

## Retrotransposons and Tandem Repeat Sequences in the Nuclear Genomes of Cryptomonad Algae

Hameed Khan,<sup>1</sup> Catherine Kozera,<sup>2</sup> Bruce A. Curtis,<sup>2</sup> Jillian Tarrant Bussey,<sup>2</sup> Stan Theophilou,<sup>1,\*</sup> Sharen Bowman,<sup>2</sup> John M. Archibald<sup>1</sup>

<sup>1</sup> Genome Atlantic and the Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 1X5

<sup>2</sup> The Atlantic Genome Centre, Institute for Marine Biosciences, Halifax, Nova Scotia, Canada

Received: 10 April 2006 / Accepted: 24 October 2006 [Reviewing Editor: Dr. Debashish Bhattacharya]

**Abstract.** The cryptomonads are an enigmatic group of unicellular eukaryotic algae that possess two nuclear genomes, having acquired photosynthesis by the uptake and retention of a eukaryotic algal endosymbiont. The endosymbiont nuclear genome, or nucleomorph, of the cryptomonad *Guillardia theta* has been completely sequenced: at only 551 kilobases (kb) and with a gene density of  $\sim 1$  gene/kb, it is a model of compaction. In contrast, very little is known about the structure and composition of the cryptomonad host nuclear genome. Here we present the results of two small-scale sequencing surveys of fosmid clone libraries from two distantly related cryptomonads, *Rhodomonas salina* CCMP1319 and *Cryptomonas paramecium* CCAP977/2A, corresponding to  $\sim 150$  and  $\sim 235$  kb of sequence, respectively. Very few of the random end sequences determined in this study show similarity to known genes in other eukaryotes, underscoring the considerable evolutionary distance between the cryptomonads and other eukaryotes whose nuclear genomes have been completely sequenced. Using a combination of fosmid clone end-sequencing, Southern hybridizations, and PCR, we demonstrate that *Ty3-gypsy* long-terminal repeat (LTR) retrotransposons and tandem repeat sequences are a prominent feature of the nuclear genomes of both organisms. The

complete sequence of a 30.9-kb genomic fragment from *R. salina* was found to contain a full-length *Ty3-gypsy* element with near-identical LTRs and a chromodomain, a protein module suggested to mediate the site-specific integration of the retrotransposon. The discovery of chromodomain-containing retroelements in cryptomonads further expands the known distribution of the so-called chromoviruses across the tree of eukaryotes.

**Key words:** Retrotransposons — Tandem repeat sequences — Nuclear genomes — Cryptomonad algae

### Introduction

Genome size in eukaryotic cells varies tremendously (Gregory 2001). Much of this variation can be attributed to differences in the abundance of repetitive DNA sequences such as transposable elements and tandem repeats, which are known to make up a significant fraction of the nuclear genomes of many plant and animal species (reviewed in Gregory 2005; Hancock 2002; Kidwell 2002). Approximately 40% of the human genome, for example, is comprised of various classes of repetitive elements (Lander et al. 2001), and as much as 90% of the genome of some plants is comprised of repetitive DNA (Flavell 1986).

Considerably less is known about the abundance, diversity, and biological significance of repetitive

\*Present address: DNA Technologies, 1721 Lower Water Street, Halifax, Nova Scotia, Canada, B3J 1S5

Correspondence to: H. Khan; email: kxanh@dal.ca

elements in the genomes of unicellular eukaryotes (protists), although preliminary investigations suggest that repeat sequence abundance correlates positively with genome size (reviewed by Wickstead et al. 2003). For example, approximately 10% of the ~60-megabase (Mb) genome of *Trypanosoma cruzi* is comprised of a 195-bp repeat sequence (Elias et al. 2003; Requena et al. 1996; Sloof et al. 1983), while the 8.3-Mb genome of the apicomplexan *Theileria parva* is largely devoid of repetitive DNA (Gardner et al. 2005; Nene et al. 1998). The recent explosion in the number of eukaryotic genome sequencing projects promises to further improve our understanding of the structure and composition of protist genomes. Multiple partial or complete genomes are now available from within the trypanosomatids (El-Sayed et al. 2005a, b), apicomplexans (Abrahamsen et al. 2004; Gardner et al. 2002, 2005; Xu et al. 2004), ciliates (Stover et al. 2006; Zagulski et al. 2004), and Amoebozoa (Loftus et al. 2005). The Genomes On Line Database provides a comprehensive list of finished and ongoing projects (Liolios et al. 2006). For practical reasons, most efforts have focused on sequencing the relatively small and compact genomes of parasitic species. Yet the genomes of eukaryotic microorganisms span an exceptionally wide range of sizes (McGrath and Katz 2004), and a large fraction of protist diversity remains unexplored from a genomic perspective.

We are studying the genomic diversity of a eukaryotic lineage of pivotal evolutionary significance, the cryptomonads. These organisms are unicellular algae that acquired their photosynthetic capabilities through the process of secondary endosymbiosis. This occurs when a eukaryotic phototroph establishes itself as a permanent resident inside a nonphotosynthetic host eukaryote (reviewed by Archibald 2005; Archibald and Keeling 2002; Delwiche 1999; Keeling 2004; Palmer 2003). Secondary endosymbiosis has given rise to a large and exceptionally diverse array of photosynthetic organisms. In addition to the cryptomonads, these include the dinoflagellate algae, haptophytes, heterokonts (e.g., diatoms and kelps), the euglenids, and the chlorarachniophytes (Delwiche 1999; Keeling 2004; Palmer 2003). The number of secondary endosymbioses that have occurred during eukaryotic evolution is controversial, but it is generally thought to have occurred at least twice and has involved both red and green algal endosymbionts (Bhattacharya et al. 2003; Bodyl 2005; Palmer 2003).

Together with the chlorarachniophytes, the cryptomonads are unusual in that they still retain the nucleus of their eukaryotic endosymbiont in a highly reduced form termed a “nucleomorph.” Genomic analyses have revealed that these tiny organelles harbor the smallest nuclear genomes known (reviewed by Gilson 2001; Gilson and McFadden 2002).

The nucleomorph genome of the model cryptomonad *Guillardia theta* has been completely sequenced and is only 551 kilobases (kb) in size (Douglas et al. 2001). The nucleomorph genome of the chlorarachniophyte *Bigeloviella natans* is even smaller, at 373 kb (Gilson et al. 2006). Both genomes are partitioned among three small chromosomes and are extremely compact, with approximately one gene per kilobase, tiny spliceosomal introns, and little in the way of repetitive sequences (Douglas et al. 2001; Gilson et al. 2006). The broad similarities of the cryptomonad and chlorarachniophyte nucleomorph genomes are especially intriguing when one considers that they are derived from independent secondary endosymbiotic events. In the case of the chlorarachniophytes, the endosymbiont is most likely derived from an ancestor of modern-day green algae (Archibald et al. 2003; Ishida et al. 1997, 1999; McFadden et al. 1995), while the cryptomonad endosymbiont is related to red algae (Archibald et al. 2001; Douglas et al. 1991; Douglas and Penny 1999; Van der Auwera et al. 1998). In both lineages, the bulk of the genes present in the original endosymbiont nuclear genome have been lost or transferred to the nuclear genome of the host cell (Gilson 2001; Gilson and McFadden 2002).

Apart from being a repository for large numbers of nucleomorph-derived genes (Archibald et al. 2003; Deane et al. 2000), very little is known about the size and structure of the host nuclear genome of cryptomonads and chlorarachniophytes. Cryptomonads are thought to belong to a very diverse “supergroup” of eukaryotes called Chromalveolates (Cavalier-Smith 1999; Keeling et al. 2005), a lineage that includes other photosynthetic organisms such as the haptophytes, heterokonts (e.g., diatoms), and dinoflagellates, as well as nonphotosynthetic members such as the apicomplexans and ciliates. Here we present the results of small-scale sequencing surveys from two distantly related cryptomonad algae, *Rhodomonas salina* and *Cryptomonas paramecium*. We are interested in both the host nucleus and the nucleomorph of these species because at ~755 and ~450 kb, respectively, their nucleomorph genomes differ significantly from the 551-kb nucleomorph genome of the model cryptomonad *Guillardia theta* (Douglas et al. 2001; Lane et al. 2006; Rensing et al. 1994). This variation may reflect differences in the amount of nucleomorph-to-host-nucleus gene transfer that has occurred since these organisms diverged from a common ancestor (Lane et al. 2006). Our results indicate that the nuclear genomes of both species contain a variety of repetitive DNA sequences, including tandem repeats and significant numbers of long terminal repeat (LTR) retrotransposons belonging to the *Ty3-gypsy* family. Phylogenetic analyses suggest that these elements are most similar to the “chromoviruses” of plants, animals, and fungi.

