Genetica **115:** 13–28, 2002.

Jam packed genomes – a preliminary, comparative analysis of nucleomorphs

Paul R. Gilson[1] & Geoffrey I. McFadden[2,*]
[1]*Centre for Cellular and Molecular Biology, School of Biological and Chemical Sciences, Deakin University, VIC, 3125, Australia;* [2]*Plant Cell Biology Research Centre, School of Botany, University of Melbourne, VIC, 3010, Australia;* **Author for correspondence (Phone: +61 3 8344 5054 (office); Fax: +61 3 9347 1071; E-mail: g.mcfadden@unimelb.edu.au)*

Key words: chloroplast, chlorarachniophyte, cryptomonad, C-value enigma, endosymbiosis, intron, mitosis, nucleomorph, photosynthesis, secondary plastid, telomere, transposable element

Abstract

There are two ways eukaryotic cells can permanently acquire chloroplasts. They can take up a cyanobacterium and turn it into a chloroplast or they can engulf an alga that already has a chloroplast. The second method is far more common and there are at least seven major groups of protists that have obtained their chloroplasts, this way. In most cases little remains of the engulfed alga apart from its chloroplast, but in two groups, the cryptomonads and chlorarachniophytes, a small remnant nucleus of the engulfed alga is still present. These tiny nuclei, called nucleomorphs, are the smallest and most compact eukaryotic genomes known and recently the nucleomorph of the cryptomonad alga *Guillardia theta*, was completely sequenced (551 kilobases). The nucleomorph of the chlorarachniophyte *Bigellowiella natans* (380 kilobases), is also being sequenced and is about half complete. We discuss some of the similarities and differences that are emerging between these two nucleomorph genomes. Both genomes contain just three chromosomes that encode mainly housekeeping genes and a few proteins for chloroplast functions. The bulk of nucleomorph gene coding capacity, therefore, appears to be devoted to self perpetuation and creating gene and protein expression machineries to make a small number of essential chloroplast proteins. We discuss reasons why both nucleomorphs are extraordinarily compact and why their gene sequences are evolving rapidly.

Abbreviations: kb – kilobase; bp – base pair; nt – nucleotide; ER – endoplasmic reticulum.

Introduction

The physical and genetic fusion of cells to create new living chimeras has been a major driver of biological innovation (Margulis & Chapman, 1998). These mergers create organisms that can exploit new environmental niches, out-compete rivals and proliferate by having more offspring. Photosynthesis, allows an organism to manufacture its own food instead of catching it and has been a major impetus for many fusion events. Cyanobacteria with the ability to harness the sun's energy to make food were no doubt preyed upon by early eukaryotes. But not all these meals were digested. Some eukaryotes probably found it advantageous to delay the digestion of their cyanobacterial prey and encourage them to keep on photosynthesising and to leak their carbohydrates. Selection eventually favoured increased integration of this arrangement and the first eukaryotic alga, replete with a plastid, was born (Figure 1). The term 'plastid' is used to generically describe any photosynthetic organelle. The word 'chloroplast', though often used in place of plastid, specifically refers to green plastids of green algae and land plants. The largest group of organisms to have acquired plastids directly this way are the Kingdom Plantae (Cavalier-Smith, 1999, 2000), which comprises red algae, glaucocystophytes and green algae (land plants are descendants of green algae)

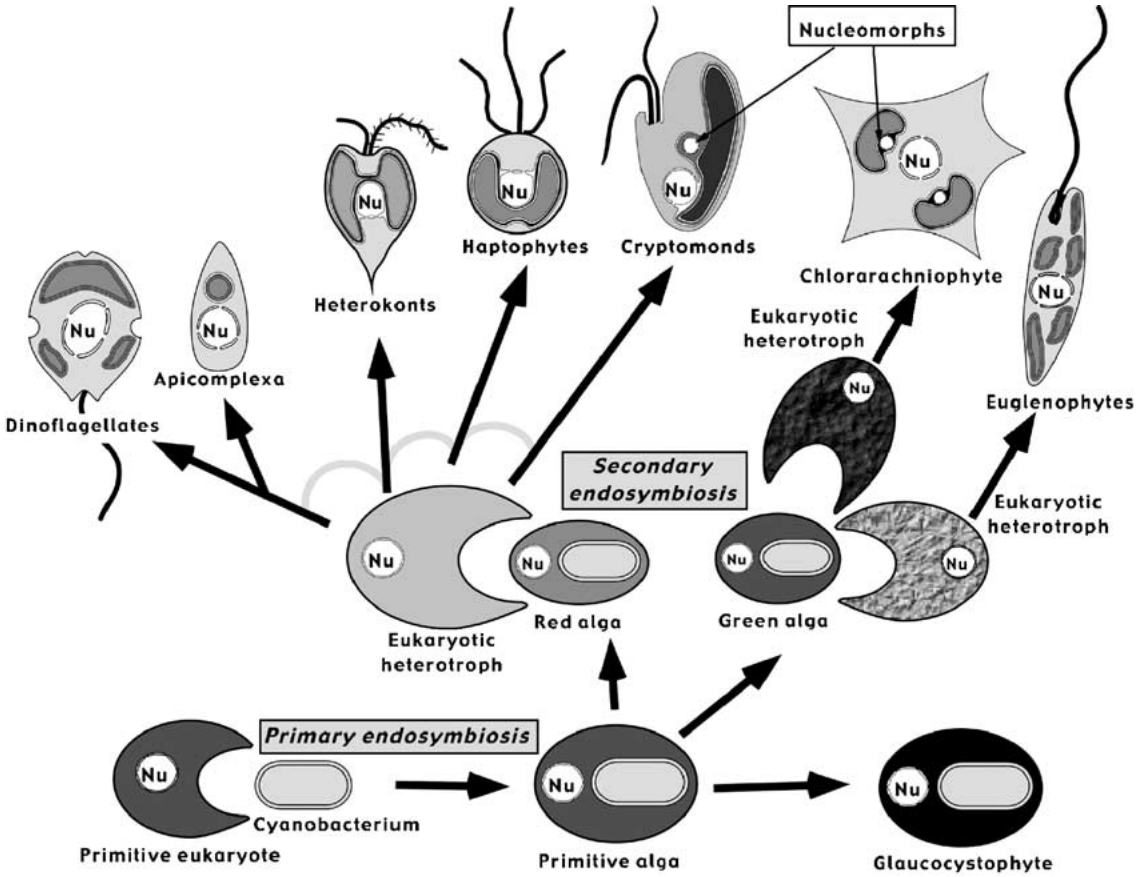

Figure 1. Proposed evolutionary scheme that has created most extant groups of algae. The primary plastids of green and red algae and glaucocystophytes share a common ancestry being derived from the engulfment of a cyanobacterium by a primitive eukaryote. Some of these algae were in turn engulfed by other eukaryotic cells to create secondary or complex plastids. Dinoflagellates/apicomplexa, heterokonts, haptophytes and cryptomonads engulfed red algae to acquire their plastids and they may be derived from a common ancestor (indicated by grey line). Chlorarachniophytes and euglenophytes obtained their plastids by independently taking up green algae. In most cases all genetic information needed to operate the plastid was transferred to the host cell nucleus from the nucleus of the engulfed algae that in turn completely disappeared. In cryptomonads and chlorarachniophytes relic nuclei of the engulfed alga (called a nucleomorphs) remain.

(Figure 1). These organisms appear to be derived from a common ancestor (Martin et al., 1998; Moreira, LeGuyader & Philipe, 2000) that obtained its plastid by modifying an engulfed cyanobacterium in a process called primary endosymbiosis (Figure 1). Primary plastids are bound by two membranes derived from the cyanobacterium's double envelope (Cavalier-Smith, 1995a).

While the Plantae are a widespread and enormously successful group, they are not the only photosynthetic eukaryotes. Several other groups have acquired a photosynthetic capacity less directly by taking up algae and using their plastids (Figure 1). We refer to this as secondary endosymbiosis and it is thought to have occurred at least twice (Cavalier-Smith, 1999) but possibly as many as seven times (Douglas, 1998; Palmer & Delwiche, 1998) (Figure 1). Intriguingly, the prevalence of secondary endosymbiotic acquisition compared to primary acquisition (of which only one seems to have persisted) suggests it is easier to acquire a plastid once it has already been 'broken in' by another eukaryote. The major groups with secondary plastids are the heterokonts, haptophytes, dinoflagellates/apicomplexa, euglenophytes, cryptophytes and chlorarachniophytes (Figure 1).

In heterokonts and haptophytes the plastid is bound by four membranes. The innermost pair is equivalent to the standard two membranes of Plantae plastids and the third membrane is the remnant plasma membrane of the engulfed alga, now known as the periplastid membrane (Figure 1). The plastid resides

inside the endoplasmic reticulum (ER) which thus forms the outermost, fourth membrane. Apart from the plastid, periplastid membrane and masses of genes transferred from the now extinct algal nucleus to the host-cell nucleus, no other residue of the engulfed alga remains.

