J. Phycol. **47,** 1454–1460 (2011)

DOI: 10.1111/j.1529-8817.2011.01071.x

ANALYSIS OF EXPRESSED SEQUENCE TAGS FROM THE GREEN ALGA *DUNALIELLA SALINA* (CHLOROPHYTA)[1]

Rui Zhao, Yu Cao, Hui Xu, Linfeng Lv, Dairong Qiao, and Yi Cao[2]

Microbiology and Metabolic Engineering Key Laboratory of Sichuan Province, College of Life Sciences, Sichuan University, Chengdu, China, 610064

The unicellular green alga *Dunaliella salina* (Dunal) Teodor. is a novel model photosynthetic eukaryote for studying photosystems, high salinity acclimation, and carotenoid accumulation. In spite of such significance, there have been limited studies on the *Dunaliella* genome transcriptome and proteome. To further investigate *D. salina*, a cDNA library was constructed and sequenced. Here, we present the analysis of the 2,282 expressed sequence tags (ESTs) generated together with 3,990 ESTs from dbEST. A total of 4,148 unique sequences (UniSeqs) were identified, of which 56.1% had sequence similarity with Uniprot entries, suggesting that a large number of unique genes may be harbored by *Dunaliella*. Additionally, protein family domains were identified to further characterize these sequences. Then, we also compared EST sequences with different complete eukaryotic genomes from several animals, plants, and fungi. We observed notable differences between *D. salina* and other organisms. This EST collection and its annotation provided a significant resource for basic and applied research on *D. salina* and laid the foundation for a systematic analysis of the transcriptome basis of green algae development and diversification.

Key index words: Dunaliella salina; **expressed sequence tag; genomics; green algae; transcriptome**

Abbreviations: **BLAST, basic local alignment search tool; EF hand, a helix-loop-helix structural domain found in a large family of calcium-binding proteins; ESTs, expressed sequence tags; KO, KEGG Orthology; UniProtKB, UniProt knowledgebase; UniSeqs, Unique sequences**

D. salina, the most well-known halophilic unicellular green alga, is a member of the class Chlorophyceae. *D. salina* is a recognized model photosynthetic organism for studying adaption to high salinity (Boetius and Joye 2009, Kim et al. 2010). *D. salina* can survive under high salt stress and has a high optimal salt concentration of ~12%. Much research concerned with osmotic balance in *D. salina* had been conducted to reveal its halotolerant mechanism (Cowan et al. 1992, Fisher et al. 1996, Katz et al. 2007, Chen et al. 2009, Kim et al. 2010), whereas relevant biotechnology applications focused more on its capacity for massive carotenoid accumulation (Jin and Melis 2003, Murthy et al. 2005, Lamers et al. 2008, Ye et al. 2008). However, in contrast to the continuous and extensive work devoted to *D. salina* (Oren 2005), there was still a small number of sequence resources in public databases such as National Center for Biotechnology Information (NCBI; 215 nucleotides, 238 proteins, 4,058 ESTs available).

The advent of modern genomics approaches and bioinformatics technologies has enabled researchers to depart from a candidate gene or pathway approach and begin to explore organismal development and evolution from a genome, transcriptome, or proteome perspective, mostly focusing on existing genetic model systems such as *Chlamydomonas reinhardtii*, a model system for studying green algae (Merchant et al. 2007). However, many key questions concerning unique organismal innovation, diversification, and evolution are difficult to address solely within the confines of typical model systems. Recent efforts have therefore begun to generate genomic and developmental genetic resources for organisms with promise as future model systems in evolutionary developmental biology and ecological genetics (Park et al. 2006, Kijimoto et al. 2009, Smith et al. 2010).

The *D. salina* whole-genome sequencing project carried out by the Department of Energy (DOE) Joint Genome Institute (http://www.jgi.doe.gov/) was started in 2006 and is approaching its completion, whereas the transcriptome projects for the purpose of *Chlamydomonas* and *Chlorella* transcriptome analysis including *D. salina* were started in 2010. However, its taxonomic relative *Chlamydomonas reinhardtii* (Asamizu et al. 1999, Jain et al. 2007, Merchant et al. 2007) had been well surveyed, and the *Volvox carteri* standard draft genome project has been completed. ChlamyCyc database (May et al. 2009) provides a curated and integrated systems biology repository that will enable and assist in systematic studies of fundamental cellular processes in *Chlamydomonas*.