## Materials and Methods

### Cell Culture and DNA Extraction

Cryptomonad cultures were obtained from public culture collections and grown in the laboratory at 25°C. *Rhodomonas salina* strain CCMP1319 was grown under a 14:10-h diurnal growth cycle in f/2-Si medium made with artificial seawater, while *Cryptomonas paramecium* strain CCAP977/2A, a nonphotosynthetic species, was grown in "Chilomonas" medium (1 g CH<sub>3</sub>COONa·3H<sub>2</sub>O, 1 g Lab Lemco powder [Oxoid; Hampshire, England], 1 liter H<sub>2</sub>O). Large-scale cultures were grown in 2-L flasks to late logarithmic phase and cells were harvested by centrifugation. Cell pellets were resuspended in a Tris-HCl digestion buffer (200 mM Tris-HCl [pH 7.5], 250 mM NaCl, 25 mM EDTA, 0.5% SDS) and incubated at 50°C for 10 min. After incubation, samples were centrifuged for 5 min at 15,000g and the aqueous phase was subjected to two rounds of protein extraction with phenol and chloroform. DNA was precipitated from the aqueous layer using 100% ethanol, centrifuged, washed with 70% ethanol, resuspended in 50 µl of dH<sub>2</sub>O, and stored at -20°C.

### Fosmid Clone Library Construction, Subcloning, and DNA Sequencing

Fosmid libraries with 35- to 45-kb inserts were generated using the CopyControl fosmid library production kit (Epicentre, Madison, WI). High molecular weight total cellular DNA from *Rhodomonas salina* and *Cryptomonas paramecium* was sheared by repeated passage through a p200 pipette tip, then end repaired with T4 DNA polymerase and T4 polynucleotide kinase, and fragments approximately 40 kb in size were isolated from 1% low-melting point (LMP) agarose gels by electrophoresis. Fragments were ligated into linearized, dephosphorylated pCC1FOS vector, packaged, and transformed into EPI300-T1 *Escherichia coli* cells. The resulting colonies were picked into 384-well plates containing LB and glycerol (25%), incubated overnight at 37°C, and stored at -80°C.

For fosmid end sequencing, overnight cultures of fosmid clones were inoculated into fresh LB + chloramphenicol (12.5 µg/ml). CopyControl induction solution was added, followed by vigorous shaking at 37°C for 5 h. Fosmid DNA was then isolated using the Perfectprep BAC 96 kit (Eppendorf, Hamburg, Germany). DNA was quality checked by digestion with *NotI* restriction enzyme followed by agarose gel electrophoresis. Purified fosmid DNA was end sequenced using ET terminator chemistry (GE Healthcare). Sequencing reactions were processed using Sera-Mag magnetic carboxylate-modified microparticles (Seradyn, Indianapolis, IN) to remove excess fluorescent terminators before loading onto GE Healthcare MegaBace capillary DNA sequencers.

Fosmid clones of particular interest were subcloned in preparation for complete sequencing as follows. Purified fosmid DNA was nebulized and blunt ended, and fragments of 1–2 kb were size-selected on 0.8% LMP agarose gels. Isolated fragments were cloned into the *SmaI* site of dephosphorylated pUC19 vector. After transformation into One Shot TOP10 Electrocompetent *E. coli* (Invitrogen), individual bacterial colonies were picked into 96-well plates containing LB/glycerol. These plates were incubated overnight at 37°C. The resulting bacterial suspensions were inoculated into lysis buffer and denatured at 95°C for 5 min. DNA from each clone was amplified using TempliPhi DNA polymerase (GE Healthcare) according to manufacturer's instructions. DNA sequencing was performed using ET terminator chemistry and MegaBace capillary sequencers as described above. The fosmid clone end-sequences determined in this study have been deposited in the GSS (genome survey sequences) division of GenBank under the following accession numbers: DX586900–DX587205 (*Rhodo-*

*monas salina*) and DX587206–DX587676 (*Cryptomonas paramecium*). A 30,933-kb contiguous sequence of *R. salina* fosmid clone rs04C02 has been deposited in GenBank under accession number DQ859722.

### High-Density Filter Construction

High-density filters containing randomly selected fosmid clones were constructed as follows. Glycerol stocks of fosmid clones were first allowed to thaw. Using an automatic colony-arraying robot (QPix2, Genetix), bacterial suspensions were spotted onto nylon filters (Hybond N+, Amersham) that were placed in contact with LB-agar containing chloramphenicol (12.5 µg/ml). A total of 1920 clones were arrayed on each filter (5 × 384 well plates). Bacterial colonies were grown on one surface of the filter by incubation for 18 h at 37°C. The bacterial colonies arrayed on the filter were then lysed in situ on reagent-soaked Whatman paper using a standard alkaline lysis procedure (<http://www.sanger.ac.uk/HGP/methods/mapping/grids/lysis.shtml>). Filters were extensively rinsed (2 × SSC/0.1% SDS for 5 min, 2 × SSC for 5 min, two washes in 50 mM Tris-Cl pH 7.4), dried, and UV cross-linked in preparation for Southern hybridization.

### PCR Amplification, Cloning, and Sequencing

PCR primers specific for *Ty3-gypsy* LTR retrotransposon sequences in *Rhodomonas salina* and *Cryptomonas paramecium* were designed based on data obtained from fosmid clones and on amino acid alignments of retrotransposon proteins from a wide range of eukaryotic species. An ~400-bp DNA fragment from *R. salina* coding for a portion of the reverse transcriptase domain of the LTR retrotransposon was amplified using the following primers: Rsal.gagpol.F1, ACACCACGAGTGGCTGGTGA; and Rsal.gagpol.R1, TTCTGGTAGAACCCAGCCAGTCC. A smaller (~250-bp) fragment corresponding to the same region from *C. paramecium* was amplified using the following primers: Cpar.RT.F1, CAAGGCCGACGGCACCTGGCGCTTC; and Cpar.RT.R1, CAGGAGYGACGACATGCCGTGCAGGCCGAA. These primers were also used to generate labeled PCR products for Southern hybridization (below) using the digoxigenin (DIG) synthesis kit (Roche Diagnostics Corp., Indianapolis, IN). PCR products were purified using the MinElute Gel Extraction Kit (Qiagen Sciences, Valencia, CA) and cloned using the Topo TA Cloning Kit (Invitrogen) according to the manufacturer's protocol. Individual clones were selected and grown overnight in 2 ml of LB medium. Plasmids were extracted using the Fastplasmid Mini Kit (Eppendorf) and insert sizes were determined by restriction enzyme digestion.

Cloned DNA fragments generated by PCR were sequenced using the CEQ Dye Terminator Cycle Sequencing (DCTS) kit (Beckman Coulter, Inc., Fullerton, CA) and run on Beckman Coulter CEQ8000 capillary DNA sequencers. The sequences of PCR products have been deposited in GenBank under accession numbers DQ641209–DQ641220 (*R. salina*) and DQ641221–DQ641235 (*C. paramecium*).

### Genomic Digests and Southern Hybridizations

Total genomic DNAs (~2 µg) from *Cryptomonas paramecium* and *Rhodomonas salina* were digested using various combinations of restriction enzymes, including *PstI*, *EcoRI*, *HindIII*, *BamHI*, *BglII*, and *MboI* (Invitrogen). Digests contained ~10 U of enzyme and were incubated overnight at 37°C. The digested samples were resolved on 1% agarose gels for 16 h at 20 V. Gels were blotted onto positively charged nylon membranes (Roche Diagnostics Corp.)

using the method described in Current Protocols in Molecular Biology (Wiley Interscience, New York). Southern hybridizations with DIG-labeled probes were performed overnight at 60°C and membranes were processed using the DIG Luminescence Detection Kit and CDP-Star substrate (Roche Diagnostics Corp.).

### Data Analysis

Fosmid clone end-sequences obtained from the *R. salina* and *C. paramecium* libraries were compared to sequences in GenBank using BLASTX and BLASTN (Altschul et al. 1997). Sequences with no obvious similarity to known genes were searched for the presence of repetitive elements using the program Tandem Repeats Finder (<http://www.tandem.bu.edu/trf/trf.html>; Benson 1999). Contiguous sequences were assembled using the Staden package (Dear and Staden 1991) and Sequencher version 4.5 (GeneCodes Corp., Ann Arbor, MI). Artemis (The Sanger Institute) was used to construct G + C content profiles using a 100-bp sliding window. Phylogenetic analysis of *Ty-3 gypsy* retrotransposons was performed on the reverse transcriptase (RT) domain of the pol protein. Amino acid sequences inferred from the new *R. salina* and *C. paramecium Ty-3-gypsy* coding sequences were added to a comprehensive RT domain alignment provided by I. Marín (Marín and Lloréns 2000) and aligned by eye. Additional sequences were retrieved from GenBank and the Joint Genome Institute web site (<http://www.jgi.doe.gov/>). Phylogenetic analyses were performed on alignments of various lengths, depending on the presence/absence of partial sequences (e.g., the PCR-generated fragments from *C. paramecium*). The final alignment (available upon request) contained 56 sequences and 111 unambiguously aligned amino acid positions. TREE-PUZZLE 5.2 (Strimmer and von Haeseler 1996) was used to construct  $\Gamma$ -corrected distance matrices using the WAG substitution matrix and a four-rate category discrete approximation to the  $\Gamma$  distribution plus an invariable sites category estimated from the data. Phylogenies were inferred from these matrices using FITCH (<http://www.evolution.genetics.washington.edu/phylip.html>) and BIONJ (Gascuel 1997). Maximum likelihood trees were inferred using PHYML version 2.3 (Guindon and Gascuel 2003) with the WAG substitution matrix and four  $\Gamma$ -distributed rate categories. Statistical support for the resulting topologies was determined by bootstrapping with 100 replicates.

