J. Eukaryot. Microbiol., 55(6), 2008 pp. 530–535

DOI: 10.1111/j.1550-7408.2008.00357.x

Nuclear Genome Sequence Survey of the Dinoflagellate *Heterocapsa triquetra*

MICHELLE McEWAN, RAHEEL HUMAYUN, CLAUDIO H. SLAMOVITS and PATRICK J. KEELING

Department of Botany, Canadian Institute for Advanced Research, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

ABSTRACT. Dinoflagellates have among the largest nuclear genomes known, but we know little about their contents or organisation. Given the interest in dinoflagellate ecology, cell biology, and evolutionary biology, there are many reasons to thoroughly investigate the contents of dinoflagellate genomes, but because of their large size the only thorough samples to date have relied on expressed sequence tag surveys to analyse cDNAs. To complement this, there are some studies of the physical properties of dinoflagellate chromosomes, but no direct survey of the nature of the sequences contained within them. To start to build a picture of the contents of these genomes, we have sequenced over 230,000 bp from the nuclear genome of *Heterocapsa triquetra*, which has been estimated to be 18–23 billion base pairs in total. The survey includes one putative gene with two relict spliced leaders, one putative pseudogene, and a small number of low-complexity repeats, transposons, and other putative selfish elements, all of which account for about 5% of the survey. Another 5% of the survey was long, complex repeats, some highly represented. By far the greatest fraction of the survey (89.5%) is made up of non-repeated sequence with no similarity to any other known sequence.

Key Words. Chromosome, dinokaryon, junk DNA, nucleus, repeat, spliced leader, transposon.

THERE is nearly a million-fold difference in size between the largest and smallest known eukaryotic nuclear genomes (Cavalier-Smith 2005). The human genome, which is often regarded as very large, is actually nearer to the mid-point of the spectrum of genome size: the smallest are a fraction of the size of many bacterial genomes whereas the largest are hundreds of times the size of the human genome. Various reasons have been suggested to explain why genome size varies so greatly between species. Early ideas focused on the lack of correlation between the inferred ''complexity'' of an organism and its genome size, the so-called ''C-value paradox.'' It is now clear that genome size does not correlate with our interpretation of cellular complexity, and other possible correlations have been suggested instead, including effective population size and cell size (Cavalier-Smith 2005; Gregory and Witt 2008; Kapraun 2005; Lynch and Conery 2003; Vinogradov 2004). These correlations appear to best hold either for subsets of the species examined or in species with moderate genome sizes, but there are always exceptions and the organisms with extremely large and small genomes appear to defy any correlations. The extremely small genomes have attracted far greater attention, in particular the so-called hyper-compacted genomes of microsporidian parasites and nucleomorphs (Keeling and Slamovits 2005). While these genomes are particularly interesting due to the unusual pressures under which they operate, they are also better studied simply because it is considerably easier to use sequencing to examine the properties of a small genome than a large one. Nevertheless, extremely large genomes also present interesting problems regarding how they came to be the way they are and how they cope with any functional stresses that relate to their large size. Unfortunately, however, very little is known about the largest nuclear genomes, and no sequence survey is available for any examples at present.

One group with particularly interesting large genomes is the dinoflagellates. Not only are dinoflagellate nuclear genomes among the largest known, they also possess a number of unique organizational features that are likely related to this expansiveness. The dinoflagellate nucleus, called the dinokaryon, lacks histones and instead contains a DNA-binding protein, perhaps related to a DNA-binding protein in bacteria (Kasinsky et al. 2001; Rizzo 1991, 2003). The chromosomes are condensed throughout the cell cycle and are segregated through an unusual mechanism of closed mitosis with an extranuclear spindle that does not pass through the nuclear envelope (Triemer and Fritz 1984). Gene expression seems to occur on loops of DNA that emerge from the central core of the chromosome, composed of nested arches of transcriptionally inactive DNA (Costas and Goyanes 2005; Sigee 1983). The physical properties of the genome of one species, *Crypthecodinium cohnii* have been examined using hydroxylapatite binding, digestion with S1 nuclease and restriction enzymes, renaturation kinetics, and electron microscopy (Allen et al. 1975; Hinnebusch et al. 1980; Moreau et al. 1998). The overall picture from these studies is that about half the genome is made up of repeated sequences, at least one species of which is present in high copy numbers. The DNA is also rich in 5-hydroxymethyluracil (Herzog, Soyer, and Daney de Marcillac 1982; Rae 1976), another indication that much of the genome may be structural in nature. Overall, the dinokaryon is so abnormal that it was once considered a possible missing link between prokaryotes and eukaryotes (Allen et al. 1975; Dodge 1966; Raikov 1995), but the abundant phylogenetic evidence placing dinoflagellates within the alveolates (Keeling et al. 2005) now shows it is in fact a highly derived feature of dinoflagellates (Fast et al. 2002; Saldarriaga et al. 2003).