The plastids of euglenophytes and dinoflagellates are wrapped in three membranes (Figure 1). Once again the inner pair are believed to be equivalent to normal plastid membranes. The origins of the third or outermost membrane are uncertain but recent insights into the targeting mechanisms that nuclear-encoded *Euglena* proteins employ to return to the plastid suggest it is a part of the host cell's endomembrane system (Sulli et al., 1999). This membrane is called the perialgal membrane and was originally a food vacuole (Cavalier-Smith, 1999). Reduction therefore is even more extreme in the dinoflagellates and euglenophytes with the periplastid membrane having been completely lost.

Apicomplexa include a number of parasitic groups such as disease-causing parasites like *Plasmodium*, the causative agent of malaria (Figure 1). The plastids in these organisms are non-photosynthetic but are thought to be retained to produce essential compounds such as fatty acids and isoprenoids (Waller et al., 1998; van Dooren et al., 2000). Apicomplexa are the sister group of the dinoflagellates (Wolters, 1991) but apicomplexan plastids are surrounded by four membranes (McFadden & Roos, 1999) whereas most dinoflagellate plastids have three bounding membranes (Figure 1). It has been suggested the plastids of both groups share a common algal ancestor but in the apicomplexa the periplastid membrane is still present (Cavalier-Smith, 1999).

Gene sequence, plastid pigment and structural analyses have proven useful in tracing the origins of all these plastids (Palmer & Delwiche, 1998). The plastids of euglenophytes and chlorarachniophytes are believed to be derived from the uptake of green algae (Van de Peer et al., 1996; Martin et al., 1998) and all the others come from red algae (Douglas, 1998; Martin et al., 1998; Palmer & Delwiche, 1998) (Figure 1). But genes and cell structure provide only vague clues about the endosymbiotic processes that have taken place to create most of these organisms. What is needed are missing links, cells in which the endosymbiotic process can be viewed mid-point before the virtual elimination of the engulfed algal cell. Fortunately such cells exist; they are the cryptomonads and chlorarachniophytes.

Cryptomonads

Ultrastructural details of cryptomonad cells were first revealed in the 1970s (Greenwood, 1974; Greenwood, Griffiths & Santore, 1977). They are biflagellated unicells whose plastid is surrounded by four membranes (Figure 1). Cryptomonad plastids resemble those of heterokonts in that the outermost membrane is contiguous with the rough ER. However, a major difference is that in cryptomonads there is an extended space between the inner and outer pair of plastid membranes. This region called the periplastidal space, is the relic cytoplasm of the engulfed alga and contains a tiny nucleus called a nucleomorph (Figure 1) as well as eukaryotic-sized ribosomes and starch grains. The nucleomorph of the engulfed alga still retains many of the trappings of an ordinary nucleus. It has a double membrane with pores, contains DNA (Hansmann et al., 1986) and even has a nucleolar-like region. The discovery of nucleomorphs was a major boost to the theory that plastids bound by more than two membranes were derived from engulfed algae (Gibbs, 1981; Whatley, 1981).

Chlorarachniophytes

Chlorarachniophytes are marine amoeboflagellates in which a nucleomorph was first revealed in 1984 (Hibberd & Norris, 1984). The nucleomorph in chlorarachniophytes, similar to cryptomonads, is located between the inner and outer pair of membranes enveloping the plastid. Nevertheless, there are several important differences between the two groups (Figure 1). Firstly, the host cells of both organisms are unrelated. Although gene phylogenies fail to firmly resolve what group of organisms are the sister group of the biflagellated cryptomonad host cell there are powerful evolutionary arguments that they may be related to the host of heterokonts, haptophytes and dinoflagellates/apicomplexa (Cavalier-Smith, 1999). GAPDH gene data also add some support to this conclusion (Fast et al., 2001). The phylogenetic affinities of chlorarachniophyte host cells are much more firmly established. They are related to other amoeba and flagellated heterotrophs called the Cercozoa (Cavalier-Smith & Chao, 1997; Cavalier-Smith, 1998), which in turn may be related to the foraminifera (Keeling, 2001).

Another important difference between chlorarachniophytes and cryptomonads is the origins of their

endosymbionts. Chlorarachniophytes and cryptomonads derived their plastids from engulfed green (Van de Peer et al., 1996) and red algae, respectively (Figure 1) (Douglas et al., 1991; Cavalier-Smith et al., 1996), although in both cases it is not clear from which sub groups within the greens or the reds they were sourced. Unlike cryptomonads, where the plastid complex resides in the ER, in chlorarachniophytes it is located in the cytoplasm situated in a modified food vacuole called the perialgal membrane (Figure 1). Despite these differences, chlorarachniophyte nucleomorphs bear an uncanny similarity to those of cryptomonads. Chlorarachniophyte nucleomorphs contain DNA, have pores and nucleolar-like regions as well as dense bodies of unknown function. Thus it was of great interest to discover how these miniature nuclei organised their DNA and what they encoded.

Nucleomorphs are tiny nuclei containing three short chromosomes

Insights into the genomic size, organisation and gene complement of nucleomorphs emerged rapidly in the early 1990s. Methods were developed to isolate cryptomonad nucleomorphs and estimate their DNA contents which at a mere 1.3–2.8 megabases (0.1% of total cellular DNA) were only 1/700 the DNA size of the nuclei of their host cells (Hansmann & Eschbach, 1990). The first gene isolated from nucleomorphs (from *Cryptomonas* Φ) encoded a eukaryotic type small subunit ribosomal RNA (srRNA). Phylogenetic analysis of this srRNA and another derived from the host cell indicated that cryptomonads were *bona fide* eukaryotic chimeras with an endosymbiont derived from a red alga (Douglas et al., 1991).

A major break-through in unravelling nucleomorph genome structure occurred when Maier et al. (1991) and Eschbach et al. (1991) separated chromosomal DNA obtained from isolated nucleomorphs by pulsed field gel electrophoresis (PFGE). In the cryptomonad *Pyrenomonas salina*, three tiny, linear chromosomes of sizes 195, 225 and 240 kb were observed (Maier et al., 1991). These chromosomes each appeared to encode srRNA genes of red algal provenance and the total genome size of the *P. salina* nucleomorph at a mere 660 kb, was the smallest eukaryotic genome identified at the time (Eschbach et al., 1991). Small nucleomorph karyotypes are common to other cryptomonads. Blots of pulsed field gels probed with srRNA gene probes have confirmed that other cryptomonad nucleomorphs harbour three chromosomes with total genome sizes ranging from 450 to 710 kb (Rensing et al., 1994). Based on the total amount of DNA calculated to reside inside the nucleomorph of *P. salina* and its haploid genome size, the ploidy of the nucleomorph is anticipated to be $2n$ and $4n$ prior to division (Hansmann & Eschbach, 1990). Could nucleomorphs, despite their diminutive size, still divide by ordinary mitosis after doubling their chromosomes?

Surprisingly, the chlorarachniophyte nucleomorph also contains three chromosomes, and they are similarly sized to the cryptomonad nucleomorph chromosomes (McFadden et al., 1994; Gilson & McFadden, 1995, 1999). It was not feasible to isolate chlorarachniophyte nucleomorphs (unlike their cryptomonads counterparts) so it was not possible to determine the karyotype of chlorarachniophyte nucleomorphs directly from a purified source. An alternative approach to identifying nucleomorph chromosomes was employed by using DNA elements expected to be present on every single nucleomorph chromosome - telomeres (Gilson & McFadden, 1995). Telomeres form protective caps on the ends of chromosomes and usually consist of numerous, simple DNA repeats. A telomere (containing multiple repeats of $TCTAGGG_n$) cloned from the nucleomorph genome (see below) hybridised to three small chromosomes sized 98, 140 and 145 kb separated by PFGE (Gilson & McFadden, 1995). Several species and strains of chlorarachniophyte have now been examined and they all have three nucleomorph chromosomes with total genome sizes ranging from 380–455 kb making them even smaller than those of the cryptomonads (Gilson & McFadden, 1999).

Why do nucleomorphs have three chromosomes?

The first plausible hypothesis as to why cryptomonad and chlorarachniophyte nucleomorphs have three similarly sized chromosomes was recently provided by Douglas et al. (2001). The genome sizes of both nucleomorphs are about the same since they encode a similar core of genes (see below). Douglas et al. (2001) calculate that once nucleomorph chromosomes are wrapped around their histone cores (and contract 1/40th their original length) to form 30 nm filaments they would be about 1.5 μm long, the same width as a nucleomorph. If there were fewer than three chromosomes they would have to be longer than their present maximum of about ∼200 kb and would be too long to segregate from each other during mitosis within

the confines of the nucleomorph/periplastidal compartment. Nucleomorph chromosome are probably too short to be packaged into higher order structures like the chromosomes of typical nuclei. At the other end of the scale, chromosomes that are less than about 100 kb may be too small to remain viable, being lost during mitosis (Murray & Szostak, 1985). Thus the two genomes may be forced to maintain chromosomes of ~100–200 kb. This combination of upper and lower bounds for chromosome size consequently forces the 300–550 genes in the two nucleomorphs to be spread across three similarly sized chromosomes.