[1]Received 19 November 2010. Accepted 11 May 2011.
[2]Author for correspondence: e-mail geneium@scu.edu.cn.

The purpose of the present article was to report a transcriptomics analysis of 6,118 ESTs obtained from *D. salina*. EST sequences were derived from different cDNA libraries at different growth conditions. Overall, 4,148 UniSeqs were identified and well annotated. The ESTs were all also compared with genomic sequences of several animals, plants, and algae.

MATERIALS AND METHODS

cDNA library construction and sequencing. The construction of the *D. salina* normalized cDNA library and generation of ESTs has been described (Li et al. 2004). Briefly, *D. salina* strain was bought from the Institute of Hydrobiology, the Chinese Academy of Sciences. The algal cells were grown in defined medium containing 1.5 mol $\cdot$ L^{-1} NaCl and at 16:8 light:dark (L:D) at 25°C. The cells were grown to a density of 6×10^5 cells $\cdot$ mL^{-1} and collected by centrifugation (Heraeus Instruments Inc., Newton, CT, USA) at 4°C, 1,360*g*, and totally transferred to a prechilled RNase-free Eppendorf (Axygen Biosciences, CA, USA) tube immediately. Approximately 6×10^6 cells were used to isolate total RNA according to RNeasy Mini Protocol (Qiagen China [Shanghai] Co. Ltd., Shanghai, China) for the isolation of total RNA from plant cells. DNA contamination was removed by DNase digestion protocol recommended in the user manual of this kit. Total RNA was checked with denaturing agarose gel electrophoresis. The concentration of RNA was determined by measuring the absorbance at 260 nm (A_{260}). Purity was assessed by the ratio of 260 nm and 280 nm (A_{260}/A_{280}). cDNA was prepared using SMART™ PCR cDNA synthesis kit (Clontech Laboratories Inc., Mountain View, CA, USA). PCR amplicons were resolved on 1.1% agarose/EtBr gels and visually compared with 1 kb DNA markers (#N3200S, NEB Inc., Ipswich, MA, USA). EST sequencing was performed at United Gene (United Gene Group Ltd., Shanghai, China) using ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, CA, USA).

Submission of ESTs to NCBI GenBank. The ESTs generated in this study have been submitted to dbEST under GenBank accession numbers HO847594–HO849875 and dbEST Id numbers 71459168–71461449.

Additional EST libraries. Besides the 2,282 ESTs sequenced from our cDNA library, we combined available *D. salina* ESTs (from six different cDNA libraries as listed in Table S1 in the supplementary material) in dbEST (http://www.ncbi.nlm.nih.gov/projects/dbEST/) in our analysis to generate a diverse set of sequences involved in different culture conditions.

Analysis and assembly of sequence data. First, quality control of sequenced ESTs and ESTs in additional EST libraries was performed as described below. Vector trimming was performed using the program crossmatch (http://www.phrap.org/). Repeats were masked by RepeatMasker (http://repeatmasker.org, Smith et al. 2010). It employs crossmatch and up-to-date repeat libraries for different species, to carry out species-specific repeat masking. Poly-A tails, low-quality segments at 5′ and 3′ cDNA ends and low-complexity regions, short ESTs (<100 bp) were trimmed using the program SeqClean (http://seqclean.sourceforge.net/).

Then, the trimmed masked sequences were clustered and assembled into contigs using CAP3 (Huang and Madan 1999, Liang et al. 2000) program, and its output was visualized using PAVE (Soderlund et al. 2009) pipeline.

Annotation of UniSeqs. EST pairs were identified by a BLASTn (Altschul et al. 1997) search with a cutoff *E*-value of 1E-50, and sequence pairs could be considered potentially identical.

We matched the UniSeqs we assembled from previous processes to the complete UniProtKB (UniProtKB/Swiss-Prot release 57.9 and UniProtKB/TrEMBL release 40.9) (Apweiler et al. 2010) data set using BLASTx 2.2.22 program with cutoff *E*-value of 1E-20 and 1E-8, respectively, and both the top match that had the lowest *E*-value and the following hits were recorded. The output information from BLASTx was parsed and loaded into a local MySQL database; at the same time, all relevant protein (that had matched with our contigs) annotations in UniProtKB were imported into this local database, with hopes that utilizing the relation database techniques would facilitate our further study as it did in many similar studies (Beldade et al. 2006, Vizcaino et al. 2006, Baxendale et al. 2009).

