to completion, including 9 eukaryotic species [www.genomesonline. org].

Analyses of genome data have diverse aims, such as the elucidation of gene homology, detection of polymorphisms, gene content, assignment of potential function (annotation), gene arrangements (clusters), expression profiles, evolutionary history, and to contrast genome data (comparative genomics). The management of large amounts of sequence data requires the use of practical informatics and computational approaches. In the last decade several computational tools for similarity and pattern searches have been developed to analyze comprehensively the large amount of information contained in genome sequences, giving rise to the field of *bioinformatics*. The spread of sequence homology-search algorithms (by similarity) such as the rigorous FASTA (Pearson and Lipman 1988) or less rigorous but faster BLAST (Altschul *et al.* 1990) to survey the public databases has had an unprecedented impact in evolutionary biology, as well as on structural and molecular biology, allowing the rapid identification and retrieval of homologous sequences from the immense databases. At the same time, the development of multiple sequence alignment programs facilitated the identification of conserved regions with potential functional or structural significance, an invaluable source of data to test hypotheses regarding molecular homology. These and other *bioinformatics* tools ultimately accelerated the field of phylogenetics and evolutionary biology.

GENOMICS SCIENCES

The 1.8 megabase pair (Mbp) genome of the gammaprotobacterium *Haemophilus influenzae* was the first to be

*Corresponding author (debashi-bhattacharya@uiowa.edu)

completed (Fleischmann *et al.* 1995). These data heralded the arrival of "genomics", and during subsequent years several other microbial genomes were sequenced to completion including that of the model bacterium *Escherichia coli* K12 (4.6 Mbp and 4,288 protein coding genes; Blattner *et al.* 1997). In addition, within an interval of 6 years (1992-1998) the 16 chromosomes (12.1 Mbp) of the budding yeast *Saccharomyces cerevisiae* were determined, constituting the first fully sequenced eukaryotic genome. Analyses demonstrated that the yeast genome resulted from an ancient duplication, and also revealed that the complete set of genes for this free living eukaryote is around 6,000 open reading frames (ORFs) (Goffeaeu *et al.* 1996) The complete genome of the first multicellular eukaryotes were assembled in the late nineties and early in this century; i.e., the nematode worm *Caenorhabditis elegans* (97 Mbp and 19,000 genes) (*C. elegans Sequence consortium*, 1998), the fruit fly *Drosophila melanogaster* (120 Mbp, ~13,600 genes) (Adams *et al.* 2000), the angiosperm *Arabidopsis thaliana* (125 Mbp and 25,500 genes) (*Arabidosis initiative* 2000) and human (2,900 Mbp and 20,000 to 25,000 genes; Lander *et al.* 2001, Venter *et al.* 2001). Currently, 102 eukaryotic genomes have been sequenced (the majority are in the assembly process) with most of them limited to fungi, animals, and land plants with relatively sparse sampling of other eukaryotic groups. In addition there are 177 projects underway for eukaryote species including several photosynthetic taxa (http://www-users.york.ac.uk/ ~ct505/PhD_ Project5/Eukaryote_Homepage.htm)

The genomes of the red alga *Cyanidioschyzon merolae* (Matsuzaki *et al.* 2004) and of the green alga *Chlamydomonas reinhardtii* comprise two primary photosynthetic eukaryotes (i.e., containing a plastid that resulted from the primary cyanobacterial endosymbiosis; Bhattacharya *et al.* 2004) that have recently been completed (genome.jgi-psf.org/Chlre3/Chlre3.info. html). Apart from their potential biomedical or economic importance, these and other algal genome projects are providing key data for groups of significant importance to understanding the evolution of photosynthetic eukaryotes. Members of the three Plantae lineages (red algae, green algae, and glaucophytes; also referred to as Archaeplastida, Adl *et al.* 2005) are currently under analysis in several genomic-scale sequencing projects (complete genome or EST libraries), and a similar situation exists for the principal groups of alveolates (ciliates, dinoflagellates, and apicomplexans) and chromists (haptophytes, cryptophytes, and stramenopiles; see Table 1). Coincident with the release of the algal genome data, some comparisons have been made to understand common characteristics of these algae in an evolutionary context. Homology-based analysis by BLAST search of the centric diatom *Thalassiosira pseudonana* complete genome (11,242 genes) against *Cyanidioschyzon merolae* (5,331 genes), *Arabidosis thaliana* (25,500 genes) and the cyanobacterium *Nostoc* sp., showed that 1,194 proteins are conserved among the three photosynthetic eukaryotes and the cyanobacterium (E value < $1e^{-5}$) with most of these involved in plastid function and being potentially derived from the ancestral endosymbiont (Ambrust *et al.* 2004). Similar comparative analyses with the genomes of *C. merolae*, *C. reinhardtii* (15,200 genes), and *A. thaliana* suggest that the majority of genes involved with basic biological process are orthologs among these taxa, but interestingly *C. reinhardtii* possesses a larger set of genes for DNA packing, the cytoskeleton, and the flagellar apparatus, which probably explains the increase in genome size in this motile alga (Misumi *et al.* 2005). Recently, a genome similarity evaluation of the pennate diatom *Phaoedactylum tricornutum* (5,108 unique ESTs) against *T. pseudonana* (Montsant *et al.* 2005), *C. merolae*, and *C. reinhardtii*, indicated that both diatoms exclusively share 820 genes (16% of *P. tricornutum* sequences analyzed by BLAST with an E value < $1e^{-4}$). As expected, both diatoms have more genes in common between them than with the other two algae. Additionally, the diatoms show a higher sequence similarity with genes in the red alga than with the green, likely indicating a red algal ancestry of genes derived from the chromalveolate secondary endosymbiosis (Montsant *et al.* 2005). However, when considering the testing of evolutionary hypotheses, BLAST analyses of genomic data comprise significantly weaker and potentially misleading criteria when compared to the use of robust phylogenetic inference. The latter approach has been shown to outperform homology search methods due to ambiguities caused by ancient gene duplications (paralogy), biased base composition, or lateral gene transfer (Eisen 1998; Zmasek and Eddy 2002; Sjolander 2004).