The architecture of nucleomorph chromosomes is remarkably similar

Once it had been established that nucleomorphs contained three chromosomes it became feasible to explore the functions of these tiny nuclei by sequencing their genomes. The chlorarachniophyte species *Bigelowiella natans* (Moestrup & Sengco, 2001), was chosen for genomic analysis because it grows to high density in aerated cultures. Efforts to map and sequence the nucleomorph initially targeted the smallest chromosome called chromosome III (98 kb) because it can be easily separated from chromosome I and II (145 and 140 kb, respectively) by PFGE (Gilson & McFadden, 1996). Plasmid libraries of restriction fragments from chromosome III were constructed and screened with a gene known to reside upon each chromosome, the srRNA gene. A complete rRNA gene cistron comprising srRNA, 5.8S and large subunit rRNA genes was isolated and revealed that nucleomorph rDNA units were similar to most other eukaryotes (Gilson & McFadden, 1995, 1996). To clone the telomeres from chromosome III another plasmid library enriched for DNA fragments from chromosome ends was created. Random sequencing identified a likely telomere clone that contained 32 repeats of the telomere-like motif $TCTAGGG_n$ (Gilson & McFadden, 1995). Curiously, mapping of the telomere and srRNA genes indicated that they were linked (Gilson & McFadden, 1995). Only one rDNA unit is linked to each telomere and there are no other copies upon each chromosome (Gilson & McFadden, 1995). Furthermore, each telomere/rDNA unit is nearly identical and in effect form 8.5 kb inverted repeats upon the ends of each chromosome (Gilson & McFadden, 1995). The telomere/rDNA units act as book ends in between which is nestled single copy DNA that encodes protein genes, tRNAs and snRNAs (Gilson & McFadden, 1996).

The cryptomonad species chosen for genome analysis was *Guillardia theta* since it had one of the smallest nucleomorphs known (Rensing et al., 1994) and its plastid and mitochondrial genomes had already been sequenced (Douglas & Penny, 1999). Mapping and sequencing the cryptomonad nucleomorph genome has proceeded in a similar way to the chlorarachniophyte project. Characterisation of cryptomonad nucleomorph rRNA genes and telomeres superficially revealed an uncanny resemblance to the chlorarachniophyte version (Zauner et al., 2000). In both genomes a single rDNA unit is linked to a telomere and this unit is repeated on each end of each chromosome book-ending the single copy DNA (Gilson & McFadden, 1995; Zauner et al., 2000). While the architecture of these chromosomes is remarkably similar, there are some marked differences in the details. The telomere repeats of the two nucleomorphs are radically different, with a plant and green algal-like motif TCTAGGG found in chlorarachniophytes and a $(AG)_7AAG_6A$ motif found in cryptomonads (at present it is unknown if this is similar to red algae). Another difference between the two genomes is that rDNA units are oriented in opposite directions with respect to the telomeres. The cryptomonad also contains a 5S rRNA gene linked to the other rRNA genes but it is transcribed from the opposite DNA strand (Zauner et al., 2000). To date no 5S rRNA gene has been found in chlorarachniophytes.

So why should the chromosome ends of genomes derived from two different sources have been so similarly sculptured by evolution? It is crucial that rRNA genes are maintained identical or nearly so to interact with a defined set of ribosomal proteins. Mechanisms that involve extensive recombination and mismatch repair preserve the same sequence in all copies of rDNA (Liao, 1999). In nucleomorphs recombination could maintain identity between different chromosome ends and in some eukaryotes subtelomeric regions are recombinational hot spots. For example, some parasites place protein genes responsible for evading the host's immune system in subtelomeric DNA where frequent recombination between different alleles can generate antigenic variability (Barry & McCulloch, 2001). Presumably recombination in nucleomorphs occurs during mitosis since meiosis is not known to occur. Recombination between subtelomeric DNA of sister chromatids occurs frequently in somatic human cells (Cornforth & Eberle, 2001) and recombination can help maintain the length of telomeres in cancerous cells which are deficient for the telomere synthesising

complex, telomerase (Dunham et al., 2000). It is interesting to note that single rRNA cistrons are also linked to the telomeres of every chromosome in the intracellular parasite, *Encephalitozoon cuniculi* and that these subtelomeric regions have high recombination activity (Brugère et al., 2000a, b). This microsporidian parasite has a small and probably secondarily reduced genome of 2.8 megabases and it will be interesting to discover if rDNA/telomere linkage is a general feature of reduced eukaryotic genomes.

Nucleomorphs: subjects of the smallest eukaryotic genome projects

Having established the basic layout of cryptomonad and chlorarachniophyte nucleomorph genomes alongside developing techniques for isolating and cloning nucleomorph DNA, it became practicable to determine the entire sequence of these genomes by shotgun sequencing approaches (McFadden et al., 1997b). Genes were identified and analysed with the program Magpie (Gaasterland & Sensen, 1996) and recently the complete nucleomorph genome sequence (551,264 bp) of *G. theta* was published (GenBank # NC_002752, NC_002753 and NC_002751 for chromosomes I, II and III, respectively) (Douglas et al., 2001). At the time of writing we have nearly finished *B. natans'* nucleomorph chromosome III and the other two larger chromosomes now partially sequenced should be completed by the end of the year. Comparative analyses of nucleomorph genome sequences will provide powerful means of understanding how eukaryote/eukaryote endosymbiosis and subsequent genome reduction occurs.

What do nucleomorphs encode?

It has been known for many years that the chloroplasts of plants only encode a fraction of the proteins that they contain (Abdallah, Salamini & Leister, 2000; The Arabidopsis Genome Initiative, 2000). The vast majority of chloroplast proteins are encoded by the plant nucleus and targeted back to the chloroplast after translation. In *Arabidopsis thaliana*, the only photosynthetic eukaryote for which a complete genome sequence is available (The Arabidopsis Genome Initiative, 2000), the nucleus encodes 1900–2500 proteins that are targeted to the chloroplast whereas the chloroplast chromosome encodes only 87 proteins (Abdallah, Salamini & Leister, 2000). Interestingly, only 35% of the nuclear encoded plastid proteins could be classified as being of cyanobacterial origin suggesting that eukaryotes have recruited or invented many other genes for plastid related functions (Abdallah, Salamini & Leister, 2000). It follows then that the nucleomorph, once the nucleus of a free living alga, should encode genes for plastid targeted proteins and indeed this has been the proposed *raison d'être* of this nucleus (McFadden et al., 1994). The recently completed nucleomorph genome of *G. theta* confirms this hypothesis but also adds a few surprising twists (Douglas et al., 2001). Although the nucleomorph encodes 531 genes, a mere 30 of these are for proteins destined to the plastid – this is even less than the number of genes encoded by the plastid genome itself (Douglas & Penny, 1999). Based on the gene density of regions of the *B. natans* nucleomorph already sequenced, we predict the genome will encode about 320 genes.

Plastid targeted nucleomorph proteins

Plastid targeted nucleomorph proteins still resemble the plastid targeted proteins found in plant nuclei (Douglas et al., 2001) in that they possess N-terminal extensions required for targeting to the plastid. These putative plastid targeting sequences, called transit peptides, appear shorter than normal transit peptides and are only sometimes predicted to act as targeting sequences by the transit peptide prediction program ChloroP (http://www.cbs.dtu.dk/services/ChloroP/) (Douglas et al., 2001). Despite this, cryptomonad plastids appear to follow standard plastid import rules. *In vivo* import assays with isolated pea or cryptomonad plastids and rubredoxin, a nucleomorph-encoded plastid protein, have shown that rubredoxin's N-terminal extension can act as a transit peptide that is cleaved after being imported (Wastl & Maier, 2000). The nucleomorph even encodes components of the plastid import machinery (Iap100 and Tic22) plus components for translocation into the thylakoid lumen (SecE) (Douglas et al., 2001).

The cryptomonad nucleomorph genome also encodes proteins required for gene expression (ribosomal proteins, and others for DNA and RNA metabolism), for plastid division (FtsZ) (Fraunholz, Moerschel & Maier, 1998), for protein folding (Cpn60, HcfI36) and protein degradation (ClpP). The nucleomorph also houses non-cyanobacterial proteins such as Iap100, Met and CbbX that were invented or

recruited by eukaryotes to operate in plastids (Maier et al., 2000). Eleven ORFs with homology to cyanobacterial genes are also present that might function in the plastid (Douglas et al., 2001). Future studies of these may reveal novel and as yet undiscovered aspects of plastid function. Interestingly, only two proteins encoded by the nucleomorph are directly involved in photosynthesis, the electron transfer molecule rubredoxin (Wastl et al., 2000; Zauner et al., 2000) and Hlip that binds carotene. It has been estimated that at least a thousand plastid protein genes must have been transferred out of the nucleomorph to the host cell nucleus (Douglas et al., 2001) but based the numbers of putative nuclear-encoded plastid proteins in higher plants (Abdallah et al., Salamini & Leister, 2000) this number could well be higher.