## Results

### Fosmid End-Sequencing Survey

Despite their evolutionary significance, the cryptomonads are a very poorly understood eukaryotic lineage. The plastid and nucleomorph genomes of the model cryptomonad *Guillardia theta* have been completely sequenced (Douglas and Penny 1999; Douglas et al. 2001), as has the mitochondrial genome of *Rhodomonas salina* (Hauth et al. 2005), but very few nucleus-encoded genes have been isolated. Those that have been studied are limited to ribosomal RNA (rRNA) genes (e.g., Hoef-Emden 2005; Hoef-Emden et al. 2002) and a handful of conserved protein genes isolated by PCR for the purposes of phylogenetic reconstruction (e.g., Harper et al. 2005).

In order to obtain a preliminary glimpse into the structure and content of the cryptomonad nuclear genome, and at the same time identify cloned frag-

ments containing nucleomorph and plastid DNA, we performed a small-scale sequence survey of large-insert fosmid libraries constructed from total DNA isolated from two distantly related cryptomonad species, *Rhodomonas salina* and *Cryptomonas paramecium*. Five hundred seventy-six (*R. salina*) and 1056 (*C. paramecium*) clones were chosen at random from each library and sequenced from one or both ends. Three hundred five of the *R. salina* clones and 471 clones from *C. paramecium* produced quality sequence data >200 bp in length. With an average read length of ~500 bp, these surveys correspond to ~150 and ~235 kb of genomic sequence data from *R. salina* and *C. paramecium*, respectively. Quality sequences were compared to GenBank using BLASTN and BLASTX searches (Altschul et al. 1997), and the results are summarized in Tables 1 and 2. Very few of the end sequences produced significant matches to known or predicted genes in the public databases. In *R. salina*, six of the randomly selected fosmids appear to contain DNA derived from the plastid genome (Table 1), as the end-sequences of these clones showed strong similarity to genes encoded in the plastid genome of the cryptomonad *Guillardia theta* (Douglas and Penny 1999). The relatively low G + C content of these sequences (32%–50%) is also consistent with a plastid origin. In contrast, none of the end-sequenced fosmids from *C. paramecium* appear to contain plastid DNA, and none of the *R. salina* or *C. paramecium* fosmid clones harbor DNA clearly derived from the nucleomorph genome. The top hit in the *C. paramecium* survey was from a clone (Cpar\_0002b06) containing a full-length histone H2A gene, while for *R. salina* the top nonplastid match was to a mastigoneme-like protein in *Entamoeba histolytica*. None of the *R. salina* genes possessed obvious spliceosomal introns, although two genes in *C. paramecium* (P-type ATPase and U6 snRNP) contained putative introns <100 bp in length with canonical GT-AG boundaries (data not shown).

### Tandem Repeat Sequences

A significant fraction of the fosmid end sequences obtained from the *Rhodomonas salina* and *Cryptomonas paramecium* library clones showed no similarity to known genes. Upon visual inspection, some of these sequences appeared to be of low complexity and showed evidence of repetitive structure. We therefore used the program Tandem Repeats Finder (Benson 1999) to systematically search for tandem repeats in all sequences that failed to produce a significant BLASTN or BLASTX hit. In sum, 2.3% and 4.7% of the *R. salina* and *C. paramecium* sequences, respectively, possessed tandemly repeated sequences (Tables 1 and 2). The repeats spanned a large range

**Table 1.** Fosmid end-sequencing survey: *Rhodomonas salina* CCMP1319

Clone ID	Accession	Putative ID (protein / gene, top hit organism)	e-value <sup>a</sup>	% G + C
<b>Plastid</b>				
Rsal_0005g11	DX587154	Photosystem I P700 chlorophyll a apoprotein A2/psaB; <i>Guillardia theta</i> (plastid)	1e-96	42
Rsal_0002a01	DX587010	RNA polymerase b" subunit/rpoC2; <i>Guillardia theta</i> (plastid)	1e-108	38
Rsal_0002a11	DX587015	Magnesium chelatase/chII (end sequence also contains psaM); <i>Guillardia theta</i> (plastid)	7e-76	33
Rsal_0002c01	DX587024	Clp protease/clpC; <i>Guillardia theta</i> (plastid)	8e-111	38
Rsal_0006b07	DX587169	Cytochrome b559 $\alpha$ subunit/psbE; <i>Guillardia theta</i> (plastid)	2e-42	33
Rsal_0005b03(f)	DX587133	16S rRNA gene; <i>Guillardia theta</i> (plastid)	0.0 (BLASTN)	50
Rsal_0005b03(r)	DX587134	Ribosomal protein L13/rpL13 (end sequence also contains rpoA); <i>Guillardia theta</i> (plastid)	2e-39	32
<b>Miscellaneous</b>				
Rsal_0003d09	DX587090	Mastigoneme-like protein; <i>Entamoeba histolytica</i>	2e-10	63
Rsal_0003c04	DX587079	Chromatin condensation factor; <i>Triticum aestivum</i>	4e-07	63
Rsal_0003g12	DX587110	Multidrug resistance-associated protein/ABC transporter; <i>Triticum aestivum</i>	6e-09	60
Rsal_0003g04	DX587105	Protein kinase; <i>Rattus norvegicus</i>	1e-05	54
<b>Retrotransposons</b>				
Rsal_0004a01	DX587116	Polyprotein; <i>Glycine max</i>	5e-65	42
Rsal_0006d06	DX587181	Retrotransposon polyprotein; <i>Ipomoea batatas</i>	3e-55	45
Rsal_0002b07	DX587020	Putative retroelement; <i>Oryza sativa</i>	2e-48	53
Rsal_0002b08	DX587021	Putative retroelement; <i>Oryza sativa</i>	5e-42	52
Rsal_0001c05(r)	DX586935	Retrotransposon protein; <i>Oryza sativa</i>	9e-44	47
Rsal_0006g12	DX587198	Ty3-gypsy putative retrotransposon protein; <i>Oryza sativa</i>	4e-33	36
Rsal_0001b07	DX586921	Retrotransposon, Ty3-gypsy subclass; <i>Oryza sativa</i>	4e-33	52
Rsal_0002f10	DX587051	Putative retroelement integrase; <i>Arabidopsis thaliana</i>	5e-18	45
Rsal_0002d02	DX587032	Putative gag-pol polyprotein; <i>Oryza sativa</i>	1e-16	59
Rsal_0003e03	DX587094	Copia-type polyprotein; <i>Arabidopsis thaliana</i>	2e-11	56
Rsal_0003c06	DX587080	Putative retroelement pol polyprotein; <i>Oryza sativa</i>	4e-12	56
Rsal_0002e07	DX587046	Retrotransposon Tca5 polyprotein; <i>Candida albicans</i>	1e-13	51
Rsal_0001f07(f)	DX586972	Gag-pol polyprotein; <i>Aspergillus flavus</i>	3e-40	43
Rsal_0001f06	DX586970	Copia-type polyprotein; <i>Arabidopsis thaliana</i>	6e-09	51
Rsal_0005d04	DX587143	Polyprotein; <i>Aspergillus flavus</i>	1e-05	50
Rsal_0005d02	DX587144	Pol polyprotein; <i>Anopheles gambiae</i>	5e-05	55
Rsal_0006b01	DX587164	Pol like protein; <i>Danio rerio</i>	7e-05	60
Rsal_0003f06	DX587098	Putative retrotransposon RIRE1 polyprotein; <i>Zea mays</i>	1e-04	50
Rsal_0006b04	DX587167	Pol protein; <i>Phanerochaete chrysosporium</i>	1e-04	45
Rsal_0001a03	DX586903	Putative endonuclease/reverse transcriptase; <i>Lymntria dispar</i>	5e-04	60
Rsal_0003h02	DX587112	Putative gag-pol polyprotein; <i>Zea mays</i>	5e-06	57
Rsal_0001a02	DX586901	Copia-type polyprotein; <i>Arabidopsis thaliana</i>	7e-09	53
Rsal_0004b06	DX587118	Retrotransposon long terminal repeat <sup>b</sup>	n/a	55
Rsal_0005f04	DX587148	Retrotransposon long terminal repeat <sup>b</sup>	n/a	47
Rsal_0005h09	DX587157	Retrotransposon long terminal repeat <sup>b</sup>	n/a	49
Rsal_0005h10	DX587158	Retrotransposon long terminal repeat <sup>b</sup>	n/a	54
Rsal_0006g08	DX587197	Retrotransposon long terminal repeat <sup>b</sup>	n/a	47
Rsal_0002c05	DX587027	Retrotransposon long terminal repeat <sup>b</sup>	n/a	49
<b>Repetitive sequences</b>				
Rsal_0005h07	DX587155	84-bp repeat (copy number = 4.7, % matches = 89, % indels = 7)	n/a	54
Rsal_0005c03	DX587141	18-bp repeat (copy number = 2.3, % matches = 95, % indels = 0)	n/a	60
Rsal_0002d05	DX587036	15-bp repeat (copy number = 4.1, % matches = 76, % indels = 2)	n/a	56
		8-bp repeat (copy number = 22.9, % matches = 91, % indels = 4)		
		3-bp repeat (copy number = 42.3, % matches = 72, % indels = 7)		
Rsal_0002b09	DX587022	14-bp repeat (copy number = 4.9, % matches = 75, % indels = 13)	n/a	57
Rsal_0003h06	DX587113	12-bp repeat (copy number = 3.1, % matches = 79, % indels = 13)	n/a	52
		2-bp repeat (copy number = 41.5, % matches = 90, % indels = 0)		
Rsal_0006g01	DX587193	12-bp repeat (copy number = 6.3, % matches = 77, % indels = 0)	n/a	51
		6-bp repeat (copy number = 57.3, % matches = 74, % indels = 13)		
Rsal_0002g07	DX587057	10-bp repeat (copy number = 5.1, % matches = 100, % indels = 0)	n/a	60
		6-bp repeat (copy number = 7.8, % matches = 100, % indels = 0)		

<sup>a</sup> BLASTX unless otherwise indicated.<sup>b</sup> Confirmed by fosmid subcloning and sequencing.