THE EMERGENCE OF PHYLOGENOMICS

Phylogenomics is the convergence of the *Genomics Sciences* (the study of the function and structure of genes and genomes) and *Molecular Phylogenetics* (the study of the hierarchical evolutionary relationships among organisms, their genes and genomes). The use of phylogenetics

Table 1. List of genome projects from algae and close relatives. The row color code is as follow: glaucophytes (purple), green algae (green), red algae (red), and chromalveolates (brown)

LINEAGE	SPECIES	PROJECT TYPE	GENOME SIZE	INSTITUTION or CONSORTIUM	WWW LINK
GLAUCOPHYTA	*Cyanophora paradoxa*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/organism_tt.php?orgID=CP
GLAUCOPHYTA	*Cyanophora paradoxa*	EST		University of Iowa	http://www.biology.uiowa.edu/debweb/html/EndosymbGeneTransferNASA.php
GLAUCOPHYTA	*Glaucocystis nostochinearum*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Mesostigma viride*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Micromonas pusilla NOUM17(RCC 299)*	genome	15 Mb	JGI	http://www.jgi.doe.gov/sequencing/DOEmicrobes2005.html
GREEN ALGAE	*Micromonas pusilla CCMP490(RCC 114)*	genome	15 Mb	JGI	http://www.jgi.doe.gov/sequencing/DOEmicrobes2005.html
GREEN ALGAE	*Volvox carteri*	genome	15 Mb	Stanford JGI	http://www-shgc.stanford.edu/datarelease/Volvox.data.html
GREEN ALGAE	*Ostreococcus* sp. CCE9901	genome	8-10 Mb	JGI	http://www.jgi.doe.gov/sequencing/why/CSP2006/Ilostreococcus.html
GREEN ALGAE	*Ostreococcus tauri*	genome	11.5 Mb	Laboratoire Arago	---
GREEN ALGAE	*Dunaliella salina* UTEX	genome	130 Mb	JGI	---
GREEN ALGAE	*Chlorella vulgaris* C-169	genome	40 Mb	JGI	---
GREEN ALGAE	*Chlamydomonas reinhardtii**	genome COM	100 Mb	JGI	http://genome.jgi-psf.org/Chlre3/Chlre3.home.html
GREEN ALGAE	*Nephroselmis olivacea*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Prototheca wickerhamii*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Scenedesmus obliquus*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Acetabularia acetabulum*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
GREEN ALGAE	*Chlamydomonas incerta*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
RED ALGAE	*Galdieria sulphuraria*	genome		Michigan State University	http://genomics.msu.edu/galdieria/
RED ALGAE	*Porphyra yezoensis*	EST		Kazusa DNA Research Institute	http://www.kazusa.or.jp/
RED ALGAE	*Cyanidioschyzon merolae* 10D	genome COM	16.5 Mb	U.Tokyo, Rikkyo U., Saitama U., Kumamoto U.	http://merolae.biol.s.u-tokyo.ac.jp/
STRAMENOPILES-Bacilliarophyta	*Pseudonitzschia multiseries*	genome	25 Mb	JGI	http://genome.imb-jena.de/ESTTAL/cgi-bin/Pseudo-nitzschia.pl
STRAMENOPILES-Bacilliarophyta	*Phaeodactylum tricornutum*	genome		JGI	---
STRAMENOPILES-Bacilliarophyta	*Thalassiosira pseudonana* CCMP 1335	genome COM		JGI	http://genome.jgi-psf.org/thaps1/thaps1.home.html
STRAMENOPILES-Chrysophyta	*Ochromonas danica*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