In the partially sequenced *B. natans* nucleomorph, we have so far identified several genes encoding putative plastid targeted proteins (three ClpPs, an ABC transporter and SecY) (Gilson & McFadden, unpublished) but the final number (perhaps 10–20 proteins) is not expected to be as great as *G. theta's* since the chlorarachniophyte nucleomorph genome is smaller.

Genetic housekeeping functions

If the small number of nucleomorph-encoded proteins destined for the plastid are essential, it follows then that the remaining nucleomorph genome is devoted to transcribing, translating and folding these essential proteins. In addition, the nucleomorph must be able to replicate its DNA, maintain a cell cycle and segregate its genome during mitosis (Douglas et al., 2001). A glance at the gene list of *G. theta's* nucleomorph confirms that it is dominated by genes encoding these so-called genetic housekeeping functions (Douglas et al., 2001).

Transcription

The extraordinary conversion of algal endosymbionts into complex plastids (organelles) has probably been occurring over the past 600 million years (Cavalier-Smith, 1995b). During this time it has been estimated that nucleomorphs have likely been miniaturised 125-fold (Beaton & Cavalier-Smith, 1999) and yet many basic functions are still remarkably conserved. For example, all three kinds of RNA polymerases are still present in the cryptomonad nucleomorph as well as proteins for promoter recognition and binding (Douglas et al., 2001). Some subunits of the three RNA polymerases have also been identified in *B. natans*' nucleomorph (Gilson & McFadden, 1996; Gilson & McFadden unpublished).

RNA metabolism

mRNAs in both nucleomorphs are polyadenlyated, and the *G. theta* genome encodes a poly A-binding protein (Pab) as well as enzymes for 5′ capping (capping enzyme Mce and cap-binding protein Cbp) (Douglas et al., 2001). Introns are present in nucleomorph genes and both genomes encode numerous spliceosomal components (Gilson & McFadden, 1996; Douglas et al., 2001), and the complete set of spliceosomal RNAs (U1, U2, U4, U5 and U6) are present in *G. theta* (Douglas et al., 2001). A nucleomorph U6 snRNA gene has been identified in *B. natans* and its transcripts probably assemble into spliceosomes because capped U6 transcripts have been detected in the nucleomorph by immunolocalization (Gilson & McFadden, 1996). The *G. theta* nucleomorph encodes not only rRNAs (srRNA, 5.8S, lrRNA and 5S) but also snoRNAs and 17 proteins of the nucleolar snoRNA machinery required for cleavage of primary rRNA transcripts and their base modification (methylation and pseudouridylation) (Douglas et al., 2001).

Early reports that the cryptomonad nucleomorph contained pores (Ludwig & Gibbs, 1985) are validated by presence of components of the nuclear pore-complex export/import machinery (importin genes *impA* and *imb1*) as well as the transport protein CRM (Douglas et al., 2001). However, many nuclear-pore complex proteins are missing, and are presumably imported from the host cell. Pores have also been observed in the chlorarachniophyte nucleomorph envelope (Hibberd & Norris, 1984) and it is anticipated that its genome will encode some pore complex genes as well. Nucleomorphs would therefore appear to have retained normal nuclear transport mechanisms for proteins and RNAs.

Translation

In addition to rRNAs, the *G. theta* nucleomorph encodes the majority of proteins typically present in a ribosome (37 large subunit and 28 small subunit proteins) (Douglas et al., 2001). However, ∼14 proteins expected to be present are missing from the nucleomorph gene list. Assuming that these proteins are required and are not encoded by any of the nucleomorph's numerous unidentified open reading frames (ORFs), they must be imported from the host cell. This discovery raises an important issue. Will these missing proteins be derived from former nucleomorph

proteins that were transferred to the host nucleus and then targeted back, or will they originate from duplicated host genes that acquired the correct targeting information to send replacement proteins to the periplastidal space? Both gene replacement scenarios occur in simple plastids and in mitochondria (Small et al., 1999).

The *G. theta* nucleomorph appears to contain standard translation initiation and elongation factors as well as a set of 37 tRNAs (Douglas et al., 2001). This set appears to be sufficient to supply all amino acids except for glutamine for which no matching tRNA was found (Douglas et al., 2001). Glutamyl-tRNA could be imported from the host cell or another tRNA could substitute by a modification of the wobble rules (Douglas et al., 2001). Data derived from the chlorarachniophyte nucleomorph have so far revealed several tRNAs and at least 41 ribosomal proteins (Gilson & McFadden, unpublished).

Interestingly, the only amino-acyl-tRNA synthetase that is present in the *G. theta* nucleomorph is specific for serine indicating that synthetases for the 19 other remaining amino acids must be imported from the host cell (Douglas et al., 2001). Intriguingly, seryl-tRNA synthetase is also the only synthetase that has been identified in the partially sequenced *B. natans* nucleomorph (Gilson & McFadden, unpublished). The endosymbiont's requirement for a large number of amino-acyl-tRNA synthetases hints that these genes were easily transferred to the host cell. Alternatively, the conserved nature of these enzymes and their substrates might mean that they could readily be replaced by host cell or even plastid equivalents once nucleomorph versions were mutationally inactivated and subsequently lost. Alternative tRNA synthetases are commonly used by plastids and mitochondria (Small et al., 1999).

Protein folding and degradation

The cornucopia of *G. theta* nucleomorph proteins responsible for protein folding and degradation tell a complex story. These proteins are not just involved in simple protein maturation and recycling but may also have roles in protein import from the host across the periplastid membrane (Hsp70), assembly and structure of the mitotic apparatus (Hsp70, Hsp90 and proteins of the CCT complex) as well as regulated turnover of cell cycle regulators (cyclin B) that may employ a ubiquitin-fusion-degradation system (Douglas et al., 2001). Ubiquitin is found fused to two ribosomal proteins and three E2 enzymes (ubiquitin conjugating) are present as well a many enzymes of the 20S core proteosome and its 19S cap (Douglas et al., 2001).

Co-chaperones that assist Hsp70, Hsp90 and CCT (Hsp40/dnaJ, hip, hop and prefoldin, respectively) are missing and either not necessary or are imported (Douglas et al., 2001). Another curious anomaly has been noted by Archibald et al. (2001). They found that the amino acid sequences of some chaperones such as Hsp70 and 90 are very conserved whilst the eight components of CCT are highly divergent. Functional constraints acting upon the heat shock proteins may be more strict than those upon CCT especially if the former work as efficient molecular ratchets to drag proteins across the periplastidal membrane. It is possible that selective pressures acting upon CCT are relaxed because the number of nucleomorph substrates that have to be folded are reduced, or the rate of protein folding is more leisurely.

The only chlorarachniophyte nucleomorph chaperone sequences available to date encode Hsp70, Hsp90 and three CCT subunits (expect eight) (Gilson & McFadden, unpublished data). The chlorarachniophyte Hsp70, unlike its cryptomonad orthologue, does not appear to possess any nuclear localisation signals (predicted by the program Psort http://psort.nibb.ac.jp/) and no heat shock regulatory elements (TTCnnGAAnnTTC) upstream of its coding sequence (Archibald et al., 2001). It will be of interest to determine if the chlorarachniophyte nucleomorph encodes a heat shock transcription factor (Hsf) like its cryptomonad counterpart and if heat and other environmental stresses can modify nucleomorph gene expression.

Mitosis

Although microscopic examination of nucleomorphs has failed to detect any trace of a mitotic spindle that could segregate the chromosomes during division, a spindle appears to be used based on genes present in the nucleomorph genome of *G. theta* (Keeling et al., 1999; Douglas et al., 2001). Spindle forming genes encoding α-, β- and γ-tubulin are present as well as proteins that might form the intranuclear centrosomes (Ranbpm, Hsp70 and Hsp90) (Douglas et al., 2001). Centromeres are still apparently used to attach nucleomorph chromosomes to the spindle because the nucleomorph encodes a histone-like centromere protein (Cenp-A). Gene free regions identified in the central regions of all cryptomonad nucleomorph chromosomes have been proposed to act as centromeric DNA (Douglas et al., 2001) although they do not ap-

pear to be similar to each other or contain any obvious head to tail repetitive elements typical of eukaryotic centromeres (Heslop-Harrison et al., 1999). Ranbpm and Cenp-A are present in the chlorarachniophyte nucleomorph indicating mitosis may also segregate these chromosomes but centromeres and tubulin genes are yet to be identified (Gilson & McFadden, unpublished).