**Table 2.** Fosmid end-sequencing survey: *Cryptomonas paramecium* CCAP977/2A

Clone ID	Accession	Putative ID (protein / gene, top hit organism)	e-value <sup>a</sup>	% G + C
<b>Miscellaneous</b>				
Cpar_0005h02	DX587399	WD40 repeat protein; <i>Nostoc punctiforme</i> ; repeat structure detected as 126-bp repeat element (copy number = 1.9, % matches = 80)	2e-42	58
Cpar_0002b06	DX587278	Histone H2A; <i>Volvox carteri</i>	4e-52	58
Cpar_0008g11	DX587538	Malonyl-CoA:ACP transacylase / fabD; <i>Guillardia theta</i>	9e-32	52
Cpar_0010d07	DX587614	P-type calcium ATPase; <i>Magnaporthe grisea</i> (coding sequence interrupted by 3 potential introns)	7e-20	50
Cpar_0006d10	DX587435	Glutamine-rich hypothetical protein; <i>Xenopus laevis</i>	7e-26	64
Cpar_0006d02	DX587429	Multidrug resistance protein; <i>Macaca mulatta</i>	2e-06	61
Cpar_0001b04	DX587213	Cystine-rich hypothetical protein; <i>Giardia lamblia</i>	3e-12	58
Cpar_0001d01	DX587227	Hypothetical protein; <i>Dictyostelium discoideum</i>	6e-08	50
Cpar_0009f12	DX587574	U6-associated snRNP; <i>Mus musculus</i> (coding sequence interrupted by potential intron)	6e-11	43
Cpar_0006c12	DX587427	Protein kinase; <i>Oryza sativa</i>	2e-10	66
Cpar_0008f09	DX587529	Dynein-like protein; <i>Pan troglodytes</i>	2e-06	65
Cpar_0006h10	DX587464	Proline-rich hypothetical protein; <i>Homo sapiens</i>	4e-07	64
Cpar_0008a10	DX587502	Transcription initiation factor IID/tfIID; <i>Encephalitozoon cuniculi</i>	2e-06	59
<b>Retrotransposons</b>				
Cpar_0006a10	DX587411	Putative retroelement pol polyprotein; <i>Arabidopsis thaliana</i>	7e-44	72
Cpar_0008h02	DX587540	Putative gag-pol polyprotein; <i>Oryza sativa</i>	8e-21	64
Cpar_0005f04	DX587385	Reverse transcriptase; <i>Ciona intestinalis</i>	4e-16	69
Cpar_0010b10	DX587600	Putative retroelement; <i>Ciona intestinalis</i>	4e-15	51
Cpar_0003h11	DX587333	Pol-like protein; <i>Danio rerio</i>	3e-14	50
Cpar_0002e10	DX587292	Pol-like protein; <i>Ciona intestinalis</i>	5e-12	60
<b>Repetitive sequences</b>				
Cpar_0009d11	DX587564	173-bp repeat (copy number = 3.4, % matches = 90, % indels = 6)	n/a	79
Cpar_0006f05	DX587444	168-bp repeat (copy number = 2.2, % matches = 96, % indels = 1)	n/a	71
Cpar_0004a12	DX587337	165-bp repeat (copy number = 3.5, % matches = 90, % indels = 5)	n/a	58
Cpar_0006a11	DX587412	54-bp repeat (copy number = 3.2, % matches = 93, % indels = 0)	n/a	65
Cpar_0009g02	DX587575	54-bp repeat (copy number = 10.1, % matches = 97, % indels = 1); overlaps with repeat element in Cpar_0010e02 and Cpar_0010a02	n/a	68
Cpar_0010a02	DX587588	54-bp repeat (copy number = 13.1, % matches = 92, % indels = 1); overlaps with repeat element in Cpar_0010e02 and Cpar_0009g02	n/a	71
Cpar_0010e02	DX587619	54-bp repeat (copy number = 9.1, % matches = 96, % indels = 0); overlaps with repeat element in Cpar_0010a02 and Cpar_0009g02	n/a	71
Cpar_0010f08	DX587629	39-bp repeat (copy number = 9.2, % matches = 99, % indels = 0); similarity to transcription elongation regulator 1-like protein; <i>Pan troglodytes</i>	5e-39	59
Cpar_0005b07	DX587366	39-bp repeat (copy number = 5.4, % matches = 90, % indels = 0) 24-bp repeat (copy number = 2.0, % matches = 87, % indels = 0) 18-bp repeat (copy number = 2.7, % matches = 93, % indels = 0)	n/a	64
Cpar_0006b08	DX587420	35-bp repeat (copy number = 2.0, % matches = 100, % indels = 0) 28-bp repeat (copy number = 2.9, % matches = 90, % indels = 0) 18-bp repeat (copy number = 2.0, % matches = 100, % indels = 0) 2-bp repeat (copy number = 14.5, % matches = 100, % indels = 0)	n/a	41
Cpar_0008b04	DX587505	27-bp repeat (copy number = 14.4, % matches = 90, % indels = 2)	n/a	71
Cpar_0006e03	DX587439	21-bp repeat (copy number = 16.2, % matches = 89, % indels = 0) 19-bp repeat (copy number = 2.1, % matches = 85, % indels = 4)	n/a	67
Cpar_0006d01	DX587428	16-bp repeat (copy number = 43.2, % matches = 85, % indels = 4)	n/a	61
Cpar_0007b06	DX587470	12-bp repeat (copy number = 3.9, % matches = 82, % indels = 0)	n/a	21
Cpar_0008f10	DX587530	10-bp repeat (copy number = 4.4, % matches = 94, % indels = 2) 10-bp repeat (copy number = 3.4, % matches = 100, % indels = 0)	n/a	40
Cpar_0001b04	DX587213	4-bp repeat (copy number = 38.3, % matches = 89, % indels = 1)	n/a	42
Cpar_0010f09	DX587630	22-bp repeat (copy number = 1.9, % matches = 100, % indels = 0) 6-bp repeat (copy number = 6.8, % matches = 94, % indels = 0) 8-bp repeat (copy number = 18.1, % matches = 78, % indels = 7) 31-bp repeat (copy number = 4.6, % matches = 75, indels = 9) 23-bp repeat (copy number = 6.0, % matches = 76, % indels = 5) 14-bp repeat (copy number = 6.8, % matches = 75, % indels = 15)	n/a	52
Cpar_0010g12	DX587643	3-bp repeat (copy number = 8.7, % matches = 100, % indels = 0)	n/a	54
Cpar_0010h02	DX587644	55-bp repeat (copy number = 3.2, % matches = 90, % indels = 4) 82-bp repeat (copy number = 2.2, % matches = 88, % indels = 4) 27-bp repeat (copy number = 6.5, % matches = 86, % indels = 6) 55-bp repeat (copy number = 2.8, % matches = 87, % indels = 6)	n/a	70
Cpar_0011a08	DX587652	55-bp repeat (copy number = 2.1, % matches = 88, % indels = 1)	n/a	64

(Continued)

**Table 2.** Continued

Clone ID	Accession	Putative ID (protein / gene, top hit organism)	e-value <sup>a</sup>	% G + C
Cpar_0011c11	DX587655	10-bp repeat (copy number = 4.2, % matches = 79, % indels = 8) 21-bp repeat (copy number = 2.7, % matches = 88, % indels = 2) 51-bp repeat (copy number = 2.0, % matches = 86, % indels = 0)	n/a	43
Cpar_0011e12	DX587662	44-bp repeat (copy number = 5.0, % matches = 97, % indels = 0)	n/a	72

<sup>a</sup> BLASTX search.

of sizes, from di- or tri-nucleotides up to more than 100 bp in length, and were repeated as few as 2 times and up to more than 30 times in both species. As expected for such sequences, many of the repeats were degenerate to some degree, though many showed identities between 95% and 100% across the full length of the array. We identified a particularly striking case of a 54-bp repeat present in end-sequences obtained from four independent *C. paramecium* fosmid clones (Cpar\_0006a11, Cpar\_0009g02, Cpar\_0010a02, and Cpar\_0010e02; Table 2). Three of these sequences contained very similar—although not identical—repeat structures and overlapped with one another to form a single contig. While the G + C contents of the tandem repeat-containing clones in *R. salina* were relatively uniform, they varied significantly in *C. paramecium*, ranging from 21% to 79% (Tables 1 and 2). Given the low G + C content of the nucleomorph genomes of cryptomonads (Douglas et al. 2001; Gilson and McFadden 2002), and organellar genomes in general, it is possible that the 21% G + C clone in *C. paramecium* is in fact derived from the nucleomorph, mitochondrial, or plastid genome.