STRAMENOPILES-Oomycota	*Phytophthora infestans*	genome	250 Mb?	MIT	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
STRAMENOPILES-Oomycota	*Phytophthora ramonum*	genome	24 Mb	JGI	http://genome.jgi-psf.org/ramorum1/ramorum1.home.html
STRAMENOPILES-Oomycota	*Phytophthora sojae*	genome	62 Mb	JGI	http://genome.jgi-psf.org/sojae1/sojae1.home.html
STRAMENOPILES-Pelagophyceae	*Aureococcus anophagefferens*	genome	32 Mb	JGI	---
STRAMENOPILES-Phaeophyceae	*Ectocarpus siliculosus*	genome		Institute de la Genomique Marine	http://www.sb-roscoff.fr/GIS-genomique-marine/
ALVEOLATA-Apicomplexa	*Toxoplasma gondii**	genome	80.Mb	TIGR	http://www.tigr.org/tdb/e2k1/tga1/intro.shtml
ALVEOLATA-Apicomplexa	*Toxoplasma gondii**	genome	80.Mb	Cambridge University, Beowulf Genomics	---
ALVEOLATA-Apicomplexa	*Perkinsus marinus*	genome	28.Mb	TIGR, University of Maryland COMB	http://www.tigr.org/tdb/e2k1/pmg/help.shtml
ALVEOLATA-Apicomplexa	*Plasmodium falciparum*	genome COM	22.9 Mb	Malaria Genome Project Consortium	http://www.sanger.ac.uk/Projects/P_falciparum/who&what.shtml
ALVEOLATA-Apicomplexa	*Plasmodium yoelii yoelii*	genome COM	23.1 Mb	DOD, TIGR, NMRC	http://www.tigr.org/tdb/e2k1/pya1/
ALVEOLATA-Apicomplexa	*Cryptosporidium parvum*	genome COM	10.4 Mb	University of Minnesota, UCSF, NIAID	---
ALVEOLATA-Apicomplexa	*Theileria parva*	genome COM	8.3 Mb	TIGR, ILRI	http://www.tigr.org/tdb/e2k1/tpa1/intro.shtml
ALVEOLATA-Ciliates	*Ichthyophthirius multifiliis* G5	EST		FUNGEN	---
ALVEOLATA-Ciliates	*Paramecium tetraurelia*	genome		GENOSCOPE	http://www.genoscope.cns.fr/externe/English/Projets/Projet_FN/FN.html
ALVEOLATA-Ciliates	*Tetrahymena thermophila.*	genome		University of Chicago, TIGR	http://www.tigr.org/tdb/e2k1/ttg/intro.shtml
ALVEOLATA-Dinophyceae	*Alexandrium tamarense* CCMP1598	EST		University of Iowa	http://www.biology.uiowa.edu/debweb/html/AlgGen_Y1.php
ALVEOLATA-Dinophyceae	*Amphidinium operculatum*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
ALVEOLATA-Dinophyceae	*Heterocapsa triquetra*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
ALVEOLATA-Dinophyceae	*Karenia brevis*	EST		University of Iowa, JGI	http://www.biology.uiowa.edu/debweb/html/Karenia_brevis.php
HAPTOPHYTA	*Pavlova lutheri*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
HAPTOPHYTA	*Isochrysis galbana* CCMP1323	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
HAPTOPHYTA	*Emiliania huxleyi* CCMP371	EST		University of Iowa	http://www.biology.uiowa.edu/debweb/html/AlgGen.php
HAPTOPHYTA	*Emiliania huxleyi* 1516	Genome	5 Mb	JGI	http://www.jgi.doe.gov/sequencing/DOEmicrobes2003.html
HAPTOPHYTA	*Emiliania huxleyi*	EST		University of Southampton	---
CRYPTOPHYTA	*Guillardia theta*	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
CRYPTOPHYTA	*Goniomonas* sp. ATCC 50108	EST		PEP	http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php
CRYPTOPHYTA	*Goniomonas* sp.	EST		University of Iowa	---