DNA folding, replication and cell cycle control

The nucleomorph of *G. theta* encodes three core histones (H2b, H3, H4) (Douglas et al., 2001). Genes encoding H2a and H1 appear to be absent from the genome but the proteins must be imported since without H1 the chromosomes would not be able to condense into 30 nm filaments (Douglas et al., 2001). Histone acetylation appears to proceed as normal because histone acetylation sites are conserved and acetyltransferase (Hat) and deacetylase (Had) enzymes are encoded by the genome (Douglas et al., 2001). Telomeres, centromeres and other regions that are transcriptionally inactive could be the target for deacetylases (Grant, 2001).

How is nucleomorph DNA replicated? The cryptomonad nucleomorph genome does not encode DNA polymerases indicating the nucleomorph polymerase gene either resides in the host nucleus or has been replaced by a host version (Douglas et al., 2001). The nucleomorph however has not lost the ability to modify its own DNA since it encodes some of its own replication co-factor (Rfc) and a repair and recombination enzyme (Rad51) (Douglas et al., 2001). Another interesting finding of Douglas et al. (2001) was that the only substantial regions of genome that are gene-free, and therefore could serve as replication origins, lie within the inverted repeats on the chromosome ends. Eukaryotic DNA replicons are typically 100 kb in length and Douglas et al. (2001) propose that if a single replication origin lay in each terminal repeat of *G. theta's* nucleomorph chromosomes they would be about the right distance apart to replicate the chromosome. A similar phenomenon could occur in chromosomes of the chlorarachniophyte nucleomorph since the largest gene free regions occur in the inverted repeats as well (Gilson & McFadden, unpublished).

The cryptomonad nucleomorph still appears to encode some components of a cell cycle and mitotic control system (Douglas et al., 2001). It produces a cyclinB/cdc2 complex known to play a role in centrosome activation. Cell cycle control proteins made by the cryptomonad nucleomorph such as replication licensing protein (Mcm2) and a cyclin-dependent Cdc2 kinase (including cyclin B) are involved in G1 to S-phase transition and the G2-M-phase checkpoint, respectively. The retention of complex ubiquitin-based protein degradation pathways confirm the importance of regulated protein turnover, especially for the degradation of cell cycle regulators (e.g., cyclinB) in the nucleomorph (Douglas et al., 2001).

Non-genetic housekeeping functions: feeding the host

The *G. theta* nucleomorph thus appears able to express, replicate and divide its own genetic information content in a regulated manner. Despite being incredibly whittled down and relying on host proteins for some functions, it still remains 'conceptually equivalent to a cell' (Douglas et al., 2001). As discussed above, the function of this vestigial cell is to express a mere handful of mostly plastid proteins that perform 'end-product functions' useful to the rest of the cell (Cavalier-Smith & Beaton, 1999). Sequencing of the *G. theta* genome has not only confirmed earlier expectations that the largest end-product functional group are involved in plastid-related functions (discussed above) but also revealed a small number of proteins that are involved in other functions (Douglas et al., 2001).

The engulfed alga feeds the host cell photosynthetically manufactured carbohydrate. Cryptomonads still store their starch in a region of the cell equivalent to where red algae store their starch, the cytoplasm or periplastidal space of the endosymbiont. As the need in the host cell arises for more glucose it may signal the periplastidal space to degrade more starch for glucose export to the host cell. Douglas et al. (2001) have identified a number of nucleomorph enzymes potentially involved in this process.

So far sequencing of the *B. natans* nucleomorph genome has not revealed any non-plastid, end product functions (Gilson & McFadden, unpublished data). In chlorarachniophytes, the endosymbionts do not store starch, unlike their green algal ancestors that store it as chloroplast localised starch grains (Hibberd & Norris, 1984). The task of carbohydrate storage has been entirely assumed by the host cell that stores it as a β-1,3 glucan rather than a starch like α-1,4 glucan (McFadden, Gilson & Sims, 1997a). Interestingly, the β-1,3 glucan is retained within vesicles that are closely appressed to a specialised region of the plastid called the pyrenoid (McFadden, Gilson & Sims, 1997a). Evidently the host has moved its carbo-

hydrate storage system as close to the endosymbiont as possible, presumably to optimise uptake of glucose from the plastid/endosymbiont. We predict that chlorarachniophyte nucleomorphs will encode few, if any, carbohydrate metabolising enzymes and that this deficit compared to cryptomonads could be one factor accounting for the smaller genome sizes of chlorarachniophyte nucleomorphs.

Unidentified ORFs

Of 464 putative protein-coding genes in the nucleomorph of *G. theta*, 188 (37%) have no homologues in the database (Douglas et al., 2001). Six percent of ORFs are conserved (31 ORFs, see above) and 57% (245) have similarity to proteins of known function. Considering that most proteins encoded by the nucleomorph are involved in well-studied genetic housekeeping functions, the percentage of nucleomorph proteins with recognised homology seems quite low. There are probably at least three reasons for this.

(1) Organisms for which we have complete genome sequences often belong to groups with many genetically well-characterised relatives which tends to boost the number of genes with recognised homologues because they are often found in relatives. An example of this is the well-studied animal *Homo sapiens*, in which 74% of proteins have homologues in other organisms, many of them other well-studied animals (International Human Genome Sequencing Consortium, 2001). By comparison cryptomonad nucleomorphs are derived from red algae for which there is a relative paucity of gene data.
(2) Nucleomorphs contain comparatively fewer recognisable proteins because their protein sequences seem to be evolving quite rapidly (see below) and this may have reduced sequence conservation to levels that are no longer recognisable.
(3) The nucleomorph/periplastidal space is a highly specialised environment and many novel genes may have been created, or recruited from pre-existing genes, for nucleomorph-specific functions.

Nucleomorphs are extremely compact

At first glance one of the most striking aspects of nucleomorphs are the extraordinary compactness with which they organise their genetic information (Gilson, Maier & McFadden, 1997; Douglas et al., 2001). They are amongst the most compact genomes in nature with up to 91% of their DNA encoding genes (Beaton & Cavalier-Smith, 1999). In comparison, the genomes of vertebrates possess extremely slovenly housekeeping with only 1.5% encoding functional gene products in the case of humans (Gilson & McFadden, 2001).

Jam-packed genomes

The eukaryotic prize winner for gene density has to be *G. theta's* nucleomorph with 1 gene per 977 bp, higher than most bacteria (Douglas et al., 2001). Runner up is the chlorarachniophyte nucleomorph. Data from chromosome III of *B. natans* indicate gene density is 1 per 1141 bp (Gilson & McFadden, unpublished data). Gene density of the gene-rich single copy region of chromosome III is somewhat higher at 1 gene per 1007 bp and would probably be similar to the cryptomonad if the substantially higher intron density of the chlorarachniophyte nucleomorph were taken into account (Gilson & McFadden, 1996).

The lengths of spacer DNA between genes in both nucleomorphs that would normally house regulatory information for gene expression are often extremely short. The average spacer lengths between non-ribosomal RNA genes in the *B. natans'* nucleomorph is 97 bp (based on chromosome III, Gilson & McFadden, unpublished data) and although spacer calculations have not been done for the *G. theta* nucleomorph they are probably similarly sized. Interestingly, in *B. natans'* nucleomorph the average spacer length varies depending on the orientation of neighbouring genes to each other. Neighbouring genes that are transcribed in the same direction (head to tail) have an average spacer length of 101 bp (Gilson & McFadden, unpublished data). Those neighbours that are on opposite DNA strands and transcribed away from each other (head to head) have the longest spacers (average 116 bp) and those that are on opposites strands but are transcribed towards each other (tail to tail) have the shortest spacers (average 77 bp). These spacer lengths suggest that gene promoter regions occupy more space than terminators or are less refractory than terminators to adopting cryptic sequences within neighbouring coding sequences.

In nucleomorphs the elimination of spacer DNA between some genes has been taken to the extreme by its complete elimination. In some cases the coding sequences of genes even overlap! In *G. theta* the coding

sequences of 44 genes overlap, in one case by up to 76 nucleotides (nt) (Douglas et al., 2001). Overlapping genes have also been discovered in the chlorarachniophyte nucleomorph but so far the longest example is a mere 12 nt (Gilson & McFadden, unpublished). As noted for terminators above, most cases of overlapping genes are in the tail to tail orientation.

The paucity of regulatory information between genes in some cases appears to have forced the utilisation of elements within the coding sequences of bordering genes as promoter and terminators. Transcripts of cryptomonad genes often begin in an upstream gene and terminate in a downstream one (Douglas et al., 2001). In chlorarachniophytes this is taken to the extreme with several genes frequently found on a single mature mRNAs in both sense and antisense orientations (Gilson & McFadden, 1996). It is not known at this stage if all the coding sequences in polycistronic mRNAs can be translated or if they have to be processed into individual transcripts or even transcribed separately. Such transcription length inaccuracy would perhaps be disastrous for a normal nucleus but in the minimally complex genetic environment of the nucleomorph such inefficiency seems to be tolerated.

Extremely short introns

Nucleomorphs still retain the intronic hallmarks of eukaryotic genes, but like the rest of the genome they have become miniaturised during endosymbiosis. The so called pygmy-sized spliceosomal introns of the *B. natans* nucleomorph genome (Gilson & McFadden, 1996) are on average the smallest of any eukaryote ranging in size from 18 to 20 nt (chromosome III; 10% 18 nt, 68% 19 nt, 22% 20 nt). They are even smaller than the ciliate, *Paramecium tetraurelia* whose average spliceosomal intron is about 26 nt (Russell, Fraga & Hinrichsen, 1994) and much smaller than the typical eukaryotic range of 40–125 nt (Deutsch & Long, 1999). It seems likely that the introns of nucleomorphs, like the rest of the genome, have been secondarily reduced in size and were not always as small. Assuming the intron sizes within the green algal ancestor of the chlorarachniophyte endosymbiont were about the same size as *Arabidopsis* (average size 168 nt) (The Arabidopsis Genome Initiative, 2000) there has been a considerable reduction in intron size probably due to DNA loss.

The introns in *G. theta's* nucleomorph range in size from 42 to 52 nt (average 49 nt) (Douglas et al., 2001). Although genomic data from which we can estimate the average intron size of the red algal ancestors of cryptomonad endosymbionts is sparse, recent intron data derived from the genes of light-harvesting proteins from *Galdieria sulphuraria* suggest red algal introns might be surprisingly small (50–74 nt) (Marquardt et al., 2000). Perhaps cryptomonad nucleomorph introns have only undergone modest reduction due to limitations imposed by the efficacy of the splicing apparatus to remove them. It will be of great interest to compare components of chlorarachniophyte nucleomorph spliceosomes (once they are sequenced) with other eukaryotes to determine what changes have occurred to allow these spliceosomes to remove such small introns.

If accelerated DNA loss has created tiny introns then it might be expected to have removed many introns altogether. However, in the *B. natans* nucleomorph this does not appear to be the case since intron density still remains quite high with an average of 3.3 introns per kilobase of protein coding sequence. Although we do not know the intron density in the green algal relatives of the endosymbiont (Van de Peer et al., 1996; Ishida et al., 1997; Gilson & McFadden, 1999; Ishida, Green & Cavalier-Smith, 1999), intron density in *Arabidopsis* is similar (The Arabidopsis Genome Initiative, 2000) indicating that intron loss in chlorarachniophyte nucleomorphs has probably been minimal.

The entire *G. theta* nucleomorph only contains 17 spliceosomal introns, which is fewer than the number of spliceosomal components the nucleomorph encodes. Unfortunately, there is little data from which estimate intron loss in cryptomonad nucleomorphs due to the lack of red algal genomic data.

Muller's ratchet and mutational hyperdrive

The first hint that nucleomorph gene sequences might be highly derived and evolving rapidly came from phylogenetic analyses of their srRNAs that showed extremely long branch lengths compared to other eukaryotes (Cavalier-Smith et al., 1996; Van de Peer et al., 1996). Protein sequences are also highly diverged from orthologues in other organisms, as recent analysis of components of the nucleomorph CCT complex and tubulins of *G. theta* have shown (Keeling et al., 1999; Archibald et al., 2001). However, there is an important difference between genes encoding structural RNAs (e.g., rRNAs, tRNAs,

snRNAs) and those of proteins. A+T-richness of structural RNAs in both nucleomorphs ranges from 50 to 65%, whereas for many protein encoding genes A+T content is 65–80% (Gilson & McFadden, 1996; Zauner et al., 2000; Douglas et al., 2001). Although there is a propensity for genomes of nucleomorphs, plastids, mitochondria and bacterial endosymbionts (e.g., *Buchnera* endosymbionts in aphids (Moran, 1996)) to become enriched for A+T bases, functional RNAs are often constrained by factors such as base pairing for secondary structure formation.

Nucleomorph genes appear to be evolving rapidly due to a Muller's Ratchet effect where the fittest alleles are lost and less fit mutations accumulate due to genetic drift in small populations (Gilson & McFadden, 2001). The effective population size of nucleomorphs is extremely small because the opportunities for genetic exchange with other nucleomorphs are next to zero. Sexual life stages in cryptomonads and chlorarachniophytes that would allow for nucleomorphs from different cells to share a common cytoplasm and exchange DNA are not well characterised and maybe uncommon (Hill & Wetherbee, 1986; Grell, 1990). Even when host cells fuse, the ability of nucleomorphs to exchange DNA is likely to be compromised by the outer pair of membranes surrounding the nucleomorphs. Why the protein genes of many endosymbionts and organelles become A+T-rich as they drift is unknown but one consequence is that where possible, amino acids with A+T-rich codons have been selected for. For instance, nucleomorph CCT proteins of *G. theta* have highly biased amino acid compositions and are enriched for amino acids that utilise A+T-rich codons (e.g., asparagine that uses AAT and AAC) (Archibald et al., 2001).

Gene spacers and introns are more A+T-rich (up to 90%) than protein coding gene regions because of reduced constraints (Gilson & McFadden, 1996; Palmer & Delwiche, 1996). Curiously, gene spacers in the inverted repeats on the chromosome ends are not especially A+T-rich (about 55%) (Douglas et al., 2001). Even though some of this spacer DNA is part of the pre-rRNA gene transcript containing srRNA, 5.8S and lrRNAs and is potentially constrained by rRNA processing requirements (folding and interaction with snoRNAs), other spacer DNA is not. Perhaps frequent recombination between chromosome ends prevents a build up of A+T substitutions or maybe the putative roles for these regions as replication origins enforce a more balanced A+T content.

Nucleomorphs and the C-value enigma

Although the total number of genes in free-living eukaryotic genomes only varies about 10-fold, total amounts of nuclear DNA can vary up to 200,000-fold (Gregory, 2001). The bulk of this size difference is due to variation in the levels of non-coding DNA, the so called C-value enigma (sometimes but less accurately referred to as the C-value paradox). No single factor appears to determine genome size and it is likely a complex interplay between many factors acting at different levels of biological organisation are responsible (Gregory, 2001). What factors may be responsible for moulding the genome sizes and densities of nucleomorphs?

It has been widely recognised that cell volume roughly scales with nuclear volume and therefore genome size, that is, big cells have big genomes (Gregory, 2000). Cavalier-Smith and Beaton (Beaton & Cavalier-Smith, 1999; Cavalier-Smith & Beaton, 1999) have found this to be true for the host-cell nuclei of different cryptomonad species but not for their nucleomorphs whose DNA contents appear similar despite large variation in cell volumes. Although, it is not known why or what maintains nuclear to cytoplasmic volume ratios (but for discussion see Gregory, 2001) it has been suggested that nucleomorphs have been released from their relationship with cell size because the host cell nucleus fulfils this role (Beaton & Cavalier-Smith, 1999; Cavalier-Smith & Beaton, 1999). We believe there is an alternative explanation. Nucleomorphs do not scale with the rest of the cell because they are separated from it by membranes. Both the host cell and the plastid are isolated from the nucleomorph and the periplastidal compartment by membrane barriers. Unlike nuclear pores that permit the free flow of small molecules and proteins in and out of the membrane-bound nucleus, the membranes that separate the nucleomorph/periplastidal space from the plastid lumen and the host cell cytoplasm are highly selective and this possibly abrogates the need for nucleomorph genome size to scale with the rest of the cell. If the continuity of small molecules between the cytoplasm and nucleus is an important factor in maintaining constant nuclear:cytoplasmic volumetric ratios then we would anticipate that nucleomorphs should scale with the small volumes of periplastidal cytoplasm surrounding them since nucleomorphs still have nuclear pores.

There are several reasons why nucleomorph genome sizes are so small and why their gene

organisation is extremely compact. If there were positive selection for nucleomorphs to replicate their chromosomes quickly then genome sizes might have shrunk. It has been suggested that competition between the replication rates of the chromosomes in plastids and mitochondria may have played a significant role in reducing amounts of non-coding DNA (Selosse et al., 2001). In these organelles, chromosomal genomes are usually multicopy and versions that contain less DNA could replicate faster than larger ones and eventually dominate the population. Selection for reduced genome sizes in organelles may also have hastened replacement of their genes by nuclear-encoded versions (Selosse, Albert & Godelle, 2001). Since nucleomorph chromosomes are diploid and are not multicopy, replication rate competition does not seem likely to be a significant factor capable of reducing genome sizes. However, it has been recently proposed that if rates of nucleomorph replication were comparatively slow and there was competition for rapid cell growth within a population then there may have been selection for a reduction in genome size (Selosse, Albert & Godelle, 2001). This proposal deserves further attention given that the restriction of replication origins to the ends of each nucleomorph chromosome or the need to import DNA polymerases from the host cell might reduce chromosomal replication rates and hence cell growth.

Alternatively, there may have been no selection for genome size reduction *per se*, rather the rates of DNA loss in nucleomorphs may simply be high. Reasons for elevated levels of DNA loses are unknown but high rates of DNA loss have been discovered in many other eukaryotes such as *Drosophila* (Petrov, Lozovskaya & Hartl, 1996). If the rates and/or sizes of deletions in nucleomorphs were particularly large and/or insertions were rare/short then these genomes could easily have been trimmed to their present sizes over the 600 million years or so they have resided within endosymbionts.

For a genome's size to remain constant, losses of old DNA must be balanced by additions of new DNA. Genomes can gain DNA by several means among which the activity transposable elements (TEs) is very important. Much of the non-coding DNA in some eukaryotes is comprised of transposable elements or their decaying sequences (International Human Genome Sequencing Consortium, 2001). Considering the apparent ubiquity of TEs in free-living organisms (Arkhipova & Meselson, 2000) it was surprising then that preliminary searches of nucleomorph DNA for transposable elements have not revealed any candidates (Gilson & Petrov, unpublished). No ORFs with apparent similarity to transposases or reverse transcriptases are recognised. TEs are known to flourish in sexually reproducing populations where presumably their introduction into new populations allows them to proliferate (Arkhipova & Meselson, 2000). A lack of sexual DNA recombination in nucleomorphs may, therefore, have acted as a barrier to the introduction of new TEs. The extra membranes surrounding nucleomorphs would also have obstructed the horizontal transmission of TEs from the host (random sequencing of DNA has revealed retrotransposon-like sequences reside in the nuclear genome, Gilson & McFadden, unpublished). Without DNA exchange to introduce new TEs into nucleomorphs, those that were originally there have probably been mutationally inactivated since there was no selection to maintain them.

In summary, nucleomorph genomes have been miniaturised by several processes. Gene sequences have been lost because they were no longer required or were replaced by versions encoded by the host cell nucleus. DNA was removed by high levels of deletion perhaps driven by selection for faster genome replication. Amplifying the rates of genome reduction have been the inactivation and removal of TEs that are important generators of new DNA.

Concluding remarks

Although, only about half the nucleomorph genome of *B. natans* has been sequenced and can be compared to the fully sequenced nucleomorph of *G. theta* a number of useful conclusions can be drawn. The most important of these is that both genomes despite being derived from very different ancestral algae have been honed towards very similar end points and are remarkable examples of convergent evolution at a genomic scale. Both genomes are similarly sized and are constrained to three chromosomes due to the upper and lower limits placed upon the length of their histone-bound chromosomes. The bulk of genes encoded by both nuclei are involved in genetic housekeeping functions whose purpose is to express a relatively small number of proteins that are useful to the rest of the cell, namely plastid-related functions. Gene organisation within the two nucleomorphs is remarkably similar with rDNA repeats bordering compactly organised, single copy genes encoding proteins, tRNAs and snRNAs.

Completion of the chlorarachniophyte nucleomorph will hopefully reveal even more surprises.

Comparison of the fully sequenced nucleomorph to its cryptomonad counterpart may reveal the minimal set of genetic housekeeping genes required to make a eukaryotic nucleus and express its information content (although we note that some nucleomorph genes must be imported from the host cell). Furthermore, a common set of genes encoding plastid proteins that are recalcitrant for transfer to host cell may also be identified in both genomes. Such genes have been identified in plastid and mitochondrial genomes (Race, Herrmann & Martin, 1999).

Acknowledgements

We thank the Australian Research Council for support, the anonymous referees for many useful comments and suggestions, and Vanessa Su for sequence data.

References

Abdallah, F., F. Salamini & D. Leister, 2000. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. Trends Plant Sci. 5: 141–142.

Archibald, J., T. Cavalier-Smith, U. Maier & S. Douglas, 2001. Molecular chaparones encoded by a reduced nucleus – the cryptomonad nucleomorph. Mol. Biol. Evol. 52: 490–501.

Arkhipova, I. & M. Meselson, 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. USA 97: 14473–14477.

Barry, J. & R. McCulloch, 2001. Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. Adv. Parasitol. 49: 1–70.

Beaton, M. & T. Cavalier-Smith, 1999. Eukaryotic non-coding DNA is functional: evidence from the differential scaling of cryptomonad genomes. Proc. R. Soc. London 266: 2053–2059.

Brugère, J., E. Cornillot, G. Metenier & C. Vivares, 2000a. Occurence of subtelomeric rearrangements in the genome of the microsporidian parasite *Encephalitozoon cuniculi*, as revealed by a new fingerprinting procedure based on two-dimensional pulsed field gel electrophoresis. Electrophoresis 21: 2576–2581.

Brugère, J.-F., E. Cornillot, G. Mètènier, A. Bensimon & C. Vivarès, 2000b. *Encephalitozoon cuniculi* (Microspora) genome: physical map and evidence for telomere-associated rDNA units on all chromosomes. Nucl. Acid Res. 28: 2026–2033.

Cavalier-Smith, T., 1995a. Membrane heredity, symbiogenesis, and the multiple origins of algae, pp. 75–114 in Biodiversity and Evolution, edited by R. Arai, M. Kato & Y. Doi. National Science Museum, Tokyo.

Cavalier-Smith, T., 1995b. Membrane heredity, symbogenesis, and the multiple origins of the algae., pp. 75–114 in Biodiversity and Evolution, edited by R. Arai, M. Kato & Y. Doi. The National Science Foundation, Tokyo.

Cavalier-Smith, T., J.A. Couch, K.E. Thorsteinsen, P.R. Gilson, J.A. Deane, D.R.A. Hill & G.I. McFadden, 1996. Cryptomonad nuclear and nucleomorph 18S rRNA phylogeny. Eur. J. Phycol. 31: 315–328.

Cavalier-Smith, T. & E.E. Chao, 1997. Sarcomonad ribosomal RNA sequences, rhizopod phylogeny, and the origin of euglyphid amoebae. Arch. Protistenk. 147: 227–236.

Cavalier-Smith, T., 1998. A revised six-kingdom system of life. Biol. Rev. Camb. Phil. Soc. 73: 227–236.

Cavalier-Smith, T., 1999. Priciples of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryotic family tree. J. Euk. Microbiol. 47: 347–366.

Cavalier-Smith, T. & M. Beaton, 1999. The skeletal function of non-genic nuclear DNA: new evidence from ancient cell chimeras. Genetica 106: 3–13.

Cavalier-Smith, T., 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5: 174–182.

Cornforth, M. & R. Eberle, 2001. Termini of human chromosomes display elevated rates of mitotic recombination. Mutagenesis 16: 85–89.

Deutsch, M. & M. Long, 1999. Intron-exon structures of eukaryotic model organisms. Nucl. Acids Res. 27: 3219–3228.

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

Douglas, S., S. Zauner, M. Fraunholz, M. Beaton, S. Penny, L.-T. Deng, X. Wu, M. Reith, T. Cavalier-Smith & U.-G. Maier, 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410: 1091–1096.

Douglas, S.E., C.A. Murphy, D.F. Spencer & M.W. Gray, 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature 350: 148–151.

Douglas, S.E. & S.L. Penny, 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J. Mol. Evol. 48: 236–244.

Dunham, M., A. Neumann, C. Fasching & R. Reddel, 2000. Telomere maintenance by recombination in human cells. Nat. Genet. 26: 447–450.

Eschbach, S., C.J.B. Hofmann, U.-G. Maier, P. Sitte & P. Hansmann, 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. Nucl. Acids Res. 19: 1779–1781.

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

Fraunholz, M.J., E. Moerschel & U.G. Maier, 1998. The chloroplast division protein FtsZ is encoded by a nucleomorph gene in cryptomonads. Mol. Gen. Genet. 260: 207–211.

Gaasterland, T. & C.W. Sensen, 1996. Fully automated genome analysis that reflects user needs and preferences: a detailed introduction to the MAGPIE system architecture. Biochimie 78: 302–310.

Gibbs, S., 1981. The chloroplasts of some algal groups may have evolved from some endosymbiotic eukaryotic algae. Ann. NY Acad. Sci. 361: 193–208.

Gilson, P. & G. McFadden, 1999. Molecular and morphological characterization of six chlorarachniophyte strains. Phycol. Res. 47: 7–19.

Gilson, P. & G. McFadden, 2001. A grin without a cat. Nature 410: 1040–1041.

Gilson, P.R. & G.I. McFadden, 1995. The chlorarachniophyte: a cell with two different nuclei and two different telomeres. Chromosoma 103: 635–641.

Gilson, P.R. & G.I. McFadden, 1996. The miniaturised nuclear genome of a eukaryotic endosymbiont contains genes that overlap,

genes that are contranscribed, and smallest known spliceosomal introns. Proc. Natl. Acad. Sci. USA 93: 7737–7742.

Gilson, P.R., U.G. Maier & G.I. McFadden, 1997. Size isn't everything – lessons in genetic miniaturisation from nucleomorphs. Curr. Opin. Genet. Dev. 7: 800–806.

Grant, P., 2001. A tale of histone modifications. Genome Biol. 2: 3.1–3.6.

Greenwood, A., 1974. The Cryptophyta in relation to phylogeny and photosynthesis, pp. 566–567 in 8th International Congress of Electron Microscopy, edited by J. Sanders & D. Goodchild. Australian Academy of Sciences, Canberra.

Greenwood, A., H. Griffiths & U. Santore, 1977. Chloroplasts and cell compartments in Cryptophyceae. Br. Phycol. J. 12: 119.

Gregory, T., 2000. Nucleotypic effects without nuclei: genome size and erythrocyte size in mammals. Genome 43: 895–901.

Gregory, T., 2001. Coincidence, coevolution, or causation? DNA content, cell size and the C-value enigma. Biol. Rev. 76: 65–101.

Grell, K., 1990. Indications of sexual reproduction in the plasmodial protist *Chlorarachnion reptans* Geitler. Z. Naturforsch. 45c: 112–114.

Hansmann, P., H. Falk, U. Sheer & P. Sitte, 1986. Ultrastructural localization of DNA in two Cryptomonad species by use of a monoclonal DNA antibody. Eur. J. Cell Biol. 42: 152–160.

Hansmann, P. & S. Eschbach, 1990. Isolation and preliminary characterization of the nucleus and the nucleomorph of a cryptomonad, *Pyrenomonas salina*. Eur. J. Cell Biol. 52: 373–378.

Heslop-Harrison, J., M. Murata, Y. Ogura, T. Schwarzacher & F. Motoyoshi, 1999. Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis* chromosomes. Plant Cell 11: 31–42.

Hibberd, D.J. & R.E. Norris, 1984. Cytology and ultrastructure of *Chlorarachnion reptans* (*Chlorarachniophyta Divisio Nova, Chlorarachniophyceae Classis Nova*). J. Phycol. 20: 310–330.

Hill, D.R.A. & R. Wetherbee, 1986. *Proteomonas sulcata* gen. et sp. nov. (Cryptophyceae), a cryptomonad with two morphologically distinct and alternating forms. Phycologia 25: 521–543.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Ishida, K., Y. Cao, M. Hasegawa, N. Okada & Y. Hara, 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., B. Green & T. Cavalier-Smith, 1999. Diversification of a chimeric algal group, the Chlorarachniophytes: phylogeny of nuclear and nucleomorph small-subunit rRNA genes. Mol. Biol. Evol. 16: 321–331.

Keeling, P., 2001. Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home? Mol. Biol. Evol. 18: 1551–1557.

Keeling, P.J., J.A. Deane, C. Hink-Schauer, S.E. Douglas, U.-G. Maier & G.I. McFadden, 1999. The secondary endosymbiont of the cryptomonad *Guillardia theta* contains alpha-, beta-, and gamma-tubulin genes. Mol. Biol. Evol. 16: 1308–1313.

Liao, D., 1999. Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. 64: 24–30.

Ludwig, M. & S. Gibbs, 1985. DNA is present in the nucleomorph of cryptomonads: further evidence that the chloroplast evolved from a eukaryotic endosymbiont. Protoplasma 127: 9–20.

Maier, U., M. Fraunholz, S. Zauner, S. Penny & S. Douglas, 2000. A nucleomorph-encoded CbbX and the phylogeny of RuBisCo regulators. Mol. Biol. Evol. 17: 576–583.

Maier, U.-G., C. Hofmann, S. Eschbach, J. Wolters & G. Igloi, 1991. Demonstration of nucleomorph-encoded eukaryotic small subunit RNA in Cryptomonads. Mol. Gen. Genet. 230: 155–160.

Margulis, L. & M. Chapman, 1998. Endosymbioses: cyclical and permanent in evolution. Trends Microbiol. 6: 342–346.

Marquardt, J., S. Wans, E. Rhiel, A. Randolf & W. Krumbein, 2000. Intron-exon structure and gene copy number of a gene encoding for a membrane- intrinsic light-harvesting polypeptide of the red alga *Galdieria sulphuraria*. Gene 255: 257–265.

Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa & K. Kowallik, 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393: 162–165.

McFadden, G., P. Gilson & I. Sims, 1997a. Preliminary characterization of carbohydrate stores from chlorarachniophytes (Division: Chlorarachniophyta). Phycol. Res. 45: 145–151.

McFadden, G.I., P.R. Gilson, C.J. Hofmann, G.J. Adcock & U.-G. Maier, 1994. Evidence that an amoeba acquired a chloroplast by retaining part of an engulfed eukaryotic alga. Proc. Natl. Acad. Sci. USA 91: 3690–3694.

McFadden, G.I., P.R. Gilson, S.E. Douglas, C.J.B. Hofmann & U.-G. Maier, 1997b. Bonsai genomics: sequencing the smallest eukaryotic genomes. Trends Genet. 13: 46–49.

McFadden, G.I. & D.S. Roos, 1999. Apicomplexan plastids as drug targets. Trends Microbiol. 6: 328–333.

Moestrup, Ø. & M. Sengco, 2001. Ultrastructural studies on *Bigelowiella natans*, gen. et sp. nov., a chlorarachniophyte flagellate. J. Phycol. 37: 624–646.

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA 93: 2873–2878.

Moreira, D., H. LeGuyader & H. Philipe, 2000. The origin of red algae: implications for the evolution of chloroplasts. Nature 405: 69–72.

Murray, A. & J. Szostak, 1985. Chromosome segregation in mitosis and meiosis. Annu. Rev. Cell Biol. 1: 289–315.

Palmer, J.D. & C.F. Delwiche, 1996. Second-hand chloroplasts and the case of the disappearing nucleus. Proc. Natl. Acad. Sci. USA 93: 7432–7435.

Palmer, J.D. & C.F. Delwiche, 1998. The origin and evolution of plastids and their genomes, pp. 375–409 in Molecular Systematics of Plants II, edited by D.E. Soltis, P.S. Soltis & J.J. Doyle. Chapman Hall, New York.

Petrov D.A., E.R. Lozovskaya & D.L. Hartl. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature 384: 346–349.

Race, H., R. Herrmann & W. Martin, 1999. Why have organelles retained genomes? Trends Genet. 15: 364–370.

Rensing, S., M. Goddemeier, C. Hofmann & U.-G. Maier, 1994. The presence of a nucleomorph *hsp*70 gene is a common feature of Cryptophyta and Chlororacnhiophyta. Curr. Genet. 26: 451–455.

Russell, C.B., D. Fraga & R.D. Hinrichsen, 1994. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. Nucl. Acids Res. 22: 1221–1225.

Selosse, M., B. Albert & B. Godelle, 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. 16: 135–141.

Small, I., K. Akashi, A. Chapron, A. Dietrich, A.-M. Duchene, D. Lancelin, L. Maréchal-Drouard, B. Menand, H. Mireau, Y. Moudden, J. Ovesna, N. Peeters, W. Sakamoto, G. Souciet & H. Wintz, 1999. The strange evolutionary history of plant mitochondrial tRNAs and their aminoacyl-tRNA synthetases. J. Hered. 90: 333–337.

Sulli, C., Z.W. Fang, U. Muchal & S.D. Schwartzbach, 1999. Topology of *Euglena* chloroplast protein precursors within the

endoplasmic reticulum to Golgi to chloroplast transport vesicles. J. Biol. Chem. 274: 457–463.

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

Van de Peer, Y., S.A. Rensing, U.-G. Maier & R. de Wachter, 1996. Substitution rate calibration of small subunit rRNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc. Natl. Acad. Sci. USA 93: 7732–7736.

van Dooren, G.G., R.F. Waller, K.A. Joiner, D.S. Roos & G.I. McFadden, 2000. Protein transport in *Plasmodium falciparum*: traffic jams. Parasitol. Today 16: 421–427.

Waller, R.F., P.J. Keeling, R.G.K. Donald, B. Striepen, E. Handman, N. Lang-Unnasch, A.F. Cowman, G.S. Besra, D.S. Roos & G.I. McFadden, 1998. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA 95: 12352–12357.

Wastl, J. & U.-G. Maier, 2000. Transport of proteins into cryptomonads complex plastids. J. Biol. Chem. 275: 23194–23198.

Wastl, J., H. Sticht, U.G. Maier, P. Rosch & S. Hoffmann, 2000. Identification and characterization of a eukaryotically encoded rubredoxin in a cryptomonad alga. FEBS Lett. 471: 191–196.

Whatley, J., 1981. Chloroplast evolution – ancient and modern. Ann. NY Acad. Sci. 361: 154–165.

Wolters, J., 1991. The troublesome parasites: molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. Biosystems 25: 75–84.

Zauner, S., M. Fraunholz, J. Wastl, S. Penny, M. Beaton, T. Cavalier-Smith, U.-G. Maier & S. Douglas, 2000. Chloropast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. Proc. Natl. Acad. Sci. USA 97: 200–205.