Protein family, domain, and function site characterization were confirmed using InterProScan Perl-based version 4.5 (Zdobnov and Apweiler 2001) tools, which also associated functional information and Gene Ontology (GO; Ashburner et al. 2000) terms with our contigs. Results from InterProScan were also integrated into previous MySQL database for further comprehensive analysis.

KEGG Orthology (KO, Kanehisa et al. 2010) annotation and pathway identification were assigned and performed using the KOBAS (Wu et al. 2006) tool. Pathway distribution applied chi-square test and false discovery rate (FDR) correction, taking *P*-value < 0.05 and false positive < 1.

Comparative genomic analysis. For comparative genomic analysis, unique sequences were additionally aligned to genome sequences (downloaded in http://genome.ucsc.edu/ and http://www.jgi.doe.gov/) from diverse organisms, using specialized alignment programs GMAP (Wu and Watanabe 2005) to facilitate genomic mapping. We used Perl scripts to help us extract useful information from all raw results produced by the above tools.

RESULTS

Production and analysis of EST sequences. We constructed a normalized cDNA library from *D. salina,* and 2,282 high-quality ESTs were generated. To make a rich set of *D. salina* ESTs, we added 3,990 sequences from additional six EST libraries (see ''Additional EST libraries'' section) into our analysis. As Table 1 shows, the average sequence length was 519 nucleotides, and ~81.1% of the ESTs were no less than 400 nucleotides, while 0.8% of the sequences were shorter than 100 nucleotides (some ESTs libraries from dbEST had been preprocessed before they were submitted, so there were fewer short ESTs in the data set than raw ESTs). The total 6,272 ESTs were then put into a quality-control process including vector trimming, repeat masking, and

TABLE 1. Summary of the expressed sequence tag (EST) analysis.

No. of high-quality ESTs	6,272
No. of trimmed masked ESTs	6,118
No. of unique sequences	4,148
Average length of unique sequences (bp)	496
No. of contigs	691
No. of singletons	3,457
No. of unique sequences	4,148
BLASTX hits (UniProtKB 2009.10) (%)	56
BLASTX hits (Swiss-Prot 2009.10) (%)	38
Identification of IPR domains (InterProScan) (%)	45
Gene Ontology (GO) terms assigned (InterProScan) (%)	36

low-quality segment trimming. Then the derived trimmed masked 6,118 ESTs were further subjected to cluster and assembly procedure using the PAVE pipeline with default parameters. A total of 6,118 ESTs were assembled into 691 contigs, while 3,457 sequences remained as singletons (totaled 4,148 unique sequences).

Functional annotation of assembled sequences. We compared ESTs from our library with those obtained from NCBI dbEST database. There were 17.5% ESTs in our library that were identical with 30.0% *D. salina* ESTs in NCBI dbEST. The inconsistency showed that the normalization process decreased the EST redundancy.

After ESTs were assembled, we subjected all UniSeqs to BLASTx searches to provide a first pass annotation for the putative function. Among the 4,148 unique sequences (691 contigs and 3,457 singletons), 1,590 (38.3%) UniSeqs (407 contigs and 1,183 singletons) listed in Table S2 (in the supplementary material) had found putative homologies in UniProtKB/Swiss-Prot protein knowledgebase (Apweiler et al. 2010) using the BLAST program with a cutoff *E*-value of 1E-20. If we change the cutoff *E*-value to 1E-8 and combine UniProtKB/TrEMBL protein database, there would be 738 more UniSeqs matched. But 663 among these 738 UniSeqs were assigned ''Predicted protein.''

While a total of 2,328 unique sequences (56%) showed sequence similarity with at least one species with an *E*-value < 1E-8, results that could be matched with two plants (*Arabidopsis thaliana* and *Oryza sativa* subsp. *indica*), two animals (*Homo sapiens* and *Drosophila melanogaster*), and green algae (*C. reinhardtii* and Chlorophyta) were picked out and are illustrated in Figure 1. There were 858 unique sequences, which were found in genomes from both green algae and animals. And a much higher number of genes (1,483) were shared between green algae and plants. Furthermore, the number of unique sequences with similarity to the green alga and plant genomes but not the animal genomes was 710. However, the number of unique sequences with similar sequences in the green algae plus in either of the two animals, but not in the plant genomes, was 85. The *D. salina* unique sequences determined to have similar sequences only in green algae species totaled 602 (14.5%).