genome COM: Completely sequenced
EST: Expressed Sequence Tags
Mb: Mega base pairs (1 x10^6 bp)

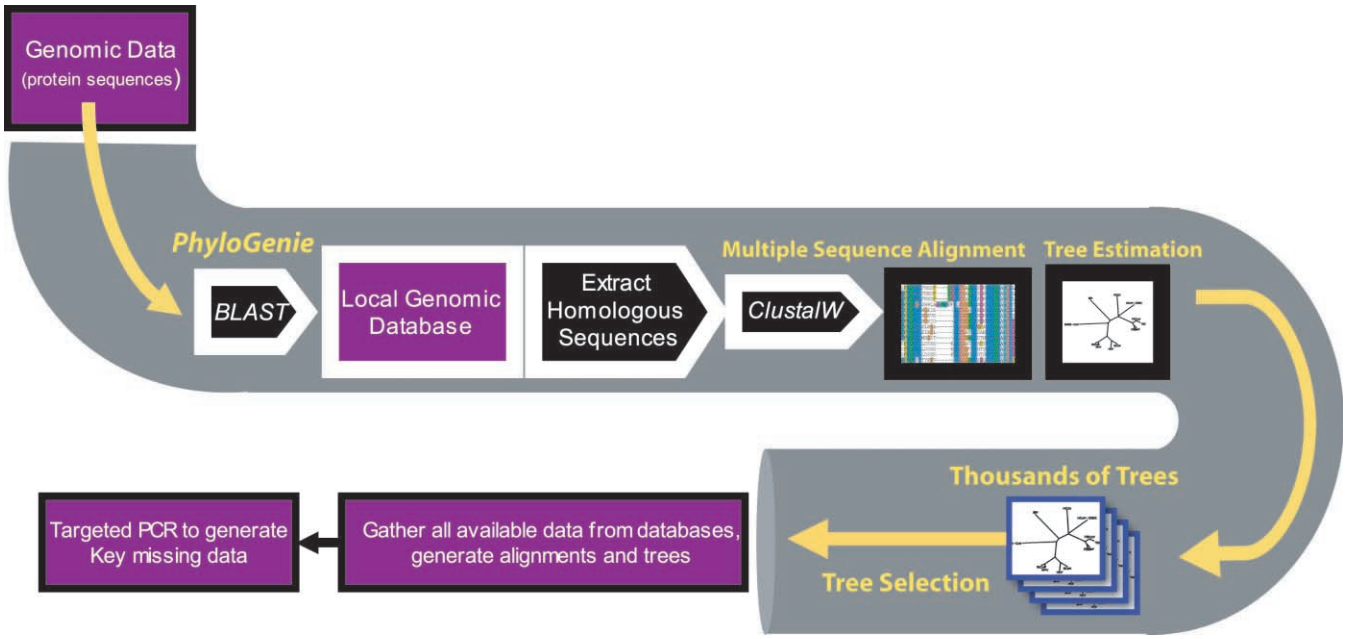

Fig. 1. A typical phylogenomic pipeline. An important characteristic of this particular pipeline is the ability to modify the local genomic database in order to test different phylogenetic hypotheses. The analysis of thousands of bootstrapped phylogenetic trees is a critical step to carry out the next round of detailed analyses using maximum likelihood or maximum parsimony methods. The assembled software is publicly available (see text for details).

to drive comparative genome analyses facilitates the reconstruction of the evolutionary history of genes, gene families, and organisms. Equally important, phylogenomics is used for gene annotation, prediction of molecular function, reconstruction of metabolic pathways, and ultimately, to correlate function with changes in molecular structure (Eisen 1998). Phylogenomics has become a relevant approach given the thousands of predicted hypothetical proteins that are identified using genome-scale projects in the absence, in most cases, of functional information or the associated annotation. From the perspective of algal phylogeny, the emerging vast amounts of sequence information from a growing number of diverse genomes creates an ideal opportunity to evaluate evolutionary relationship among algal groups (e.g., Plantae, Rodriguez-Ezpeleta *et al.* 2005; chromalveolates, Li *et al.* 2006) through the automated phylogenetic analyses of whole protein repertoires (i.e., the phylomes).

In addition to providing access to genome data, phylogenomics requires adequate *bioinformatics* tools to manage and process the sequence information and to generate an efficient high-throughput approach for genome-wide phylogenetic analyses. Once the query genome data has been defined, organized and annotated, a set of fundamental requirements can be identified to carry out the first round of a phylogenomic analysis: 1) The design and assembly of adequate genome database to be used as reference for the phylogenetic analyses (i.e., including predicted ingroup and outgroup taxa). For the algae very few complete sequences are presently available, but the results of a number of EST projects can be included (e.g., Hackett *et al.* 2004; Yoon *et al.* 2005; Li *et al.* 2006), along with reference genomes such as Opisthokonta, bacterial, and archaeal genomes. 2) Identification and gathering of homologous sequences by similarity searches (BLAST search) against the genome database. The similarity cut-off (e.g., BLAST E value < $1e^{-10}$) is critical to obtain accurate results in the subsequent steps (Martin *et al.* 2002). 3) A fundamental requirement to obtain reliable phylogenetic hypotheses is the generation of high quality multiple sequence alignments. The most widely used algorithm for multiple sequence alignment is the global progressive pairwise strategy (e.g., CLUSTALW, Chenna *et al.* 2003), but the use of any accurate alignment algorithm is feasible. In this regard, it has been suggested that protein sequences with > 30% identity in pairwise comparisons allows satisfactory alignment matches that are consistent with structural conservation (Sjolander 2004). 4) Construction of phylogenetic trees. The primary phylogenetic inference method for several thousand multiple alignments should use a fast method such as neighbor joining; i.e., given the high number of expected topologies and the subsequent bootstrap analyses. 5) Selection of the trees with significant statistical support for partic-

ular sub-trees. The essential step of tree selection requires an exhaustive analysis to identify all candidate phylogenies. Thereafter, the alignments of interest are used for in-depth analyses that incorporate the available taxon sampling and are analyzed under other optimality criterion such as maximum parsimony or maximum likelihood to obtain accurate phylogenetic results. Figure 1 summarizes using a flow diagram the bioinformatic pipeline presently in place in our lab to analyze genome data.

SOFTWARE TOOLS

There are several freely available computational tools that can be used to assemble a *computational pipeline* for phylogenomics. In the following section we do not describe the software used to process raw genome sequence data but assume the availability of processed and annotated genome sequences that are ready for the phylogenomics pipeline (Figure 1).

Some hardware considerations

Today it is possible to store complete genome sequence data to a local hard drive for subsequent analysis. For example, the greater than 25,000 genes of *Arabidopsis thaliana* (www.ncbi.nlm.nih.gov/Genomes/) comprise a single text file that is only 15 megabytes in size. A typical local genome database that includes most relevant prokaryotic and eukaryotic taxa (e.g., 25 individual sets of genome and EST data) has a size of around 250 megabytes and is easily handled by modern computers. The critical space requirements come from the phylogenetic analyses and the massive output files that contain similarity search results, multiple sequence alignments, and thousands of phylogenetic trees. These may require up to 10 gigabytes of hard drive space for a typical analysis.