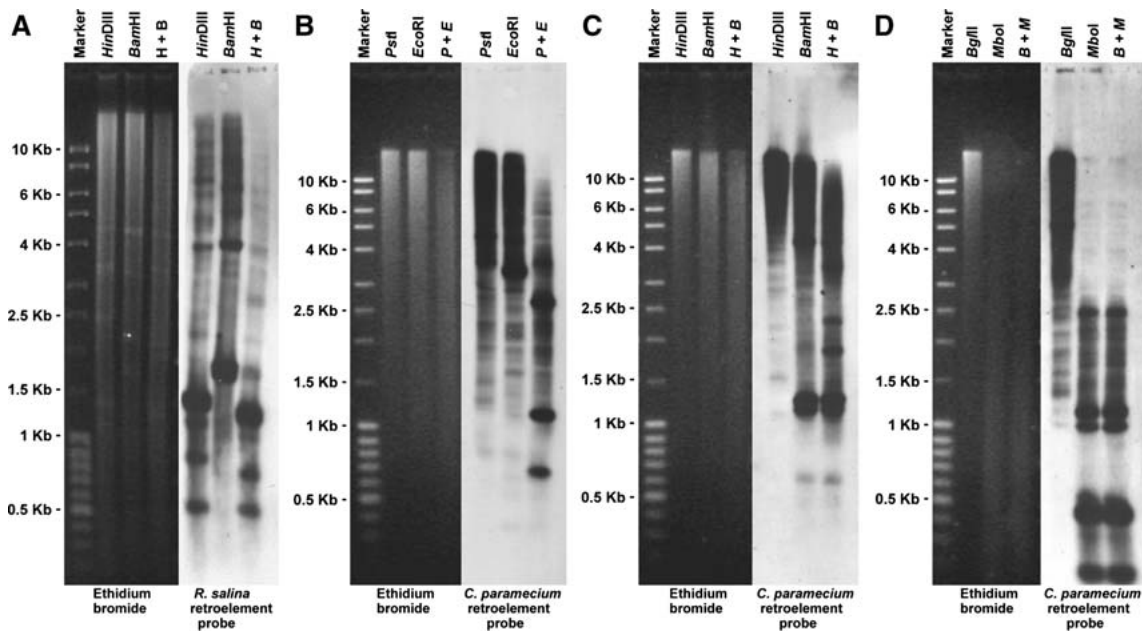
#### *Abundance and Diversity of Ty3-gypsy LTR Retrotransposons*

Nine and seven-tenths percent of successful end-sequences obtained from the *R. salina* fosmid library and 1.3% of the *C. paramecium* sequences produced significant BLASTX matches to retrotransposons (Tables 1 and 2). In both species, the majority of these clones showed similarity to the various domains of the *Ty3-gypsy* class of long terminal repeat (LTR) retrotransposons in plants. Significant hits to *Ty1-copia* and other retroviral elements were also observed in *R. salina*. To further explore the possibility that mobile genetic elements are a prominent feature of the nuclear genomes of these organisms, we performed a variety of Southern hybridization experiments using probes designed to the region of the retrotransposon gene encoding the conserved RT domain. Figure 1 shows the results of hybridizations performed on *R. salina* and *C. paramecium* gDNAs digested with a variety of different restriction enzymes, singly and in combination. Multiple hybrid-

ization bands were apparent in all cases, suggesting that LTR retrotransposon genes are indeed present in multiple copies in the nuclear genomes of both species.

To further assess the abundance of retroelements in the *Rhodomonas salina* nuclear genome, we performed additional hybridizations against randomly selected fosmid clones printed in duplicate on high-density filters. Using the *R. salina* retrotransposon RT domain probe described above, approximately 27% (521 of 1921) of the fosmids produced strong hybridization signals on both filters (Fig. 2), suggesting that at least one RT domain coding region from the *Ty3-gypsy* family is present on the ~40-kb insert in each of the clones. To rule out the possibility that the *R. salina* fosmid clone library used in our sequencing survey and hybridization experiments is artificially enriched for retrotransposon-containing genomic regions, we randomly selected 15 of the 521 positive clones for end-sequencing. None of these sequences showed any similarity to one another. Technical difficulties prevented us from assessing the abundance of retroelements in the *C. paramecium* genome using Southern hybridizations against high-density filters.

We also used PCR to explore the diversity of LTR retrotransposons in the *Rhodomonas salina* and *Cryptomonas paramecium* nuclear genomes. Primers were designed against conserved regions of the RT domain and used to amplify retrotransposon gene fragments from genomic DNA of both organisms. PCR products of the expected size (~400 and ~250 bp for *R. salina* and *C. paramecium*, respectively) were cloned and nine independent clones were sequenced for both species. None of the PCR-generated clones proved to be identical to the retrotransposon genes uncovered in the fosmid sequencing surveys, and within the two species, interclonal identities ranged between 91% and 99% (*R. salina*) and between 80% and 100% (*C. paramecium*; data not shown). Interestingly, most of the clonal heterogeneity observed in the retrotransposon gene fragments amplified from both species was in the form of silent substitutions, suggesting that purifying selection is a factor in the evolution of these genes. In both *R. salina* and *C. paramecium*, a single clone was found to encode an in-frame stop



**Fig. 1.** Presence of multiple copies of *Ty3-gypsy* LTR-retrotransposons in the nuclear genomes of *Rhodomonas salina* and *Cryptomonas paramecium*. Southern hybridizations were performed against restriction enzyme-digested genomic DNAs. **A** *R. salina* genomic DNAs hybridized with a dig-labeled probe specific to the reverse transcriptase domain of the *Ty3-gypsy* LTR-retrotranspo-

son of this species (see text). **B–D** Southern hybridizations using *C. paramecium* genomic DNAs and a *C. paramecium* reverse transcriptase-specific probe. The presence of numerous hybridizing bands suggests that retrotransposons are present in multiple copies in both the *R. salina* and the *C. paramecium* nuclear genomes.

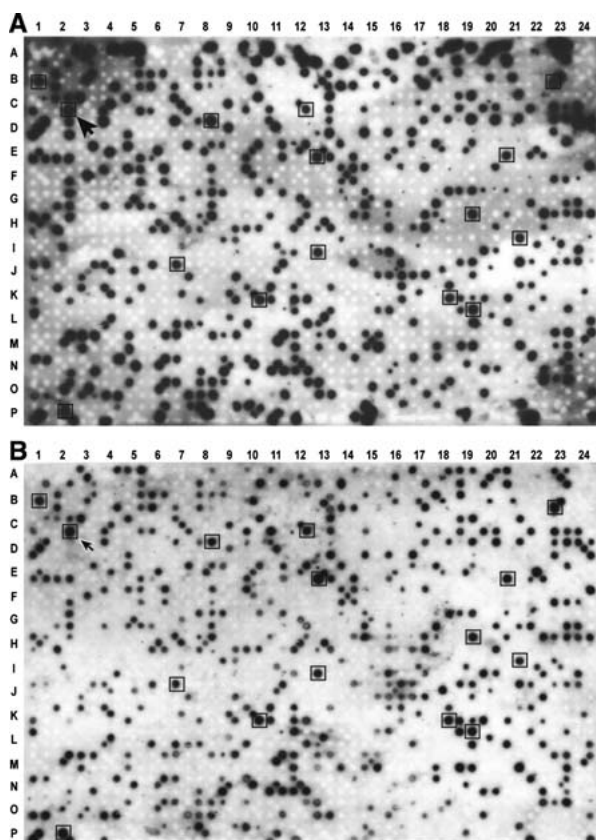
codon; it is unclear whether these represent PCR-generated errors or were amplified from bona fide degenerate retrotransposon genes in the *R. salina* and *C. paramecium* genomes. This caveat aside, our data suggest that significant numbers of these mobile elements may be active at the same time in both genomes.

#### *Structure of the Ty3-gypsy LTR Retrotransposon in Rhodomonas salina*

Long terminal repeat (LTR) retrotransposons constitute one of the two major families of retrotransposons (the other being the non-LTR retrotransposons [Havecker et al. 2004]). As their name suggests, the LTR retrotransposons are characterized by the presence of direct sequence repeats: these repeats flank the *gag* and *pol* genes, which encode proteins involved in the process of transposition. The *gag* gene encodes a structural protein, whereas the *pol* gene encodes a protein with several functions, including protease (PR), RT, and integrase (INT) activities. Approximately 50% of the LTR retrotransposons studied contain a single open reading frame (ORF) fusing the *gag* and *pol* genes, a feature that is particularly prominent in plant sequences (Havecker et al. 2004). A 40- to 50-amino acid-long domain called the “chromodomain” (CD) is also present at the C-terminal region of the INT domain in a variety of plant, animal, and fungal elements (Marín and Lloréns 2000): *Ty3-gypsy* elements that possess a CD are often

referred to as chromoviruses (Kordis 2005). The chromodomain itself is present in a variety of different proteins and is believed to be involved in methylation, chromatin remodeling, and gene expression (Eissenberg 2001; Nielsen et al. 2002). In the context of retrotransposons, the CD domain has been suggested to be involved in targeting the chromovirus to specific sites in the genome (Malik and Eickbush 1999).