Then the presence of protein family, domain, and function site databases with our UniSeqs was confirmed. There were 1,883 (45.4%) UniSeqs (463 contigs and 1,420 singletons), which had found InterPro (Apweiler et al. 2001) signatures (Table S2), while 2,145 (51.7%) UniSeqs had been observed in putative protein family, domain, or functional site annotations, and 1,505 UniSeqs had been assigned GO terms to better characterize function category (Fig. 2).

Furthermore, we identified 106 KEGG pathways among all UniSeqs, and the result was compared with *A. thaliana* as a background distribution facilitated by the KOBAS tool. Pathways that were different between *D. salina* and *A. thaliana* were listed in Table S3 (in the supplementary material). Significantly enriched pathways mainly covered photosynthesis, carbon fixation in photosynthetic organisms, and porphyrin and chl metabolism, whereas processes like phenylpropanoid biosynthesis; stilbenoid, diarylheptanoid, and gingerol biosynthesis; limonene and pinene degradation; naphthalene and anthracene degradation; methane metabolism; and so forth were significantly less represented in our list.

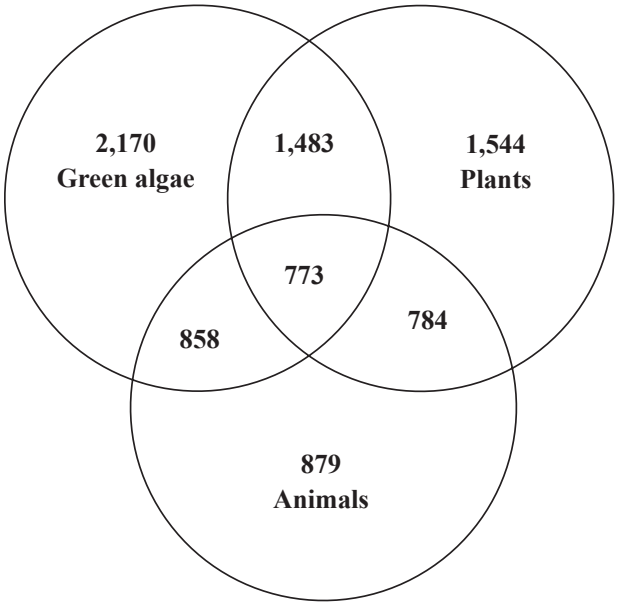

FIG. 1. Venn diagram of the *Dunaliella salina* unique sequences with similar sequences (4,148) in other green algae, plants, and animals. The numbers for the different taxa are shown at the level *E*-value < 1E-8.

Comparative analysis of the Dunaliella *transcriptome.* GMAP tool was used to study the presence of similarity in the collections of genome sequences from nine model organisms (*A. thaliana*, *C. reinhardtii*, cow, dog, human, *Ostreococcus tauri*, *V. carteri*, yeast, zebrafish). As illustrated in Table 2, unique sequences had much higher similarity with two green algae (*C. reinhardtii* and *V. carteri*) than other species. Yeast genome had the lowest similarity with *D. salina* unique sequences, while *A. thaliana* (a model plant organism) and *O. tauri* (a marine unicellular green alga of the Prasinophyceae clade) showed comparable similarity as the other four animal organisms did.

DISCUSSION

D. salina, the most-studied species in the genus *Dunaliella*, is increasingly being recognized as an emerging model system in abiotic stress tolerance studies (Cowan et al. 1992, Boetius and Joye 2009,