The core programs

The Similarity search for homolog identification is typically performed using BLAST (Basic Local Alignment Search Tool), which is available for local setup under most commonly used operating systems (www.ncbi.nlm.nih.gov/BLAST/download.shtml). BLAST is a very versatile tool and allows the use of DNA or protein sequences as the query or as a database. The extended use of and free access to the CLUSTAL series of programs (Chenna *et al.* 2003), makes them the obvious choice for the multiple alignment of identified homologoues sequences (ftp-igbmc.u-strasbg.fr/pub/ClustalX/). However, there-are other options, such as T-COFFEE (Notredame *et al.* 2000; igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html) or MUSCLE (Edgar 2004; www.drive5.com/muscle/), which are also publicly available. The subsequent phylogenetic analyses require the use of distance criteria as a first estimation (Neighbor-Joining or Minimum Evolution) to assess the confidence in topologies within a reasonable time frame. There are no theoretical limitations on the use of parsimony or likelihood methods in phylogenomics, but the current algorithms, including the fastest (e.g., PHYML, http://atgc.lirmm.fr/phyml/), would increase significantly the amount of computing resources and the processing time. Several types of public domain software are available to compute distance matrices and to construct distance phylogenetic trees. The final output of the phylogenomic analysis should include thousands of trees and the ideal approach with these data is to select the trees of particular interest. To do this, tools are needed that automate the tree search and testing for the presence/absence of specific topologies. The absence of these computing tools is currently a limiting step. However there are some existing algorithms that are designed to search for patterns among unordered trees (where the branching order of sister clades inside nodes does not matter) that could be useful for phylogenomic tree inspection (cs.nyu.edu/cs/faculty/sasha/papers/treesearch.html and www.aria.njit.edu/biotool), which has been modified to retrieve information from the public phylogenetic tree catalog TreeBASE (www.treebase.org). Our group is currently working on methods to link these types of algorithms for automated tree selection to the phylogenomics output.

The idea of a phylogenomic *pipeline* implies the creation of small but efficient computing programs (*scripts*) that automatically link the core software for high throughput analyses (Fig. 1). These scripts should ideally contain points of control to check the generated files during or after the process. There are some *pipeline scripts* available, and all of them share the central scheme of identifying and selecting homologs in a query genome, to generate multiple sequence alignments, and to estimate phylogenetic trees and their branch support. For example, *PyPhy* (Sicheritz-Ponten and Anderson 2001; www.cbs.dtu.dk/staff/thomas/pyphy/) is a Python script that incorporates an useful graphic interface and incorporates the use of PAUP* for phylogenetics (Swofford 2001), which permits distance or parsimony

tree estimation. The use of PAUP* however requires a private license. *PyPhy* makes use of the accurate annotated protein database Swissprot /TrEMBL (ca.expasy.org/sprot/) as the sampling reference. Additionally, *PyPhy* allows the search for particular tree categories through the hierarchical classification of results and provides links to the annotation tables and biochemical pathway information included in KEGG (www.genome.jp/kegg/). Another application is *RIO* (Zmasek and Eddy 2002: www.genetics.wustl.edu/eddy/forester/), a Perl script that connects several publicly available programs and focuses primarily on automated phylogenomics for the functional organization of orthologs. *RIO* uses the bootstrap resample gene trees approach to assess the consistency of orthology assignment, taking into account the information of high quality alignments and pre-calculated distance matrix from the *Pfam* database (www.sanger.ac.uk/Software/Pfam/).

In our lab, phylogenomics research has been carried out using the flexible *PhyloGenie* (Frickey and Lupas 2004; http://protevo.eb.tuebingen.mpg.de/download) Perl script, that, in contrast to *PyPhy* and *RIO*, allows the user to assemble particular genome database as reference, which offers the flexibility to test specific phylogenetic hypotheses. This facilitates the choice of reference genomes to avoid taxon sampling bias and the inclusion of uninformative or redundant genome data. The phylogenetic estimation capabilities of *PhyloGenie* include distance (Neighbor-Joining) and maximum likelihood criteria (TreePuzzle) (Schimidt *et al.* 2002; www.tree-puzzle.de). As described above, the pipeline scripts require local installation of the core software for sequence analyses.

ALGAL PHYLOGENOMICS

Currently, genome projects (complete or EST) for 50 algae and their non-photosynthetic relatives are underway. The 41 genera are distributed among the major photosynthetic eukaryote lineages, such as Plantae (13 greens + 3 reds + 2 glaucophytes), Chromista (6), Alveolates (5 apicomplexans + 4 dinoflagellates + 3 ciliates), Haptophyta (3), and Cryptophyta (2; see Table 1).

Recent findings

The proposed photosynthetic ancestry of alveolates (i.e., the "chromalveolate" hypothesis; Cavalier-Smith 1999) provides a framework to search for cyanobacterial-derived genes in the genome of non-photosynthetic members of this lineage such as the apicomplexans, in particular for species lacking the remnant apicoplast. A recent phylogenomic analysis of the predicted 5,591 proteins of the apicomplexan *Cryptosporidium parvum* revealed the presence of 7 proteins potentially transferred from the ancestral secondary plastid genome to the nucleus. Interestingly, all of these proteins lack the typical transit peptide for import into the apicoplast, providing evidence for the loss of this organelle in *C. parvum*. Additionally, 24 genes of bacterial origin (non-cyanobacterial) from probable horizontal gene transfers (HGTs) were identified in *C. parvum* (Huang *et al.* 2004). The genes of cyanobacterial (i.e., endosymbiotic) origin constitute only 0.7% of the 954 trees generated with the *PyPhy* database, indicating that loss of the apicoplast resulted in the attendant loss of nuclear genes encoding apicoplast targeted proteins. In a broader context it is possible that the complete loss of the plastid in taxa such as the ciliates may also result in the massive (or complete) loss of plastid targeted proteins.