In order to determine the structure and organization of the cryptomonad retrotransposons, and to gain further insight into the genomic organization and density of the cryptomonad nuclear genome, we selected 1 of the 15 *Rhodomonas salina* fosmids producing a strong hybridization signal in our high-density filter hybridizations for subcloning and complete sequencing. Shotgun sequencing of this fosmid to a depth of eightfold coverage produced three contigs, the largest of which was 30.9 kb in size. As anticipated, this fragment contained a full-length *Ty3-gypsy* element (Fig. 3) that was determined to be 9097 bp long, on par with such elements found in other nuclear genomes (Kumar and Bennetzen 1999). The LTRs were each 787 bp long and essentially identical to one another (differing at only one nucleotide position). As is the case in LTRs characterized from other organisms, the *R. salina* LTRs possess 5'-TG...CA-3' dinucleotide end sequences. The LTRs were also found to be highly similar (though not identical) to the end sequences of six independent clones analyzed in the *R. salina* fosmid end-sequencing survey (Table 1). These clones pre-



**Fig. 2.** High abundance of *Ty3-gypsy* LTR-retrotransposons in the *Rhodomonas salina* nuclear genome. A duplicate set (A and B) of 1920 randomly chosen fosmid clones was spotted onto high-density filters and used in Southern hybridizations with a *R. salina* probe designed against a portion of the retrotransposon gene encoding the reverse transcriptase domain. Five hundred twenty-one of the 1920 clones (~27%) produced reliable hybridization signals. Fifteen fosmids (boxed) were chosen at random for end-sequencing; none of these clones were found to be identical. One of these fosmids (clone rs04C02; highlighted by an arrow) was sub-cloned and completely sequenced.

sumably also encode partial or full-length retroelements.

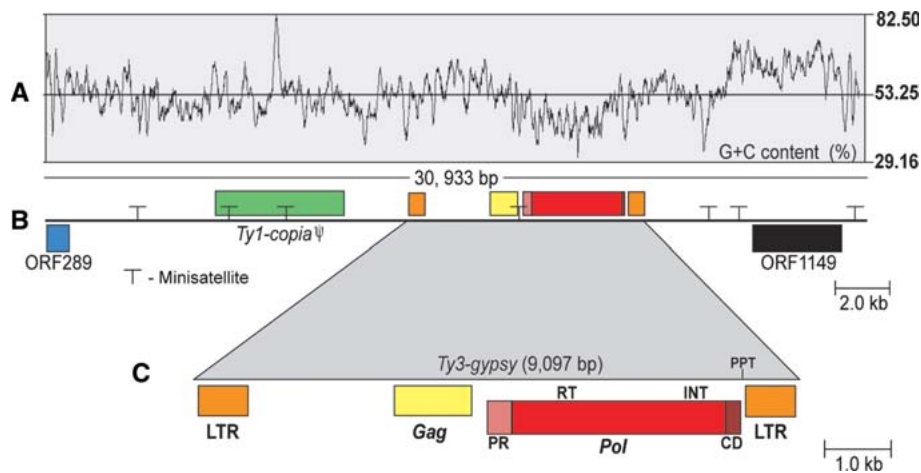
The structure of the full-length *R. salina Ty3-gypsy* retroelement gene and its surrounding genomic context is shown in Fig. 3. A region encoding a canonical zinc finger (ZF) binding domain is observed in the *gag* gene (~1.1 kb), and as has been observed in other systems, a frameshift exists between the *gag* and the *pol* genes. This suggests that the mechanisms regulating the abundance of the two gene products are similar to those present in other organisms. The overall structure of the *R. salina pol* protein is typical of *Ty3-gypsy* elements in that the RT domain precedes the INT domain and a polypurine tract (PPT) is just upstream of the 3' LTR. Although no obvious primer binding site (PBS) with complementarity to the 3' end of initiator methionine tRNA is apparent, the sequence TGG appears immediately downstream of the 5' LTR, indicative of a primer-dependent mechanism of replication (Levin 1995). Intriguingly,

the PR motif at the 5'-end of the *pol* coding region does not contain a start codon and is separated from the RT domain by an in-frame stop codon. This region of the contig was covered by sequences from seven independent sequences in our shotgun assembly, so the stop codon is thus unlikely due to a sequencing error. To determine if this is a general characteristic of the *R. salina Ty3-gypsy* elements or is unique to this clone, we PCR-amplified, cloned, and sequenced DNA fragments spanning the gap between the *gag* and the *pol* coding regions: in none of four independent clones was a stop codon present in this position. A similar situation exists at the 3'-end of the *pol* gene. A putative chromodomain coding region was found in-frame with the *R. salina pol* coding sequence but was separated from the INT domain by a stop codon. The CD coding region also lacked a start codon, and six PCR-generated clones overlapping this area did not possess a stop codon between the INT and the CD coding regions (see Discussion).

In addition to a full-length *Ty3-gypsy* retrotransposon, the ~31-kb fragment of the *R. salina* nuclear genome possesses a *Ty1-copia* element, although its protein coding regions are interrupted by termination codons and minisatellite sequences. LTR sequences were undetectable in the regions adjacent to the gene, suggesting that this particular retroelement is not functional. Two hypothetical ORFs predicted to encode proteins of 1149 and 289 amino acids are also present (Fig. 3). A strong codon bias was observed in both ORFs, suggesting that they are bona fide genes despite having no obvious similarity to known genes in public databases. While the average G + C content of the ~31-kb fragment is ~53%, the G + C content is below average in the region encoding the *pol* protein and above average for ORF1149. The fragment was also found to contain several tandem repeat sequences (Fig. 3).

#### *Phylogenetic Origin of the Cryptomonad Ty3-gypsy LTR Retrotransposons*

In order to infer the evolutionary history of the cryptomonad *Ty3-gypsy* LTR retrotransposons, we constructed phylogenetic trees from amino acid alignments of the RT domain from diverse eukaryotic *Ty3-gypsy* elements and representative viral sequences. Despite its high degree of conservation, the small size of the RT domain makes the construction of robust phylogenies difficult. Our preliminary analyses using partial *C. paramecium* sequences and a truncated alignment (79 unambiguously aligned residues) showed the *R. salina* and *C. paramecium* genes to branch together (data not shown). In the interest of maximizing the amount of informative data, we therefore removed *C. paramecium* and focused on



**Fig. 3.** Schematic of an ~31-kb fragment of the *Rhodomonas salina* nuclear genome. **A** Average G + C content of *R. salina* fosmid clone rs04C02 calculated using a 100-bp sliding window. **B** Putative coding regions. Boxes shown above the line are transcribed left to right; those below, right to left. The locations of minisatellite sequences are highlighted. **C** Closeup of the *Ty3-gypsy* coding region. The various domains of the pol protein are indicated, as is the location of the polypurine tract (PTT). RT, reverse transcriptase; INT, integrase; PR, protease; CD, chromodomain.

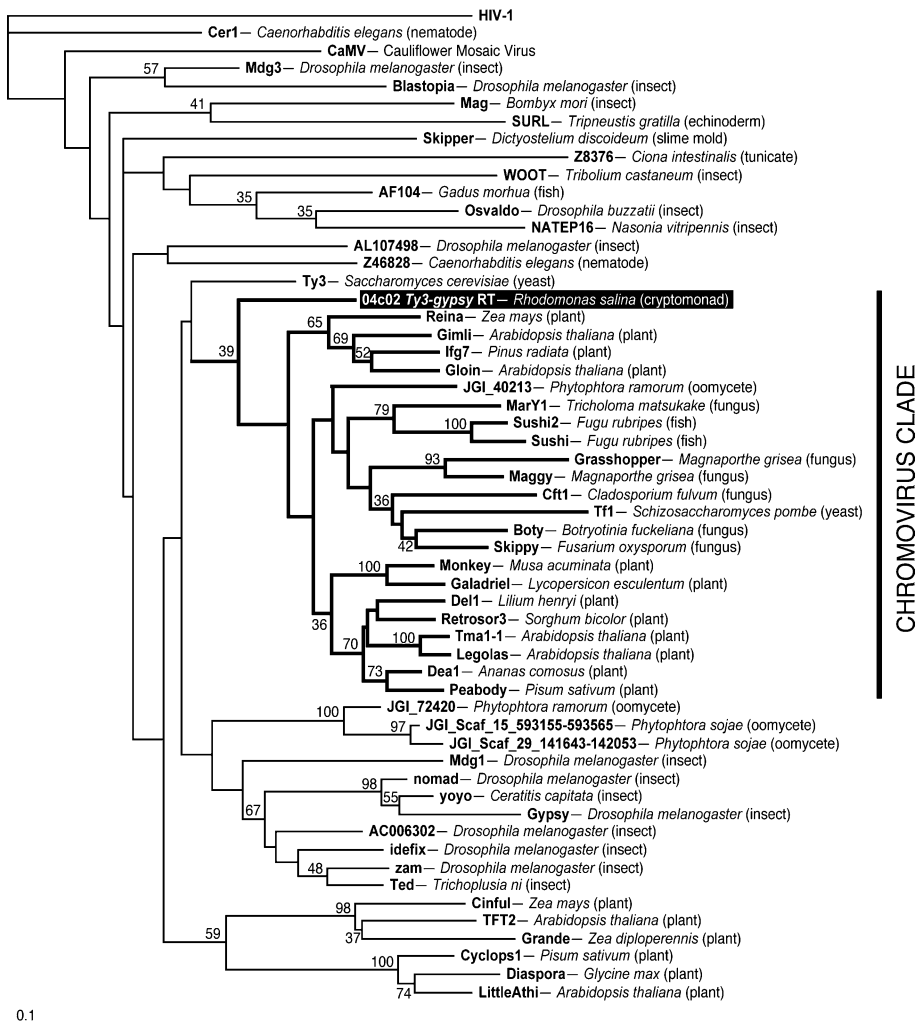
more rigorous analyses of a larger alignment (111 sites) using the *R. salina* fosmid clone sequence as the sole cryptomonad representative. Figure 4 shows a maximum likelihood (ML) phylogeny taking into account among-site rate heterogeneity. Even using the maximum number of confidently aligned sites, statistical support for individual nodes was generally very low. A monophyletic grouping of the chromodomain-containing retroelements (the “chromoviruses”) was obtained using both ML and ML-distance methods (with the exception of the *Skipper* element of *Dictyostelium discoideum*), consistent with the analysis of Marín and Lloréns (2000). Sequences from the oomycete *Phytophthora* branched in two different positions in the phylogeny, within and outside the chromovirus clade. Significantly, the *R. salina* *Ty3-gypsy* sequence determined in this study branched at the base of the cluster of chromodomain-containing sequences, although it did not branch with the *Phytophthora* sequence in this group, as might be expected based on the increasingly accepted relationship among cryptomonads, heterokonts (including oomycetes), and other chromalveolate taxa (Keeling et al. 2005).