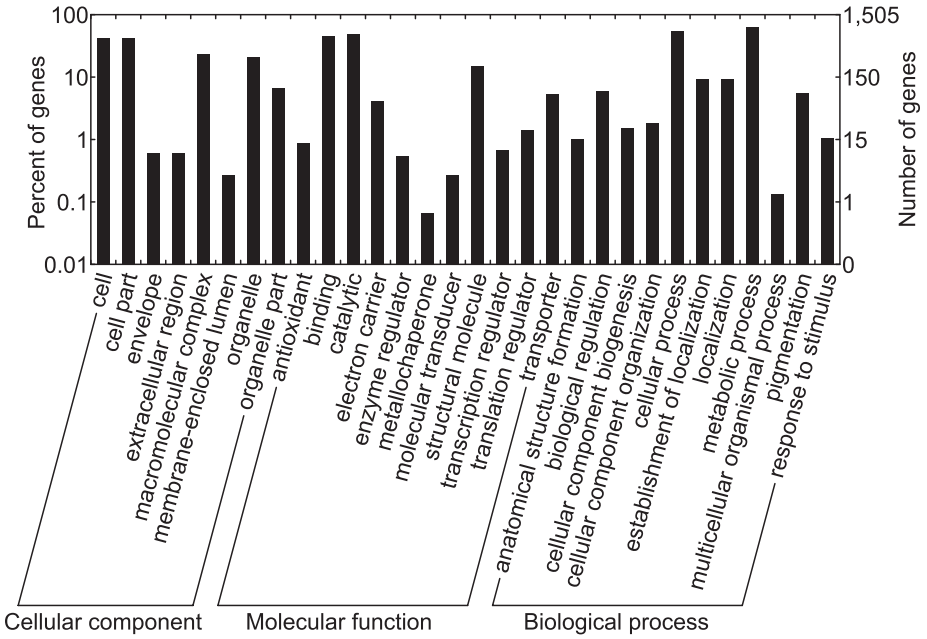

FIG. 2. Gene Ontology electronic annotation of UniSeqs from *Dunaliella salina*. Among the 1,505 UniSeqs, largest proportion of annotated UniSeqs was assigned as ''cell part'' (634) in ''Cellular component,'' ''catalytic'' (732)/''binding'' (675) in ''Molecular function,'' and ''metabolic process'' (949)/''cellular process'' (811) in ''Biological process.''

TABLE 2. Unique sequences were aligned to genome sequences from nine organisms. Each aligned UniSeq indicates at least one region of the organism's genome sequence could be match with the UniSeq using GMAP tool.

Organism	Aligned UniSeqs	%
Chlamydomonas reinhardtii	812	19.6
Volvox carteri	634	15.3
Dog	187	4.5
Human	175	4.2
Cow	169	4.1
Ostreococcus tauri	161	3.9
Arabidopsis thaliana	141	3.4
Zebrafish	140	3.4
Yeast	26	0.6
Any above	1,058	25.5

Chen et al. 2009, Tian and Yu 2009, Kim et al. 2010). In the following section, we discuss the major findings of our study and the applicability to ongoing and future research efforts in *D. salina* and beyond.

EST abundance. The expressed sequences resources presented here provide a valuable entry point for studies in *D. salina*. The 6,272 high-quality EST sequences from different libraries were assembled into 4,148 nonredundant sequences (contigs and singletons). The cross library assembly resulted in a sample of sequences derived from a wide range of biological functions.

Sequencing of random cDNA clones allows studies of mRNA abundance. Thus, analysis of the frequency of specific ESTs that form individual contigs can provide information with respect to the expression levels of particular genes under different experimental conditions. Table 3 displays the highly expressed genes (as represented by contigs made up of 20 or more ESTs together). Of the 14 highly expressed ESTs presented in Table 3, two were of unknown function, four were photosystem-related products, and six were ribosomal components. The data illustrate genes encoding ribosomal proteins and photosystem components dominating the frequency table. Among the first 100 most highly expressed genes shown in Table S2 (ds_0001 $\sim$ ds_0100, six with unknown function not listed here), 52 were ribosomal proteins, 26 were photosystem-related, 16 were for other genes, and six were unknown genes. As rRNA is extremely abundant and makes up 80% of the 10 mg $\cdot$ mL^{-1} RNA distributed in a typical eukaryotic cytoplasm (Kampers et al. 1996), ESTs from additional cDNA libraries such as Cushman, J. C.'s and EonSeon, Jin's (Park et al. 2006) did not apply normalization process that would represent a high level of rRNA and tRNA. So, rRNA genes here took a large fraction in highly expressed genes. Photosystem-related genes included many chloroplastic precursors. As a unicellular photosynthetic eukaryote, photosystem-related genes were more highly expressed than higher plants such as *A. thaliana* (Asamizu et al. 2000). We searched for the six unknown genes against UniProtKB/TrEMBL protein database. Two were matched with *C. reinhardtii* LciD gene (Wang and Spalding 2006) and *V. carteri* lciB gene (Nematollahi et al. 2006), one was predicted open reading frame in *C. reinhardtii*, and three still have no hit.