Recent findings in our lab have documented the migration of typical plastid encoded genes to the nuclear genome of the peridinin containing dinoflagellate *Alexandrium tamarense*, which probably occurred after the split of dinoflagellates and apicomplexans. The analysis of a unique set of 6,480 ESTs identified 48 typical plastid genes encoded in the nuclear genome of *A. tamarense*, 15 of which are encoded in the plastid genome of all other photosynthetic eukaryotes. Moreover, the phylogenetic analyses indicated different origins for plastid-targeted proteins; i.e., the genes *atp*I, *atp*F and *psb*O have the expected red-algal ancestry whereas, *ALA dehydratase* and *tuf*A have a green-algal origin (Hackett *et al.* 2004). Other plastid-targeted proteins have an unresolved ancestry, but indicate likely green or red algal origins. These data combined with the well-established plastid replacements in some photosynthetic dinoflagellates (Chesnik *et al.* 1996; Watanabe 1987; Hackett *et al.* 2003) open several questions about genome evolution. For example, it is not yet known whether the genes acquired from the original red algal secondary endosymbiont were conserved in the nucleus or ortholog substitutions occurred after the establishment of the new plastid (for details, see Hackett *et al.* 2004).

Our most recent phylogenomic analysis focused on identifying cases of endosymbiotic gene transfer in the chromalveolates, using a 5,081 EST unigene set from the haptophyte alga *Emiliania huxleyi* (Li *et al.* 2006) and genome data from other chromalveolates such as *A.*

tamarense (Hackett *et al.* 2004), *Karenia brevis* (Lidie *et al.* 2005) and *T. pseudonana* (Armbrust *et al.* 2004). The results of this study indicate that the majority of the nuclear encoded plastid-targeted proteins have a red-algal origin (17 proteins), whereas two genes have a green-algal ancestry. These results support a red algal origin of the chromalveolate plastid with a relatively minor contribution from green algae (potentially through lateral gene transfers), thereby reinforcing evidence of a common origin of the plastid from other studies (e.g., Fast *et al.* 2001; Harper *et al.* 2005; Yoon *et al.* 2002; Hackett *et al.* 2004).

In summary, although the data are still relatively meager, the emerging trend is for a continued expansion of algal genome sequences in the coming years. The availability of these and other genomes will undoubtedly drive the development of phylogenomics tools, which stand to become one of the pre-eminent approaches for understanding algal gene and genome origin and evolution.

REFERENCES

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., George R.A., Lewis S.E., Richards S., Ashburner M., Henderson S.N., Sutton G.G., Wortman J.R., Yandell M.D., Zhang Q., Chen L.X., Brandon R.C., Rogers Y.H., Blazej R.G., Champe M., Pfeiffer B.D., Wan K.H., Doyle C., Baxter E.G., Helt G., Nelson C.R., Gabor G.L., Abril J.F., Agbayani A., An H.J., Andrews-Pfannkoch C., Baldwin D., Ballew R.M., Basu A., Baxendale J., Bayraktaroglu L., Beasley E.M., Beeson K.Y., Benos P.V., Berman B.P., Bhandari D., Bolshakov S., Borkova D., Botchan M.R., Bouck J., Brokstein P., Brottier P., Burtis K.C., Busam D.A., Butler H., Cadieu E., Center A., Chandra I., Cherry J.M., Cawley S., Dahlke C., Davenport L.B., Davies P., de Pablos B., Delcher A., Deng Z., Mays A.D., Dew I., Dietz S.M., Dodson K., Doup L.E., Downes M., Dugan-Rocha S., Dunkov B.C., Dunn P., Durbin K.J., Evangelista C.C., Ferraz C., Ferriera S., Fleischmann W., Fosler C., Gabrielian A.E., Garg N.S., Gelbart W.M., Glasser K., Glodek A., Gong F., Gorrell J.H., Gu Z., Guan P., Harris M., Harris N.L., Harvey D., Heiman T.J., Hernandez J.R., Houck J., Hostin D., Houston K.A., Howland T.J., Wei M.H., Ibegwam C., Jalali M., Kalush F., Karpen G.H., Ke Z., Kennison J.A., Ketchum K.A., Kimmel B.E., Kodira C.D., Kraft C., Kravitz S., Kulp D., Lai Z., Lasko P., Lei Y., Levitsky A.A., Li J., Li Z., Liang Y., Lin X., Liu X., Mattei B., McIntosh T.C., McLeod M.P., McPherson D., Merkulov G., Milshina N.V., Mobarry C., Morris J., Moshrefi A., Mount S.M., Moy M., Murphy B., MurphyL., Muzny D.M., Nelson D.L., Nelson D.R., Nelson K.A., Nixon K., Nusskern D.R., Pacleb J.M., Palazzolo M., Pittman G.S., Pan S., Pollard J., Puri V., Reese M.G., Reinert K., Remington K., Saunders R.D., Scheeler F., Shen H., Shue B.C., Siden-Kiamos I., Simpson M., Skupski M.P., Smith T., Spier E., Spradling A.C., Stapleton M., Strong R., Sun E., Svirskas R., Tector C., Turner R., Venter E., Wang A.H., Wang X., Wang Z.Y., Wassarman D.A., Weinstock G.M., Weissenbach J., Williams S.M., WoodageT., Worley K.C., Wu D., Yang S., Yao Q.A., Ye J., Yeh R.F., Zaveri J.S., Zhan M., Zhang G., Zhao Q., Zheng L., Zheng X.H., Zhong F.N., Zhong W., Zhou X., Zhu S., Zhu X. and Smith H.O. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185-95.