Genome-wide analyses have now been carried out on both the mitochondria and plastids from several dinoflagellates (Barbrook et al. 2001; Jackson et al. 2007; Nash et al. 2007; Slamovits et al. 2007; Zhang, Green, and Cavalier-Smith 1999), but all large-scale examinations of the nuclear genome have to date been limited to the genes encoded there due to the intractability of a whole genome sequence or survey. Expressed sequence tag (EST) projects have been carried out in several dinoflagellates (Bachvaroff et al. 2004; Hackett et al. 2004; Nosenko and Bhattacharya 2007; Patron et al. 2005; Patron, Waller, and Keeling 2006), providing a great deal of information about the evolution of individual genes and classes of genes, but by design do not address the nature of most of the genome. Accordingly, between our understanding of the physical properties of these genomes and the sequences of their genes lies a large gap in our knowledge about the overall nature of dinoflagellate genome sequences.

Here, we describe a low-redundancy genome sequence survey (GSS) from *Heterocapsa triquetra*, one of the better-studied dinoflagellates at the molecular level (Jackson et al. 2007; Patron et al. 2005; Waller, Patron, and Keeling 2006; Zhang et al. 1999; Zhang, Green, and Cavalier-Smith, 2000). This survey is by no means intended to provide deep coverage of the genome, but rather to provide a random sample of the nature of the sequences encoded there, and what we might anticipate from a large-scale sequencing project. As expected, we find few genes, but several gene fragments. There is also ample evidence of transposons and other selfish elements as well as low complexity sequence and

Corresponding Author: P. J. Keeling, Department of Botany, Canadian Institute for Advanced Research, University of British Columbia, 6270 University Blvd, Vancouver, BC, Canada V6T 1Z4—Telephone number: +1 604 822 4906; FAX number: +1 604 822 6089; e-mail: pkeeling@interchange.ubc.ca

micro-repeats, but again none of these forms a large portion of the sequence. Instead, the majority of the sequence is apparently random, non-repetitive sequence.

MATERIALS AND METHODS

Strain and culture conditions. *Heterocapsa triquetra* strain CCMP449 was acquired from the Culture Collection of Marine Plankton (West Boothbay Harbor, ME) and grown axenically in f/2–Si medium at 18 °C with a 16/8-h light/dark cycle. Cultures (250 ml) were grown to high density and used to inoculate 4-L cultures, which were subcultured every 10–20 days after checking for purity by light microscopy. Cells were harvested from successive 4-L cultures by centrifugation and disrupted by grinding under liquid nitrogen. DNA was purified using the DNeasy Plant DNA isolation kit (Qiagen, Mississauga, ON).

Library construction and sequencing. The individual DNA isolates with the highest concentrations and purity according to spectrophotometer readings were pooled for a total of 25 µg of total DNA. Total DNA was sheared by nebulization using the pCR 4-Blunt TOPO Shotgun Cloning kit (Invitrogen, Carlsbas, CA) for 40 s at 69 kPa. Sheared DNA was separated on a 1% agarose gel, and fragments ranging from 1 to 4 kbp were purified, concentrated, and cloned using the Invitrogen pCR4 according to the manufacturer's directions. Fragments were dephosphorylated before cloning to select against the insertion of two fragments into a single clone (because two dephosphorylated inserts cannot be joined to one another by the topoisomerase). This procedure does not rule out the possibility that sequence fragments may have sustained deletions or rearrangements after cloning. Similarly, there is no way to detect a bias in the cloning (e.g. favouring or disfavouring gene sequences over non-coding sequences), but there is no evidence from the overall composition of the survey for such a bias. The entire ligation was used to transform *Escherichia coli* in 10 individual transformations, resulting in approximately 2,200 insert-containing CFUs. Five-hundred and fifty colonies were screened by PCR for insert size, resulting in 281 clones with insert size >1 kbp. Clones with inserts up to 1.3 kbp were sequenced in one direction, while those with larger inserts (35 clones) were sequenced in both directions, and where needed, completed by primer walking.

Sequence analysis. Sequences were trimmed for low quality manually and sequences derived from a single clone were assembled using Sequencher 4.2 (GeneCodes, Ann Arbor, MI). Sequences derived from different clones were not assembled because the genomic coverage of the survey was so low that we expect identical sequences to be due to genomic repeats rather than overlapping coverage. The one exception to this was two clones that were identical in sequence and sharing the same ends, which are most likely identical clones, so one of these was discarded from subsequent analysis (only one case of this was found).