## Discussion

We have performed end-sequencing surveys of fosmid libraries constructed from total cellular DNA of *Rhodomonas salina* and *Cryptomonas paramecium*, two distantly related species of cryptomonad algae for which very little molecular data are currently available. The G + C content and (when apparent) gene content of the clones suggest that most are derived from the host nuclear genomes of these organisms, as expected from the stoichiometry of nuclear to organellar DNA in cryptomonads (Douglas et al. 2001). Furthermore, because the fosmid libraries were constructed using nebulized DNA, we have no

reason to believe that the sequenced clones represent a significantly nonrandom sampling of the *R. salina* and *C. paramecium* genomes. Therefore, although small in scale, these surveys provide a first glimpse of the structure and composition of the cryptomonad nuclear genome.

The fosmid clone end-sequencing surveys of *Rhodomonas salina* and *Cryptomonas paramecium* clearly show that the nuclear genomes of both organisms contain numerous retrotransposons and tandem repeat sequences. In the case of retrotransposons, these observations were confirmed using Southern hybridizations against restricted genomic DNAs (Fig. 1) and high-density filters containing fosmid clones (Fig. 2). In *R. salina*, the bulk of the retrotransposon sequences belong to the *Ty3-gypsy* family, and the most significant BLASTX matches were to retrotransposons in various plant species (especially *Oryza sativa*). Given that the *Ty3-gypsy* retroelements appear to be relatively rare outside animals, fungi, and plants (Dusan 2005; Gorinsek et al. 2004), this raises the intriguing possibility that the cryptomonad *Ty3-gypsy* elements are derived from the nuclear genome of the red algal endosymbiont that donated the plastid to this lineage. Unfortunately, this could not be tested using phylogenetic analysis due to the fact that the complete genome of the red alga *Cyanidioschyzon merolae* (Matsuzaki et al. 2004) does not contain these elements. Although two small fragments of *Ty3-gypsy* elements are available from *Porphyra yezoensis*, they were too short to include in our analyses. With respect to other chromalveolate taxa, *Ty3-gypsy* elements appear to be absent in apicomplexans, and although they do exist in the complete genome of the diatom *Thalassiosira pseudonana* (Armbrust et al. 2004), they contain numerous stop codons and are extremely divergent. The only chromalveolate *Ty3-gypsy* elements useful for comparison were from two species of the oomycete *Phytophthora*, and, as is the case for cryp-



**Fig. 4.** Protein maximum likelihood phylogeny of the reverse transcriptase domain of diverse *Ty3-gypsy* retrotransposons, rooted with HIV-1. The chromodomain-containing sequences are labeled, as is the *Rhodomonas salina Ty3-gypsy* sequence determined in this study. The PHYML bootstrap values > 35% are indicated. The scale bar indicates the inferred number of amino acid substitutions per site.

tomonads, these have been reported to possess a chromodomain (Dusan 2005). Based on this observation, it seems probable that the chromoviruses were present in the common ancestor of cryptomonads and heterokonts, and perhaps in the common ancestor of all chromalveolates, although none of the *Phytophthora* sequences branch specifically with *R. salina* in our phylogenies (Fig. 4). More comprehensive genome sequence data from *Ty3-gypsy*-containing red algae and additional chromalveolate species will be needed to better resolve the origin and diversification of the *Ty3-gypsy* retrotransposons in cryptomonads and their closest relatives.

Despite the present uncertainty surrounding their evolutionary origin, the *Ty3-gypsy* elements in cryptomonads are interesting from a functional perspective. The *R. salina* elements possess canonical *gag* and *pol* coding regions separated by a frameshift. Functional studies of the mechanism of LTR retrotransposition have shown that much more Gag protein is required than Pol (Gao et al. 2003). While ~50% of examined LTR retrotransposons contain a single open reading frame (ORF) fusing the *gag* and *pol*

genes (Havecker et al. 2004), LTR retrotransposons can also use a “recoding” mechanism for regulating dosage. In this case, the *gag* and *pol* ORFs are separated by either a frameshift or a stop codon, which are occasionally ignored by the translation machinery, resulting in the synthesis of *pol* protein. The presence of a frameshift between the *gag* and the *pol* ORFs in the *R. salina* LTR retrotransposons suggests that a similar recoding mechanism appears to exist in *R. salina*. In addition to this frameshift, we found that in the full-length *Ty3-gypsy* retroelement obtained by complete fosmid sequencing (Fig. 3), the protease (PT) and chromodomain (CD) regions are separated from the reverse transcriptase (RT) and integrase (INT) domains of *pol* by in-frame stop codons, and neither the PT nor the CD coding regions possess start codons themselves. In contrast, the sequences of multiple PCR-amplified *Ty3-gypsy* element fragments lacked stop codons in these regions. The fact that the 5' and 3' LTRs on the fosmid-derived element are essentially identical to one another (differing at one position over 787 bp) suggests that this particular element is the product of a very recent

insertion. Although the in-frame stop codons separating the PT-RT and INT-CD domains may be the first signs of degeneration in this element, it is also possible that they represent a biologically significant regulatory mechanism present in a subset of *Ty3-gypsy* retroelements in the cryptomonad nuclear genome.

The fact that so few of the *R. salina* and *C. paramecium* end sequences determined in this study match known genes/proteins is most likely due to the large evolutionary distance between the cryptomonads and other eukaryotes for which genome sequences are available. Outside of the apicomplexans, the diatom *Thalassiosira pseudonana* is currently the only chromalveolate whose genome has been completely sequenced (Armbrust et al. 2004), and although cryptomonads and heterokonts (to which diatoms belong) have long been thought to be related, robust evidence for their common ancestry has been slow in coming (reviewed by Archibald and Keeling 2005; Bodyl 2005; Keeling 2004; Palmer 2003). From the perspective of gene identification/annotation, the wide evolutionary gulf between the cryptomonads and other eukaryotes means that many bona fide protein-coding genes may go unrecognized, particularly in an end-sequencing survey such as ours where read lengths are short and unedited sequences are used as queries in BLAST searches. This “sampling” phenomenon was apparent in the annotation of the *T. pseudonana* genome: of the 11,242 protein coding genes predicted in *T. pseudonana* genome, almost a third showed no similarity to those in other proteomes (Armbrust et al. 2004).

Another factor to consider is genome size. Beaton and Cavalier-Smith (1999) used flow cytometry to estimate the nuclear DNA content of 17 diverse cryptomonad species and observed a wide variation, between 0.72 and 8.8 pg per cell. Assuming that these species are predominantly haploid (Beaton and Cavalier-Smith 1999) and that 1 pg = 980 Mb, these numbers correspond to genome sizes of between ~700 and 8600 Mb. *Cryptomonas paramecium* was not examined in the study, but *Rhodomonas salina* CCMP1319 was estimated to have a DNA content of 3.72 pg per cell, corresponding to a presumed 1C nuclear genome size of ~3600 Mb. Gene density in a genome of this size would presumably be very low. We have been unable to confirm or refute the genome size predictions of Beaton and Cavalier-Smith for *R. salina*, but for several other cryptomonads, including *Guillardia theta* CCMP327 and *Hemiselmis rufescens* CCMP644, our pulsed-field gel electrophoresis and flow cytometric data suggest that their results are a significant overestimate of DNA content and genome size (Archibald Lab, unpublished data). A more complete picture of the size, structure, and composition of cryptomonad nuclear genomes will ultimately

depend on rigorous genome size surveys and, where possible, complete genome sequences.

**Acknowledgments.** We thank Iñaki Ruiz-Trillo and two anonymous reviewers for their helpful comments on the manuscript, Ignacio Marín for kindly providing a protein alignment of *Ty3-gypsy* reverse transcriptase domains, and Jessica Leigh for assistance with phylogenetic analysis. This work was supported by Genome Atlantic and an NSERC discovery grant (283335-2004) awarded to J.M.A. J.M.A. is a Scholar of the Canadian Institute for Advanced Research, Program in Evolutionary Biology.