Adl S.M., Simpson A.G., Farmer M.A., Andersen R.A., Anderson O.R., Barta J.R., Bowser S.S., Brugerolle G., Fensome R.A., Fredericq S., James T.Y., Karpov S., Kugrens P., Krug J., Lane C.E., Lewis L.A., Lodge J., Lynn D.H., Mann D.G., McCourt R.M., Mendoza L., Moestrup O., Mozley-Standridge S.E., Nerad T.A., Shearer C.A., Smirnov A.V., Spiegel F.W. and Taylor M.F. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* **52:** 399-451.

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

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Armbrust E.V., Berges J.A., Bowler C., Green B.R., Martinez D., Putnam N.H., Zhou S., Allen A.E., Apt K.E., Bechner M., Brzezinski M.A., Chaal B.K., Chiovitti A., Davis A.K., Demares M.S., Detter J.C., Glavina T., Goodstein D., Hadi M.Z., Hellsten U., Hildebrand M., Jenkins B.D., Jurka J., Kapitonov V.V., Kroger N., Lau W.W., Lane T.W., Larimer F.W., Lippmeier J.C., Lucas S., Medina M., Montsant A., Obornik M., Parker M.S., Palenik B., Pazour G.J., Richardson P.M., Rynearson T.A., Saito M.A., Schwartz D.C., Thamatrakoln K., Valentin K., Vardi A., Wilkerson F.P. and Rokhsar D.S. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306:** 79-86.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J. and Wheeler D.L. 2006. GenBank. *Nucleic Acids Res.* **34:** D16-D20.

Bhattacharya D., Yoon H.S. and Hackett J.D. 2004. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays* **6:** 50-60.

Blattner F.R., Plunkett G. 3rd, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B. and Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453-74.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012-2018.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting

in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol*. **46:** 347-366.

Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J., Higgins D.G. and Thompson J.D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. **31:** 3497-3500.

Chesnick J.M., Morden C.W. and Schmieg A.M. 1996. Identity of the endosymbiont of *Peridinium foliaceum* (Pyrrophyta): analysis of the rbcLS operon. *J. Phycol*. **32:** 850-857.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32:** 1792-1797.

Eisen J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis *Genome Res*. **8:** 163-167.

Fast N.M., Kissinger J. C., Roos D. S. and Keeling P. J. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol.* **18:** 418-426.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science* **269:** 496-512.

Frickey T. and Lupas A.N. 2004. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res*. **32:** 5231-5238.

Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., Galibert F., Hoheisel J.D., Jacq C., Johnston M., Louis E.J., Mewes H.W., Murakami Y., Philippsen P., Tettelin H. and Oliver S.G. 1996. Life with 6000 genes. *Science* **274:** 546, 563-567.

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

Hackett J.D., Maranda L., Yoon H.S. and Bhattacharya D. 2003. Phylogenetic evidence for the cryptophyte origin of the plastid of *Dinophysis* (Dinophysiales, Dinophyceae). *J. Phycol*. **39:** 440-448.

Harper J.T., Waanders E., and Keeling P.J. 2005. On the monophyly of chromalveolates using a six protein phylogeny of eukaryotes. *Int. J. Syst. Evol. Microbiol*. **55:** 487-496.

Huang J., Mullapudi N., Lancto C.A., Scott M., Abrahamsen M.S. and Kissinger J.C. 2004. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum. Genome Biol*. **5:** R88.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J.C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R.H., Wilson R.K., Hillier L.W., McPherson J.D., Marra M.A., Mardis E.R., Fulton L.A., Chinwalla A.T., Pepin K.H., Gish W.R., Chissoe S.L., Wendl M.C., Delehaunty K.D., Miner T.L., Delehaunty A., Kramer J.B., Cook L.L., Fulton R.S., Johnson D.L., Minx P.J., Clifton S.W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J.F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R.A., Muzny D.M., Scherer S.E., Bouck J.B., Sodergren E.J., Worley K.C., Rives C.M., Gorrell J.H., Metzker M.L., Naylor S.L., Kucherlapati R.S., Nelson D.L., Weinstock G.M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Smith D.R., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee H.M., Dubois J., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis R.W., Federspiel N.A., Abola A.P., Proctor M.J., Myers R.M., Schmutz J., Dickson M., Grimwood J., Cox D.R., Olson M.V., Kaul R., Raymond C., Shimizu N., Kawasaki K., Minoshima S., Evans G.A., Athanasiou M., Schultz R., Roe B.A., Chen F., Pan H., Ramser J., Lehrach H., Reinhardt R., McCombie W.R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J.A., Bateman A., Batzoglou S., Birney E., Bork P., Brown D.G., Burge C.B., Cerutti L., Chen H.C., Church D., Clamp M., Copley R.R., Doerks T., Eddy S.R., Eichler E.E., Furey T.S., Galagan J., Gilbert J.G., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson L.S., Jones T.A., Kasif S., Kaspryzk A., Kennedy S., Kent W.J., Kitts P., Koonin E.V., Korf I., Kulp D., Lancet D., Lowe T.M., McLysaght A., Mikkelsen T., Moran J.V., Mulder N., Pollara V.J., Ponting C.P., Schuler G., Schultz J., Slater G., Smit A.F., Stupka E., Szustakowski J., Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y.I., Wolfe K.H., Yang S.P., Yeh R.F., Collins F., Guyer M.S., Peterson J., Felsenfeld A., Wetterstrand K.A., Patrinos A., Morgan M.J., de Jong P., Catanese J.J., Osoegawa K., Shizuya H., Choi S., Chen Y.J.and International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

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

Lidie K.B., Ryan J.C., Barbier M. and Van Dolah F.M. 2005. Gene expression in Florida red tide dinoflagellate *Karenia*

brevis: analysis of an expressed sequence tag library and development of DNA microarray. *Mar. Biotechnol.* **7:** 481-493.

Martin W., Rujan T., Richly E., Hansen A., Cornelsen S., Lins T., Leister D., Stoebe B., Hasegawa M. and Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA* **99:** 12246-12251.

Matsuzaki M., Misumi O., Shin-I T., Maruyama S., Takahara M., Miyagishima S.Y., Mori T., Nishida K., Yagisawa F., Nishida K., Yoshida Y., Nishimura Y., Nakao S., Kobayashi T., Momoyama Y., Higashiyama T., Minoda A., Sano M., Nomoto H., Oishi K., Hayashi H., Ohta F., Nishizaka S., Haga S., Miura S., Morishita T., Kabeya Y., Terasawa K., Suzuki Y., Ishii Y., Asakawa S., Takano H., Ohta N., Kuroiwa H., Tanaka K., Shimizu N., Sugano S., Sato N., Nozaki H., Ogasawara N., Kohara Y. and Kuroiwa T. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428:** 653-657.

Misumi O., Matsuzaki M., Nozaki H., Miyagishima S.Y., Mori T., Nishida K., Yagisawa F., Yoshida Y., Kuroiwa H. and Kuroiwa T. 2005. Cyanidioschyzon merolae genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. *Plant Physiol.* **137:** 567-585.

Montsant A., Jabbari K., Maheswari U. and Bowler C. 2005. Comparative genomics of the pennate diatom *Phaeodactylum tricornutum*. *Plant Physiol.* **137:** 500-513.

Notredame C., Higgins D. and Heringa J. 2000.T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* **302:** 205-217.

Pearson W.R. and Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85:** 2444-2448.

Rodriguez-Ezpeleta N., Brinkmann H., Burey S.C., Roure B., Burger G., Loffelhardt W., Bohnert H.J., Philippe H. and Lang B.F. 2005 Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* **15:** 1325-1330.

Schmidt H.A., Strimmer K., Vingron M. and Haeseler A.V. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502-504.

Sicheritz-Ponten T. and Anderson S.G. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29:** 545-552.

Sjolander K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* **20:** 170-9.

Swofford, D. L. 2001. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., Levine A.J., Roberts R.J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A.E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T.J., Higgins M.E., Ji R.R., Ke Z., Ketchum K.A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G.V., Milshina N., Moore H.M., Naik A.K., Narayan V.A., Neelam B., Nusskern D., Rusch D.B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., Yao A., Ye J., Zhan M., Zhang W., Zhang H., Zhao Q., Zheng L., Zhong F., Zhong W., Zhu S., Zhao S., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng M.L., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers Y.H., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E., Thomas R., Tint N.N., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril J.F., Guigo R., Campbell M.J., Sjolander K.V., Karlak B., Kejariwal A., Mi H., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang Y.H., Coyne M., Dahlke C., Mays A., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M., Wu D., Wu M., Xia A., Zandieh A. and Zhu X. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Watanabe M.M., Takeda Y., Sasa T., Inouye I., Suda S., Sawaguchi T. and Chihara M. 1987. A green dinoflagellate with chlorophylls a and b morphology fine structure of the chloroplast and chlorophyll composition. *J. Phycol.* **23:** 382-389.

Yoon H.S., Hackett J.D., Van Dolah F.M., Nosenko T., Lidie K.L. and Bhattacharya D. 2005. Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Mol. Biol. Evol.* **22:** 1299-1308.

Yoon H.S., Hackett J.D., Pinto G. and Bhattacharya D. 2002. The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci. USA* **99:** 15507-15512.

Zmasek C.M. and Eddy S.R. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **16:** 3:14.

Received 6 February 2006
Accepted 26 February 2006