Sequences were compared with public databases using BLASTN and BLASTX, including dbEST, which holds the *H. triquetra* EST survey. All putative open reading frames (ORFs) with any similarity to genes in public databases were examined manually to identify potential start and end points and to determine the likely phylogenetic origin of the gene. Sequences repeated within the survey were identified from BLASTN, and tandem and micro-repeats were also identified using RepeatMasker 3.0 (Smit, Hubley, and Green 2004). Transposons were identified and classified using RepeatMasker, and other potentially transposon/viral-derived protein-coding genes were identified using BLASTX. The GC/AT content of all individual sequences was determined using GeeCee (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=geecee). To compare this with coding sequences, a random sample consisting of the same number of sequences from the *H. triquetra* EST survey (Patron et al. 2005) was selected and analysed in the same way.

RESULTS

Sequence complexity and overall characteristics of the *Heterocapsa triquetra* genome. Screening 550 colonies for insert size revealed 281 with inserts apparently >1 kbp. Those estimated to be <1.3 kbp were sequenced from one end while those larger were completely sequenced by primer walking, in total yielding 233,046 bp of sequence on 216 fragments. The genome size of *H. triquetra* has been estimated using flow cytometry to be between 18.6 and 23.6 billion base pairs (LaJeunesse et al. 2005; Veldhuis, Cucci, and Sieracki 1997); so our sample represents only a fraction of a percent of the whole genome. Below, we summarise the tendencies of this sample, and stress that caution must be used in interpreting such a small fraction of the genome. Moreover, although we observed no evidence that cloning was biased in favour of certain fractions of the genome over others (e.g. coding vs. non-coding, or heavily modified vs. unmodified), we cannot exclude the possibility that the survey is skewed due to a cloning bias. Such a bias must at least be relatively slight, because our findings do not diverge much from what was expected based on physical measurements of other dinoflagellate genomes.

The average GC content over the entire sample is 54.18%, as opposed to 61.75% estimated from 1,816,929 bp of EST data (excluding poly-A tails). Calculating GC% for each individual clone and comparing these with an equivalent number of randomly chosen cDNAs from the *H. triquetra* EST survey (Patron et al. 2005) shows the range of GC% of the genomic DNA is not only lower but also much broader than that of the cDNAs, ranging from 25% to 72% as opposed to the 35% to 74% for cDNAs (Fig. 1).

The complexity of the genome of *C. cohnii* has been analysed by several methods, which showed that about 50% of the genome is made up of repeated sequence, some very highly repeated (Allen et al. 1975; Hinnebusch et al. 1980; Moreau et al. 1998). We analysed the GSS by several means to identify repeats and determine the complexity of the sequence sample. All sequences were analysed by RepeatMasker to identify known families of

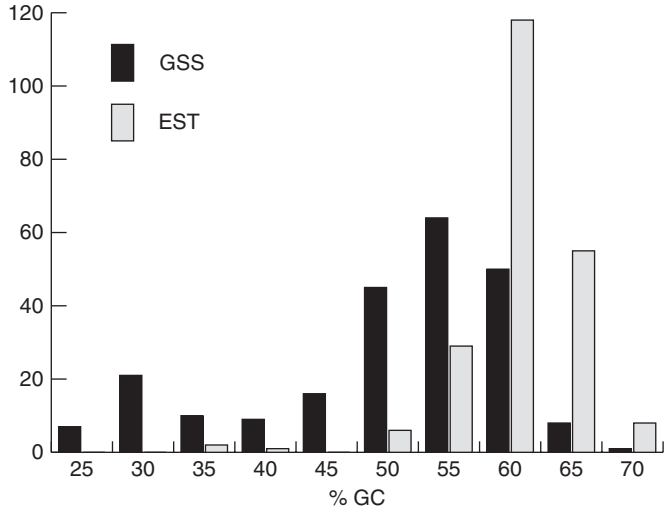

Fig. **1.** Distribution of GC bias between individual fragments of the genome sequence survey (GSS) and an equivalent number of randomly chosen cDNAs from *Heterocapsa triquetra*. The GSS sequences (black) have a lower average GC content than do cDNAs (grey), and also a wider distribution. EST, expressed sequence tag.

repeated elements (transposons and viruses), and also micro-repeats and low-complexity sequence (high AT or GC regions). Fourteen clones contained blocks of simple repeats, 2–5 bp in length, and an additional 13 clones contained tracts of low-complexity sequence, together making up 0.7% of the survey (Fig. 2).

The abundance of complex repeats was also assessed by comparing the survey sequences with one another using BLASTN. This revealed 26 classes of repeats scattered over many clones in sometimes complex patterns. Figure 3A shows the most pervasive cluster of repeated sequence in the survey. Here, each of 11 clones share regions of similarity of varying sizes with at least one other, but often flanked by unique regions. Other clusters of sequences sharing smaller regions of similarity were also found (e.g. Fig. 3B), as well as many cases of two clones sharing an otherwise unique region. None of these encoded long ORFs show any detectable similarity to known sequence. Altogether such regions make up 5.1% of the survey (Fig. 2).

The remainder of the non-coding, non-repeated, high-complexity sequence makes up an astonishing proportion of the survey. At 209,343 bp, 89.8% of the sequence is apparently simply sequence with no distinguishing features (Fig. 2). This appears to be what makes up the bulk of our sample. Some of this sequence is in all likelihood repeated elsewhere in the genome that has not been surveyed, but the majority of it likely does not make up high-copy repeats.

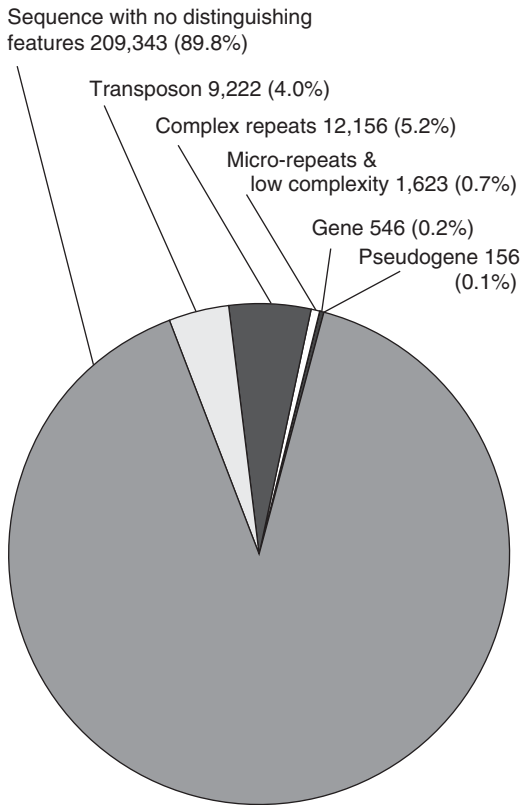

Fig. **2.** Breakdown of the genome sequence survey data by sequence type. Only 0.2% (546 bp) of the survey is putative gene sequence, while 0.1% (156 bp) is pseudogene, 0.7% (1,623 bp) is low-complexity sequence or micro-repeats, 5.2% (12,156 bp) is complex repeats, 4.0% (9,222 bp) is selfish elements (transposon- and virus-derived sequence), and 89.8% (209,343 bp) is complex, non-repeated, non-coding sequence, referred to as sequence with no distinguishing features.

Coding content of genomic DNA. Comparing all genomic DNA clones to public databases using BLASTN and BLASTX revealed 24 clones with credible hits to known sequences (Table 1). Of these, all but two sequences matched transposable elements or fragments of genes associated with selfish elements, such as transposons or viruses. These include a number of reverse transcriptases, DNA modification enzymes (methyltransferases and a methylase), and viral structural proteins and polyproteins. Analysis by RepeatMasker identified many of the same transposon-related sequences, and Table 1 includes data from both analyses with transposons classified according to the RepeatMasker identification. The presence of three DNA methylation enzymes is interesting given the unusual and widespread methylation of DNA in dinoflagellate genomes (Rae 1976). However, none of the enzymes identified here is of a class that performs the type of methylation common in dinoflagellate DNA. Rather, all these enzymes are most closely related to homologues from bacteria, phage, or eukaryotic viruses.

Only two clones contained sequence that could be recognised as dinoflagellate ''genes.'' One 2,656-bp clone encodes a fragment of a formate/nitrite transporter, several copies of which were identified in the *H. triquetra* EST survey (Patron et al. 2005), and also found in *Prorocentrum minimum* (GenBank Accession ABI14400). The genomic copy is not complete: the 5′-end is truncated at the end of the clone and the 3′-end is absent from the remaining sequence (Fig. 3C). As the remaining sequence is extensive and there is no sign of a splice junction, we conclude this is a pseudogene fragment (although we must note that dinoflagellate introns sometimes use non-canonical splice junctions). Interestingly, the region immediately downstream of the transporter fragment is highly structured compared with most of the survey. The region contains four copies of a direct, imperfect 199-bp repeat (Fig. 3C, D). A truncated copy of this same series of direct repeats is also found in another clone that lacks a recognizable ORF. Repeat units are more similar to units in the corresponding position in the other series of repeats than they are to repeats within their own series (Fig. 3D), and the region of similarity between the two clones extends past the last repeat, altogether suggesting the entire tandem repeat duplicated more recently than did the individual units.

The second clone with a recognizable gene is a 1,814-bp fragment that encodes a tandemly duplicated, spliced leader (SL) of *H. triquetra*. It has recently been shown that nuclear-encoded mRNAs in a wide diversity of dinoflagellates are trans-spliced to a universally conserved 22-bp fragment at their 5′-end (Lidie and van Dolah 2007; Zhang et al. 2007). The genes for the SL have been identified in *Karlodinium micrum*, *Pfiesteria piscicida*, and *P. minimum*, where they have been shown to appear singly or in clusters in the genome (Lidie and van Dolah 2007; Zhang et al. 2007). The 22-bp exon fragment is followed by a GT splice junction and an intronic fragment including a recognizable spliceosome-binding sequence, all of which is capable of forming a complex secondary structure. It has also been shown that many dinoflagellate cDNAs encode tandem duplicates of this leader, where the downstream copies are truncated and degenerate. mRNAs with these relict leaders are transcribed from genes that derived from processed cDNAs that were re-integrated into the genome following the addition of the SL, and they occur in a wide variety of dinoflagellates, including *H. triquetra* (Slamovits and Keeling 2008). The SL in the *H. triquetra* genomic clone appears 66-bp upstream of a 546-bp ORF with no similarity to any other known gene, but has been concluded to be an expressed gene due to the presence and nature of the relict SL (Slamovits and Keeling 2008). This is not likely to be a gene for the SL itself because it lacks the GT splice junction that must follow the 22 bp of the SL sequence and because the relict SL is itself a tandem duplicate,

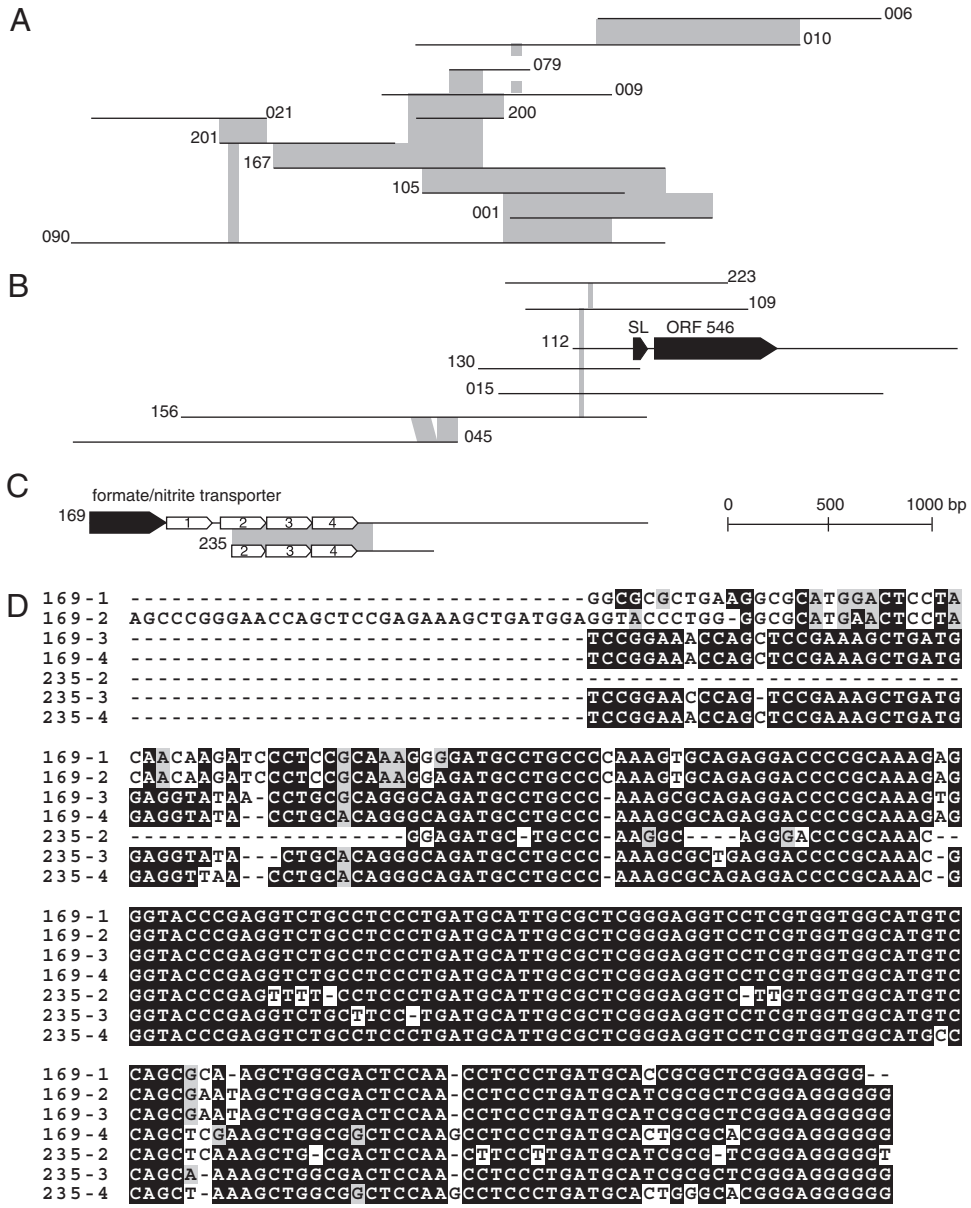

Fig. **3.** Examples of repeated sequences in the *Heterocapsa triquetra* genome. (**A**) A complex distribution of repeated sequences involving 11 individual fragments. Fragments are indicated by lines and numbers, and regions of shared sequence similarity with other fragments are indicated by grey shading between the lines. (**B**) A smaller region of highly conserved sequence is also shared between the fragment encoding the spliced leader (SL) and several fragments with no detectable similarity to anything else or one another. One of these also shares a region with an additional clone. (**C**) The region downstream of the formate/nitrite transporter fragment is highly structured, with four tandem repeats, three of which are also found on another clone. The sequences of repeats (**D**) are more similar to the corresponding repeat on the other clone than they are to their neighbour repeats, suggesting the tandem repeat was duplicated as a whole.

suggesting there are three SL sequences on the mRNA for this gene. Overall, therefore, we conclude this ORF is a legitimate gene.

Interestingly, the SL-encoding region is also flanked by structured DNA. Immediately upstream of the SL gene is a 120-bp region of relatively high AT content, which is made up of many short repeats, and might perhaps represent a transcriptional control sequence. Further upstream is a 45-bp region that is also found on several other clones, each of which also has other short regions in common with still other clones (Fig. 3B).

DISCUSSION

At 18–23 billion base pairs, it is one of the larger genomes recorded to date, but the sequences from the survey suggests that by far most of the DNA is not classifiable—it is not repeated sequence, selfish elements, pseudogenes, or any other kind of sequence that we can recognizably label. Physical analyses of the *C. cohnii* genome (Allen et al. 1975; Hinnebusch et al. 1980; Moreau et al. 1998) suggest a much higher proportion of repeated sequences than we found (about 50% and about 10%), but given

Table **1.** Sequence identification by homology.

Clone	Type	Identity
112	Putative gene	Spliced leader
169	Pseudogene	Nitrite/formate transporter
108	Transposon	LTR/Ngaro
136	Transposon	LTR/Gypsy
178	Transposon	LTR/Gypsy
029	Transposon	LTR/Copia
089	Transposon	LTR/Copia
118	Transposon	LTR/Copia
173	Transposon	LTR/Copia
186	Transposon	LTR/Copia
192	Transposon	LTR/Copia
213	Transposon	LTR/Copia
272	Transposon	LTR/Copia
278	Transposon	LTR/Copia
281	Transposon	LTR/Copia
145	Transposon	LINE/L1/Zorro
030	Transposon	LINE/RTE-RTE/R2
010	Transposon	DNA/TcMar-Tigge
206	Transposon	DNA/Maverick
135	Transposon/virus protein	RNA-directed DNA polymerase
020	Transposon/virus protein	DNA adenine methylase
070	Transposon/virus protein	Type III restriction endonuclease
078	Transposon/virus protein	Cytosine-specific methyltransferase
204	Transposon/virus protein	Cytosine DNA methyltransferase

our small sample size it is likely that many of the unique sequences we sampled are likely repeated somewhere else in the genome. Indeed, a low-redundancy survey of a genome with a large proportion of low-copy repeats would yield a high proportion of apparently unique sequence, which may be the case here. Even so, the survey shows at least some sequences must be present in large numbers throughout the genome, as we sampled some long, complex repeats numerous times. Roughly the same proportion of the survey (about 4%) is made up of several sequences with similarity to known transposons or viruses, which is a relatively small proportion of the genome, but extrapolated to the whole genome still accounts for a large number of selfish elements.