The D. salina *transcriptome.* We are investigating the transcriptome of *D. salina*, a green alga of great interest in biotechnology (Jin and Melis 2003, Stauber and Hippler 2004, Tafreshi and Shariati 2009). EST sequencing provides information about

TABLE 3. The most abundantly represented genes.

UniSeq ID	EST count	Annotation (homology in UniProtKB/Swiss-Prot)	*E*-value
ds_0001	45	Ubiquitin	1.0E-35
ds_0002	30	Chl *a*/*b*–binding protein of LHCII type I, chloroplastic precursor	7.0E-122
ds_0003	29	Unknown	N/A
ds_0004	26	Glyceraldehyde-3-phosphate dehydrogenase A, chloroplastic precursor; NADP-dependent glyceraldehydephosphate dehydrogenase subunit A	3.0E-167
ds_0005	26	Chl *a*/*b*–binding protein 1, chloroplastic precursor; LHCII type I CAB-1	1.0E-96
ds_0007	25	Light-harvesting complex I 17 kDa protein; P21 protein; PSI reaction center subunit III, chloroplastic precursor; PSI-F	2.0E-76
ds_0006	24	60S ribosomal protein L27a-3	3.0E-56
ds_0009	23	40S ribosomal protein S17-4	6.0E-46
ds_0008	22	40S ribosomal protein S9-2	3.0E-78
ds_0010	22	Unknown(contains cysteine peptidase active site)	N/A
ds_0016	22	60S ribosomal protein L10; EQM	2.0E-93
ds_0011	21	40S ribosomal protein S4	3.0E-106
ds_0012	20	60S ribosomal protein L11-2; L16	9.0E-79
ds_0013	20	Oxygen-evolving enhancer protein 1, chloroplastic precursor (PSII manganese-stabilizing protein PsbO)	3.0E-108

EST, expressed sequence tag.

functional identification of genes, gene structure, and gene expression patterns. As in other recent studies carried out in green algae (Shrager et al. 2003, Forster et al. 2006, Jain et al. 2007), a strategy of sequencing a variety of different libraries was used to maximize the number of unique genes.

Sequence redundancy within our study was 33.9%. This level of redundancy was lower than comparable studies (Stanley et al. 2005, Jain et al. 2007) carried out in *C. reinhardtii* and green alga *Ulva linza*: 93.6% and 41.8%. In the first one, 246,972 EST and cDNA sequences were collected from diverse sources. They had a separate assembly process, which is based on assembly of contiguous ESTs verified on genome (ACEGs). However, in the second one, a unique growth condition was used, and the number of sequenced ESTs was lower (1,898). This result is not comparable since it did not use multilibrary assembly strategy. Taking into account that the first one collected many more sequences than we did and sequence redundancy would increase when the total number of sequenced ESTs made an order-of-magnitude increase, we consider that our strategy, designed to maximize the number of unique genes without increasing in excess the number of sequences, worked properly.

BLASTx searches indicated that 56.1% of the unique sequences had sequence similarity (with *E*-value 1E-8) to at least one entry in the UniProtKB database. This percentage is higher than what has been observed in a similar study in *U. linza* (48.1%) (Stanley et al. 2005). New sequence data from complete genome projects of several algae species had made this increase possible, although a high proportion of the hits were annotated as hypothetical proteins. As also shown in this study, nearly half of the unique sequences could not be identified by database matches. This could be due to a number of reasons. One is that green algae may harbor a large number of unique proteins. Many studies (Bhattacharya and Medlin 1998, Turmel et al. 2002, Lewis and McCourt 2004, Merchant et al. 2007) showed that green algae genes could be traced to alga-plant or plant-animal common ancestors. It seems that the green algae may encode many unique proteins, but the low level of similarities detected could also be due to our lack of knowledge of these organisms. However, it could also be a feature of the library normalization procedure where rarer sequences and unidentified proteins have been enriched. It could also be because there are fewer sequences from green algae registered on the public databases compared with those of higher plants. In the Arabidopsis Genome Initiative 2000, 31% of the predicted gene products could not be assigned a function. Thirty-nine percent of the sequences of the green flagellate *Scherffelia dubia* showed no matches in the databases (Becker et al. 2001), and with *C. reinhardtii,* 70% of the nonredundant ESTs published by Asamiziu et al. (2000) contained no matches.