## References

- Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 304:441–445
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Archibald JM (2005) Jumping genes and shrinking genomes—probing the evolution of eukaryotic photosynthesis using genomics. *IUBMB Life* 57:539–547
- Archibald JM, Keeling PJ (2002) Recycled plastids: a green movement in eukaryotic evolution. *Trends Genet* 18:577–584
- Archibald JM, Keeling PJ (2005) On the origin and evolution of plastids. In: Saap J (ed) *Microbial phylogeny and evolution*. Oxford University Press, New York, pp 238–260
- Archibald JM, Cavalier-Smith T, Maier U, Douglas S (2001) Molecular chaperones encoded by a reduced nucleus—the cryptomonad nucleomorph. *J Mol Evol* 52:490–501
- Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ (2003) Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. *Proc Natl Acad Sci USA* 100:7678–7683
- Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86
- Beaton M, Cavalier-Smith T (1999) Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. *Proc R Soc Lond B* 266:2053–2059
- Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580
- Bhattacharya D, Yoon HS, Hackett JD (2003) Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays* 26:50–60
- Bodyl A (2005) Do plastid-related characters support the chromalveolate hypothesis? *J Phycol* 41:712–719
- 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

- Deane JA, Fraunholz M, Su V, Maier UG, Martin W, Durnford DG, McFadden GI (2000) Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. *Protist* 151:239–252
- Dear S, Staden R (1991) A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res* 19:3907–3911
- Delwiche CF (1999) Tracing the thread of plastid diversity through the tapestry of life. *Am Nat* 154(Suppl):S164–S177
- Douglas SE, Penny SL (1999) The plastid genome from the cryptomonad alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J Mol Evol* 48:236–244
- Douglas SE, Murphy CA, Spencer DF, Gray MW (1991) Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* 350:148–151
- Douglas SE, Zauner S, Fraunholz M, Beaton M, Penny S, Deng L, Wu X, Reith M, Cavalier-Smith T, Maier U-G (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* 410:1091–1096
- Dusan K (2005) A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene* 347:161–173
- Eissenberg JC (2001) Molecular biology of the chromo domain: an ancient chromatin module comes of age. *Gene* 275:19–29
- El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Wortley EA, Delcher AL, Blandin G, Westerberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaude F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazalina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B (2005a) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309:409–415
- El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Wortley EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaude F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westerberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N (2005b) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309:404–409
- Elias MC, Vargas NS, Zingales B, Schenkman S (2003) Organization of satellite DNA in the genome of *Trypanosoma cruzi*. *Mol Biochem Parasitol* 129:1–9
- Flavell RB (1986) Repetitive DNA and chromosome evolution in plants. *Philos Trans R Soc Lond B Biol Sci* 312:227–42
- Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF (2003) Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* 9:1422–1430
- Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511
- Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJ, Sato S, Ralph SA, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Pertea M, Allen J, Nierman WC, Taracha EL, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 309:134–137
- Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695
- Gilson PR (2001) Nucleomorph genomes: much ado about practically nothing. *Genome Biol* 2:REVIEWS1022
- Gilson PR, McFadden GI (2002) Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. *Genetica* 115:13–28
- Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI (2006) From the cover: complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. *Proc Natl Acad Sci USA* 103:9566–9571
- Gorinsek B, Gubensek F, Kordis D (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol Biol Evol* 21:781–798
- Gregory TR (2001) Coincidence, coevolution or causation? DNA content, cell size, and the C-value enigma. *Biol Rev* 76:65–101
- Gregory TR (2005) The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann Bot* 95:133–146
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704
- Hancock JM (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* 115:93–103
- Harper JT, Waanders E, Keeling PJ (2005) On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int J Syst Evol Microbiol* 55:487–496
- Hauth AM, Maier UG, Lang BF, Burger G (2005) The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region. *Nucleic Acids Res* 33:4433–4442
- Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. *Genome Biol* 5:225
- Hoef-Emden K (2005) Multiple independent losses of photosynthesis and differing evolutionary rates in the genus *Cryptomonas* (Cryptophyceae): combined phylogenetic analyses of DNA sequences of the nuclear and the nucleomorph ribosomal operons. *J Mol Evol* 60:183–195
- Hoef-Emden K, Marin B, Melkonian M (2002) Nuclear and nucleomorph SSU rDNA phylogeny in the Cryptophyta and the evolution of cryptophyte diversity. *J Mol Evol* 55:161–179
- Ishida K, Cao Y, Hasegawa M, Okada N, Hara Y (1997) The origin of chlorarachniophyte plastids, as inferred from phylogenetic comparisons of amino acid sequences of EF-Tu. *J Mol Evol* 45:682–687
- Ishida K, Green BR, Cavalier-Smith T (1999) Diversification of a chimaeric algal group, the chlorarachniophytes: phylogeny of nuclear and nucleomorph small-subunit rRNA genes. *Mol Biol Evol* 16:321–331
- Keeling PJ (2004) Diversity and evolutionary history of plastids and their hosts. *Am J Bot* 91:1481–1493

- Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW (2005) The tree of eukaryotes. *Trends Ecol Evol* 20:670–676
- Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115:49–63
- Kordis D (2005) A genomic perspective on the chromodomain-containing retrotransposons: chromoviruses. *Gene* 347:161–173
- Kumar A, Bennetzen JL (1999) Plant retrotransposons. *Annu Rev Genet* 33:479–532
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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 JP, 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 JC, Mungall A, Plumb R, Ross M, Showkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lane CE, Khan H, MacKinnon M, Fong A, Theophilou S, Archibald JM (2006) Insight into the diversity and evolution of the cryptomonad nucleomorph genome. *Mol Biol Evol* 23:856–865
- Levin HL (1995) A novel mechanism of self-primed reverse transcription defines a new family of retroelements. *Mol Cell Biol* 15:3310–3317
- Liolios K, Tavernarakis N, Hugenholtz P, Kyrpidis NC (2006) The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* 34:D332–334
- Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M, Ackers J, Tannich E, Leippe M, Hofer M, Bruchhaus I, Willhoeft U, Bhattacharya A, Chillingworth T, Churcher C, Hance Z, Harris B, Harris D, Jagels K, Moule S, Mungall K, Ormond D, Squares R, Whitehead S, Quail MA, Rabbinowitsch E, Norbertczak H, Price C, Wang Z, Guillen N, Gilchrist C, Stroup SE, Bhattacharya S, Lohia A, Foster PG, Sicheritz-Ponten T, Weber C, Singh U, Mukherjee C, El-Sayed NM, Petri WA Jr, Clark CG, Embley TM, Barrell B, Fraser CM, Hall N (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature* 433:865–868
- Malik HS, Eickbush TH (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J Virol* 73:5186–5190
- Marin I, Lloréns C (2000) Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. *Mol Biol Evol* 17:1040–1049
- Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, 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, Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653–657
- McFadden GI, Gilson PR, Waller RF (1995) Molecular phylogeny of chlorarachniophytes based on plastid rRNA and *rbcL* sequences. *Arch Protistenkd* 145:231–239
- McGrath CL, Katz LA (2004) Genome diversity in microbial eukaryotes. *Trends Ecol Evol* 19:32–38
- Nene V, Morzaria S, Bishop R (1998) Organisation and informational content of the *Theileria parva* genome. *Mol Biochem Parasitol* 95:1–8
- Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416:103–107
- Palmer JD (2003) The symbiotic birth and spread of plastids: How many times and whodunnit? *J Phycol* 39:4–11
- Rensing SA, Goddemeier M, Hofmann CJ, Maier UG (1994) The presence of a nucleomorph hsp70 gene is a common feature of Cryptophyta and Chlorarachniophyta. *Curr Genet* 26:451–455
- Requena JM, Lopez MC, Alonso C (1996) Genomic repetitive DNA elements of *Trypanosoma cruzi*. *Parasitol Today* 12:279–283
- Sloof P, Bos JL, Konings AF, Menke HH, Borst P, Gutteridge WE, Leon W (1983) Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. *J Mol Biol* 167:1–21
- Stover NA, Krieger CJ, Binkley G, Dong Q, Fisk DG, Nash R, Sethuraman A, Weng S, Cherry JM (2006) Tetrahymena Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. *Nucleic Acids Res* 34:D500–D503
- Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* 13:964–969
- Van der Auwera G, Hofmann CJB, De Rijk P, De Wachter R (1998) The origin of red algae and cryptomonad nucleomorphs: A comparative phylogeny based on small and large subunit rRNA sequences of *Palmaria palmata*, *Gracilaria verrucosa*, and the *Guillardia theta* nucleomorph. *Mol Phylogenet Evol* 10:333–342
- Wickstead B, Ersfeld K, Gull K (2003) Repetitive elements in genomes of parasitic protozoa. *Microbiol Mol Biol Rev* 67:360–375
- Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA (2004) The genome of *Cryptosporidium hominis*. *Nature* 431:1107–1112
- Zagulski M, Nowak JK, Le Mouel A, Nowacki M, Migdalski A, Gromadka R, Noel B, Blanc I, Dessen P, Wincker P, Keller AM, Cohen J, Meyer E, Sperling L (2004) High coding density on the largest *Paramecium tetraurelia* somatic chromosome. *Curr Biol* 14:1397–1404