The survey yielded a single putative gene and one pseudogene. This suggests a very low gene density, which is not surprising for such a large genome, but how many genes would one expect to find in such a sample? In the completely sequenced genome of the diatom *Thalassiosira pseudonanna*, 11,000 protein-coding genes were annotated. If *H. triquetra*, also an alga with a red secondary plastid, encoded roughly the same number of genes, then we would predict a gene density in the neighbourhood of one gene per 2,000,000 bp or about 10 times the size of our sample. The fact that one (recognizable) gene was sampled is therefore unexpected, but not terribly surprising. It is less surprising still when one considers that many genes may be present in multiple copies. Indeed, the many EST surveys that have been carried out on dinoflagellates (Bachvaroff et al. 2004; Hackett et al. 2004; Nosenko and Bhattacharya 2007; Patron et al. 2005, 2006) have all shown many genes to be present in multiple distinct copies. Other analyses of gene copy number of specific genes have found a great deal of variability, but some very highly repeated genes (Le et al. 1997; Zhang, Hou, and Lin, 2006). We speculate that this is not due to extended functionality of the proteins and is not part of the cause of the large genome size, but is rather a consequence of the genome size: the large size of the genome allows genes to exist in multiple copies with no ill effect. One possible mechanism for the massive duplication of many genes has also been recently described. Dinoflagellate genomes appear to be taking up processed cDNAs at a high frequency, and because this process is duplicative, it provides an obvious route for the expansion of copy number for many genes (Slamovits and Keeling 2008).

While the characteristics of the genome of *H. triquetra* suggested by this survey are interesting and consistent with what little is known about the genomes of dinoflagellates, we should stress once more in closing that the small fraction of the genome sampled here makes it difficult to draw firm quantitative conclusions about the content of the genome. Moreover, this sample provides little insight into the overall structure of the genome at the sequence level. For example, we do not know if genes or other recognizable sequence elements are clustered or evenly spread, if there is some periodicity to the organisation of sequences on the genome, or if there are zones of differing characteristics. These features require larger fragments, or perhaps whole chromosomes to be sequenced.

ACKNOWLEDGMENTS

This work was supported by a grant from the Canadian Institutes for Health Research (MOP-84256). We thank Audrey de Koning for technical help with library construction, and Juan Saldarriaga for important discussions. C. H. S. is supported by a grant from the Tula Foundation to the Centre for Microbial Diversity and Evolution. P. J. K. is a Fellow of the Canadian Institute for Advanced Research and a Senior Scholar of the Michael Smith Foundation for Health Research.

LITERATURE CITED

Allen, J. R., Roberts, M., Loeblich, A. R. III & Klotz, L. C. 1975. Characterization of the DNA from the dinoflagellate *Crypthecodinium cohnii* and implications for nuclear organization. *Cell*, **6**:161–169.

Bachvaroff, T. R., Concepcion, G. T., Rogers, C. R., Herman, E. M. & Delwiche, C. F. 2004. Dinoflagellate expressed sequence tags data indicate massive transfer of chloroplast genes to the nuclear genome. *Protist*, **155**:65–78.

Barbrook, A. C., Symington, H., Nisbet, R. E., Larkum, A. & Howe, C. J. 2001. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. *Mol. Genet. Genomics*, **266**:632–638.

Cavalier-Smith, T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot.*, **95**:147–175.

Costas, E. & Goyanes, V. 2005. Architecture and evolution of dinoflagellate chromosomes: an enigmatic origin. *Cytogenet. Genome Res.*, **109**:268–275.

Dodge, J. D. 1966. The Dinophyceae. *In*: Godward, M. B. E. (ed.), The Chromosomes of the Algae. Arnold, London. p. 96–115.

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

Gregory, T. R. & Witt, J. D. S. 2008. Population size and genome size in fishes: a closer look. *Genome*, **51**:309–313.

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

Herzog, M., Soyer, M. O. & Daney de Marcillac, G. 1982. A high level of thymine replacement by 5-hydroxymethyluracil in nuclear DNA of the primitive dinoflagellate *Prorocentrum micans* E. *Eur. J. Cell Biol.*, **27**:151–155.

Hinnebusch, A. G., Klotz, L. C., Immergut, E. & Loeblich, A. R. III 1980. Deoxyribonucleic acid sequence organization in the genome of the dinoflagellate *Crypthecodinium cohnii*. *Biochemistry*, **19**:1744–1755.

Jackson, C. J., Norman, J. E., Schnare, M. N., Gray, M. W., Keeling, P. J. & Waller, R. F. 2007. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biol.*, **5**:41.