As seen in the results, in a number of cases, clones that did not cluster (considered as unique sequences) displayed sequence similarity to the same gene product. Several factors could account for this. Among them are that (i) the clones are different genes that are orthologous genes, which may share identical sequence region; (ii) the clones align with different regions of the same search hit but do not overlap (or have too small of an overlap) with each other; and (iii) the clones represent different splice variants of the same gene. This would also increase the level of unidentified proteins.

The degree of annotation was extended through the identification of protein motifs, using InterProScan searches of the InterPro database. This extension resulted in annotation of a 51.7% (2,145) of the unique sequences, including 1,883 with InterPro annotation. This percentage was similar to the BLASTx result.

According to the HMMPfam (http://hmmer.janelia.org/) scan results against Pfam database (http://pfam.sanger.ac.uk/), the five most predominant protein families (have >20 UniSeqs matches) are WD domain (G-beta repeat), mitochondrial carrier protein, ankyrin repeat, chl *a/b*–binding protein, and a helix-loop-helix structural domain found in a large family of calcium-binding proteins (EF hand). Three of them (WD domain, ankyrin repeat, and EF hand) were involved in signal transduction, and two were relevant to energy transfer processes. We inferred that *D. salina* had an active signaling system, which enables it to adapt quickly to a wide variety of salt concentrations. The state of active energy transfer did agree with the analysis of metabolism survey. Among the four significantly enriched pathways, all were important pathways involved in energy transfer. The photosynthetic production like glycerol then adapted *Dunaliella* to osmotic changes (Tafreshi and Shariati 2009).

Comparative genomic analysis. We were interested in determining how related the genome of *D. salina* is to genomes of other species. As Table 2 revealed, unique sequences aligned to two green algae genomes numbered many more than to other genomes, while yeast have obviously the least aligned unique sequences. However, it was unexpected that there were even more unique sequences aligned to animal genomes (dog, human, and cow) than *O. tauri* (a unicellular green alga) and *A. thaliana* (model plant), which did not agree with what Figure 1 shows—*D. salina* had more common sequences with plants than animals. This inconsistency may be due to the different methods conducted between these two analyses. Results in Figure 1 were based on BLASTx analysis between *D. salina* unique sequences and protein sequences from green algae, plants, and animals. However, the comparative genomic analysis was based on alignment between unique sequences and genome sequences from different species. As many studies (Bhattacharya and Medlin 1998, Turmel et al. 2002, Lewis and McCourt 2004, Merchant et al. 2007) have shown, green algae genes can be traced to alga-plant or plant-animal common ancestors. *Dunaliella* genes shared by animals and plants were derived from the last plant-animal common ancestor, and many of these had changed in their nucleotide sequence but reserved functional site in protein sequence because of codon degeneracy. This finding may explain why there would be more common sequences in BLASTx analysis than in genomic analysis. It also indicated that green algae may have a number of unique genes.

There were 812 (19.6%) unique sequences that could be aligned to *C. reinhardtii* genome, while 2,064 (49.8%) unique sequences found putative homologies in *C. reinhardtii*. It may be due to sequence divergence between orthologs in *D. salina* and *C. reinhardtii* or alignment sensitivity.

D. salina has already been a model organism for study of photosynthesis, β-carotene synthesizing, and salt adaptation. This study provides an extensive survey of *Dunaliella* transcriptome. This data set is a valuable resource that will be of great help for *Dunaliella* functional investigation, comparative genomics, and evolutionary studies. Recent studies on *Dunaliella* transcriptome (Park et al. 2006, Kim et al. 2010) and genome (Smith et al. 2010) have made an important contribution to the ongoing research. The challenge in the future will be the further expansion of our knowledge of the basic components of different algal systems. At the genome level, the complete genome sequence of *D. salina* will be made available soon; steady progress is being made with the ever-increasing number of algal genome projects being set up or completed.

This work was supported by the National Natural Science Foundation of China (30871321, 30771312, and 309718.17), national special basic research projects of China (SB2007FY400-4), National Basic Research Program of China (2009CB125910).

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

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29:37–40.

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., et al. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38:D142–8.

Asamiziu, 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.

Asamizu, E., Nakamura, Y., Sato, S. & Tabata, S. 2000. A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res.* 7:175–80.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25:25–9.

Baxendale, S., Chen, C. K., Tang, H. Z., Davison, C., Van Hateren, L., Croning, M. D. R., Humphray, S. J., Hubbard, S. J. & Ingham, P. W. 2009. