J. Phycol. **44,** 99–102 (2008)

DOI: 10.1111/j.1529-8817.2007.00444.x

NOTE

ANALYSIS OF EXPRESSED SEQUENCE TAGS FROM THE MARINE MICROALGA *NANNOCHLOROPSIS OCULATA* (EUSTIGMATOPHYCEAE)[1]

Juan Shi, Kehou Pan[2]*, Jianzhong Yu, Baohua Zhu*

Ministry of Education Key Laboratory of Mariculture, Ocean University of China, Qingdao 266003, China

Guanpin Yang

College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China

Wengong Yu

Ministry of Education Key Laboratory of Marine Drugs, Ocean University of China, Qingdao 266003, China

and Xinyu Zhang

Ministry of Education Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, China

Nannochloropsis oculata **(Droop) D. J. Hibberd (Eustigmatophyceae), a marine eukaryotic unicellular alga, is widely used in mariculture as live feed. It is considered to be of high nutritional value owing to its high content of proteins; polyunsaturated fatty acids (PUFAs), especially eicosapentaenoic acid (EPA, C20:5n3); and diverse pigments. Previous studies of this microalga focused on its taxonomy, culture, and biochemistry, but little is known at the molecular level. Establishing a molecular base is vital to understand the biological processes of this alga. Therefore, we constructed a cDNA library using algal cells grown at exponential growth phase and carried out expressed sequence tag (EST) analysis. A total of 1,960 nonredundant sequences (NRSs) were generated for *N. oculata* clone CS-179. Only 32.5% of NRSs showed significant similarity (*E* < 1e-04) to proteins registered in the GenBank nonredundant protein database. The KOG (clusters of euKaryotic Orthologous Groups) profile database returned significant hits for 490 NRSs. Analysis revealed that a large proportion of NRSs could be unique to this microalga.**

Key index words: **Eustigmatophyceae; expressed sequence tag; genomics; lipid metabolism; *Nannochloropsis oculata***

Abbreviations: **EPA, eicosapentaenoic acid (C20:5n3); EST, expressed sequence tag; KOG, clusters of euKaryotic Orthologous Groups; NRSs, nonredundant sequences; PUFAs, polyunsaturated fatty acids**

Nannochloropsis is a unique marine microalgal genus belonging to the Eustigmatophyceae (Hibberd 1981, Maruyama et al. 1986). It includes several species, which are generally regarded as being picoeukaryotic planktonic since their sizes range from 2 to 5 μm (Hu and Gao 2003). Though previously known as ''marine *Chlorella*,'' *Nannochloropsis* contains no chl other than chl *a* (Adl et al. 2005). This microalga is distributed throughout the world's oceans and plays a significant role in the global carbon and mineral cycles, especially in oligotrophic seawater (Fogg 1995). Furthermore, it has considerable potential for commercial exploitation. It is widely used in mariculture as a feed source and is proposed as an alternative source of fish oil due to its high content of PUFAs, especially EPA (Volkman et al. 1993, Tonon et al. 2002). Diverse pigments are also produced in high amounts (Lubián et al. 2000).

Though there are some morphological, biochemical, and biophysical studies on *Nannochloropsis*, little is known at the molecular level. Only a few related nucleotide sequences are listed with the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA), and most of the full-length records are variants of the 18S rRNA gene. Considering that this alga has great ecological and commercial value, greater effort should be made to investigate its genome, which might help us better understand the molecular mechanism of its various biological processes. Generating ESTs is an efficient way to obtain information on expressed genes, and this method has been widely used with algae (Grossman 2005, Walker et al. 2005).

Construction of cDNA library. N. oculata clone CS-179 (CSIRO Collection of Living Microalgae, Tasmania, Australia) was grown in sterilized seawater (31 p.p.t., pH 8.1) enriched with f/2 medium (Guillard and Ryther 1962) at 20 ± 1°C and 70 μmol

[1]Received 11 December 2006. Accepted 7 May 2007.
[2]Author for correspondence: e-mail khpan@ouc.edu.cn.

photons·m^{-2}·s^{-1} white fluorescent light on 12:12 light:dark (L:D) rhythm. Cells were collected at the logarithmic growth phase. Total RNA was extracted by the guanidine thiocyanate-phenol-chloroform extraction method (Chomczynski and Sacchi 1987); mRNA was further purified using the Oligotex mRNA Extraction Kits (Qiagen, Hilden, Germany), according to the manufacturer's instructions. Total RNA and mRNA were run on agarose gel to investigate the quality. cDNA synthesis was performed using the pBluescript II SK(+) XR cDNA Library Construction Kit (Stratagene Cloning Systems, La Jolla, CA, USA). The library was created in Uni-Zap XR vector using oligo (dT)18 primers and directionally inserted into *Eco*RI–*Xho*I sites of pBluescript.

EST sequencing. Sequencing was done from the 5′-end using standard T3 primer. Reactions were performed using the DYEnamic™ ET Dye Terminator Kit (Amersham Biosciences, Little Chalfont, Buckinghamshire, UK) and run on the MegaBACE 1000 DNA sequencing system (Amersham Pharmacia Biotech, Piscataway, NJ, USA).

Clustering and generation of unigene set. The sequences obtained were processed to remove vector-derived sequence using Cross_match programs (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html, Green), and all sequences were self-blasted to remove redundancy and clustered. The longest sequence in each cluster was selected and pooled to form the nonredundant collection. These NRSs were then subjected to annotation.

Codon usage. The codon usage was determined using coding regions from 861 NRSs without mitochondrial and plastidal sequences. Comparison of codon usage between *N. oculata* and other species (obtained from http://www.kazusa.or.jp/codon and http://avesthagen.sznbowler.com/) was performed using chips and codcmp from the EMBOSS package (http://emboss.sourceforge.net/).

Sequence annotation. To examine sequence similarity to known genes, BLASTX (Stephen et al. 1997) was run against the nr protein sequence databases, and BLASTN (Stephen et al. 1997) was run against both the nt nucleotide sequence database and the EST database. All BLAST matches were filtered with the expectation value $E < 1e\text{-}04$. The KOG database for clusters of orthologous proteins from eukaryotes was searched through RPS BLAST (Marchler-Bauer et al. 2002). The filter condition was that the identity should be >40% and the HSP length should >30.

Findings and conclusions. In this study, we present the first characterization of ESTs from *N. oculata* clone CS-179, the typical species of the genus. A cDNA library consisting of 1.5×10^5 clones was constructed using algal cells grown at exponential growth phase. It was not normalized to reduce the abundance of cDNAs appearing in high copy number. So EST redundancy is likely to more accurately reflect gene-expression level in exponentially growing cultures.

A total of 5,812 clones were then sequenced from the 5′ end. After filtering for vector contaminants and poor-quality runs, a collection of 5,315 EST reads was obtained. Sequence lengths varied from 100 to 716 nucleotides (nt) with an average of 420 nt. The trimmed reads were clustered, and 1,960 sequences were selected to form a nonredundant collection. The tag frequency in the library was also analyzed; 508 clusters contained more than two EST reads. Among them, four clusters consisted of up to 100 or 200 EST reads.

Only several dozen hits were observed in the nt database. There were 551 NRSs (28.1%) with matches in the EST database. Only 637 (32.5%) sequences showed significant similarity to proteins registered in the nr database, and the remainder showed only weak similarity or no similarity. This finding may indicate that *N. oculata* harbors a large number of unique proteins. Similar features had been reported in the EST analyses of other algae. For example, >51.9% of the NRSs of *Ulva linza* could not be identified (Stanley et al. 2005), and 70% of the NRSs of *Chlamydomonas reinhardtii* had no significant hits (Asamizu et al. 1999, 2000). In contrast, 31% of the NRSs of *Arabidopsis* showed no homology to known proteins (Arabidopsis Genome Initiative 2000). One reason could be that there is less genomic information available for algae than for higher eukaryotes in the public databases. Another possibility is that some algae are evolutionary divergent even though they belong to the same taxonomic group, so genes specific to each alga with highly divergent sequences may be present. For example, *C. reinhardtii* and *Scherffelia dubia* share very few of their ESTs, although they are both flagellate taxa within the Chlorophyta (Becker et al. 2001). *N. oculata* bears a superficial resemblance to some green algae, in both color and cell morphology. For this reason, strain CS-179 had been incorrectly referred to as ''marine *Chlorella*'' (Chlorophyceae) until Maruyama et al. (1986) showed that it was identical to the previously described *N. oculata* according to its cell ultrastructure, pigment content, and fatty acid composition. The results of EST analysis also suggested that genetic (and phylogenetic) heterogeneity between *Nannochloropsis* and *Chlorella* might be much larger than previously thought since few sequences exhibited similarity between these two algae.

In our study, 21% of NRSs had a 5′ untranslated region (UTR) with a length of more than 100 base pairs (bp). The codon usage was determined using coding regions from 861 NRSs. The total G/C content of coding sequences was 54.28%, with the first, second, and third letters being 53.4%, 47.91%, and 61.32% G/C, respectively (Table 1). Comparison of codon usage between *N. oculata* and other species showed that this microalga may have a closer genetic relationship with *C. reinhardtii* than other species (Table 1).

TABLE 1. Codon usage comparison between *Nannochloropsis oculata* and other species.

	G/C content (%)				Sum squared difference (vs. *N. oculata*)
	Triplets	1st letter	2nd letter	3rd letter	
N. oculata	54.28	53.63	47.91	61.32	–
Chlamydomonas reinhardtii	66.26	64.71	47.83	86.24	4.082
Ostreococcus tauri	59.05	59.54	46.01	71.61	1.861
Phaeodactylum tricornutum	51.72	56.54	41.92	56.69	1.575
Thalassiosira pseudonana	49.60	54.11	41.94	52.76	1.245
Arabidopsis thaliana	44.60	50.88	40.52	42.39	1.818

Codon usage was calculated for *N. oculata* using 861 coding sequences (CDSs), *C. reinhardtii* using 748 CDSs, *O. tauri* using 61 CDSs, *P. tricornutum* using 859 CDSs, *T. pseudonana* using 465 CDSs, and *A. thaliana* using 73,223 CDSs.

The KOG database was searched through RPS BLAST. A total of 490 NRSs (25%) could be sorted into 24 KOG functional categories based on their putative cellular function. The number of sequences falling into different functional groups has been summarized in Table 2. Among them, sequences characteristic of general function prediction only were the most abundant. The majority of the rest showed similarity to proteins that are required for translation, ribosomal structure and biogenesis, and posttranslational modification. It has been shown that protein synthesis is very active in exponentially growing cells. Among the highly expressed genes, we found a known sequence encoding *Nannochloropsis* sp. violaxanthin/chl *a* binding protein precursor (Vcp), which had been characterized previously (Sukenik et al. 2000).

TABLE 2. Classification of the *Nannochloropsis oculata* non-redundant sequences (NRSs) with similarity to known protein genes by their functional categories.

Functional categories	No. NRSs
Amino acid transport and metabolism	11
Carbohydrate transport and metabolism	23
Cell-cycle control, cell division, chromosome partitioning	5
Cell-wall/membrane/envelope biogenesis	1
Chromatin structure and dynamics	8
Coenzyme transport and metabolism	7
Cytoskeleton	6
Defense mechanisms	2
Energy production and conversion	29
Inorganic ion transport and metabolism	12
Intracellular trafficking, secretion, and vesicular transport	14
Lipid transport and metabolism	9
Nuclear structure	1
Nucleotide transport and metabolism	4
Posttranslational modification, protein turnover, chaperones	58
Replication, recombination, and repair	5
RNA processing and modification	33
Secondary metabolites biosynthesis, transport, and catabolism	13
Signal transduction mechanisms	53
Transcription	21
Translation, ribosomal structure, and biogenesis	63
Mitochondria/chloroplast-located protein	15
General function prediction only	76
Function unknown	21
Total	490

Several sequences with similarity to proteins involved in lipid transport and metabolism were obtained, including long-chain acyl-CoA synthetases, long chain acyl-CoA transporter, fatty acid desaturase, and so forth. They presumably participate in the biosynthesis of PUFAs in *N. oculata*. Considering that this microalga is rich in EPA, identification of these genes will be very important.

In this study, we present the first characterization of the EST data set from the eustigmatophyceaen alga *N. oculata*, which offers the molecular basis for further research into its biochemical, physiological, cellular, and other biological processes. The ESTs are useful in gene discovery, polymorphism analysis, gene prediction, and analysis of gene expression in response to environmental factor changes using microarray technology.

The EST sequences reported in this paper have been registered in GenBank/EMBL (European Molecular Biology Laboratory)/DDBJ (DNA Data Bank of Japan) with the accession numbers EE109295–EE111252. Requests for bulk queries should be addressed to K. Pan.

The authors are grateful to Prof. Shuanglin Dong for his valuable suggestions on project design, to Changfeng Li and Yongjun Fang for their assistance with sequencing and preliminary sequence analysis, and to Beijing Cofly Bioinformatics Co. Ltd. (http://www.co-fly.net) for bioinformatics analysis. This work was supported by the National 973 Program (No. 2005CCA02400).

Adl, S. M., Simpson, A. G., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* 52:399–451.

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

Asamizu, E., Miura, K., Kucho, K., Inoue, Y., Fukuzawa, H., Ohyama, K., Nakamura, Y. & Tabata, S. 2000. Generation of expressed sequence tags from low-CO_2 and high-CO_2 adapted cells of *Chlamydomonas reinhardtii*. *DNA Res.* 7:305–7.

Asamizu, E., Nakamura, Y., Sato, S., Fukuzawa, H. & Tabata, S. 1999. A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res.* 6:369–73.

Becker, B., Feja, N. & Melkonian, M. 2001. Analysis of expressed sequence tags (ESTs) from the scaly green flagellate *Scherffelia dubia* Pascher emend. Melkonian et Preisig. Protist 152:139–47.

Chomczynski, P. & Sacchi, N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* 162:156–9.

Fogg, G. E. 1995. Some comments on picoplankton and its importance in the pelagic ecosystem. *Aquat. Microb. Ecol.* 9:33–9.

Grossman, A. R. 2005. Paths toward algal genomics. *Plant Physiol.* 137:410–27.

Guillard, R. R. L. & Ryther, J. H. 1962. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt and *Detonula conferraceae* Cleve. *Can. J. Microbiol.* 8:229–39.

Hibberd, D. J. 1981. Notes on the taxonomy and nomenclature of the algal classes Eustigmatophyceae and Tribophyceae (synonym Xanthophyceae). *Bot. J. Linn. Soc.* 82:93–119.

Hu, H. & Gao, K. 2003. Optimization of growth and fatty acid composition of a unicellular marine picoplankton, *Nannochloropsis* sp., with enriched carbon sources. *Biotechnol. Lett.* 25:421–5.

Lubián, L. M., Montero, O., Moreno-Garrido, I., Huertas, E., Sobrino, C., González-del Valle, M. & Parés, G. 2000. *Nannochloropsis* (Eustigmatophyceae) as a source of commercially valuable pigments. *J. Appl. Phycol.* 12:249–55.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 30:281–3.

Maruyama, I., Nakamura, T., Matsubayashi, T., Ando, Y. & Maeda, T. 1986. Identification of the alga known as 'marine *Chlorella*' as a member of the Eustigmatophyceae. *Jpn. J. Phycol.* 34:319–25.

Stanley, M. S., Perry, R. M. & Callow, J. A. 2005. Analysis of expressed sequence tags from the green alga *Ulva linza* (Chlorophyta). *J. Phycol.* 41:1219–26.

Stephen, F. A., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, J. D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–402.

Sukenik, A., Livne, A., Apt, K. E. & Grossman, A. R. 2000. Characterization of a gene encoding the light-harvesting violaxanthin-chlorophyll protein of *Nannochloropsis* sp. (Eustigmatophyceae). *J. Phycol.* 36:563–70.

Tonon, T., Harvey, D., Larson, T. R. & Graham, I. A. 2002. Long chain polyunsaturated fatty acid production and partitioning to triacylglycerols in four microalgae. *Phytochemistry* 61:15–24.

Volkman, J. K., Brown, M. R., Dunstan, G. A. & Jeffrey, S. W. 1993. The biochemical composition of marine microalgae from the class Eustigmatophyceae. *J. Phycol.* 29:69–78.

Walker, T. L., Collet, C. & Purton, S. 2005. Algae transgenics in the genomic era. *J. Phycol.* 41:1077–93.