Kapraun, D. F. 2005. Nuclear DNA content estimates in multicellular green, red and brown algae: phylogenetic considerations. *Ann. Bot.*, **95**:7–44.

Kasinsky, H. E., Lewis, J. D., Dacks, J. B. & Ausio, J. 2001. Origin of H1 linker histones. *FASEB J.*, **15**:34–42.

Keeling, P. J. & Slamovits, C. H. 2005. Causes and effects of nuclear genome reduction. *Curr. Opin. Genet. Dev.*, **15**:601–608.

Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., Roger, A. J. & Gray, M. W. 2005. The tree of eukaryotes. *Trends Ecol. Evol.*, **20**:670–676.

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

Le, Q. H., Markovic, P., Hastings, J. W., Jovine, R. V. & Morse, D. 1997. Structure and organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*. *Mol. Gen. Genet.*, **255**:595–604.

Lidie, K. B. & van Dolah, F. M. 2007. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. *J. Eukaryot. Microbiol.*, **54**:427–435.

Lynch, M. & Conery, J. S. 2003. The origins of genome complexity. *Science*, **302**:1401–1404.

Moreau, H., Geraud, M. L., Bhaud, Y. & Soyer-Gobillard, M. O. 1998. Cloning, characterization and chromosomal localization of a repeated sequence in *Crypthecodinium cohnii*, a marine dinoflagellate. *Int. Microbiol.*, **1**:35–43.

Nash, E. A., Barbrook, A. C., Edwards-Stewart, R. K., Bernhardt, K., Howe, C. J. & Nisbet, R. E. 2007. Organisation of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. *Mol. Biol. Evol.*, **24**:1528–1536.

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

Patron, N. J., Waller, R. F. & Keeling, P. J. 2006. A tertiary plastid uses genes from two endosymbionts. *J. Mol. Biol.*, **357**:1373–1382.

Patron, N. J., Waller, R. F., Archibald, J. M. & Keeling, P. J. 2005. Complex protein targeting to dinoflagellate plastids. *J. Mol. Biol.*, **348**:1015–1024.

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

Raikov, I. B. 1995. The dinoflagellate nucleus and chromosomes—Mesokaryote concept reconsidered. *Acta Protozool.*, **34**:239–247.

Rizzo, P. J. 1991. The enigma of the dinoflagellate chromosome. *J. Protozool.*, **38**:246–252.

Rizzo, P. J. 2003. Those amazing dinoflagellate chromosomes. *Cell Res.*, **13**:215–217.

Saldarriaga, J. F., McEwan, M. L., Fast, N. M., Taylor, F. J. R. & Keeling, P. J. 2003. Multiple protein phylogenies show that *Oxyhrris marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int. J. Sys. Evol. Microbiol.*, **53**:355–365.

Sigee, D. C. 1983. Structural DNA and genetically active DNA in dinoflagellate chromosomes. *Biosystems*, **16**:203–210.

Slamovits, C. H. & Keeling, P. J. 2008. Widespread recycling of processed cDNAs in dinoflagellate genomes. *Curr. Biol.*, **18**:550–552.

Slamovits, C. H., Saldarriaga, J. F., Larocque, A. & Keeling, P. J. 2007. The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *J. Mol. Biol.*, **372**:356–368.

Smit, A. F. A., Hubley, R. & Green, P. 2004. RepeatMasker Open-3.0. http://www.repeatmasker.org

Triemer, R. E. & Fritz, L. 1984. Cell cycle and mitosis. *In*: Spector, D. L. (ed.), Dinoflagellates. Academic Press, Orlando. p. 149–179.

Veldhuis, M. J. W., Cucci, T. L. & Sieracki, M. E. 1997. Cellular DNA content of marine phytoplankton using two new fluorochromes: taxonomic and ecological implications. *J. Phycol.*, **33**:527–541.

Vinogradov, A. E. 2004. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr. Opin. Genet. Dev.*, **14**:620–626.

Waller, R. F., Patron, N. J. & Keeling, P. J. 2006. Phylogenetic history of plastid-targeted proteins in the peridinin-containing dinoflagellate *Heterocapsa triquetra*. *Int. J. Syst. Evol. Microbiol.*, **56**:1347–1439.

Zhang, Z., Green, B. R. & Cavalier-Smith, T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature*, **400**:155–159.

Zhang, Z., Green, B. R. & Cavalier-Smith, T. 2000. Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. *J. Mol. Evol.*, **51**:26–40.

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

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

Received: 04/25/08, 06/06/08, 06/20/08; accepted: 06/21/08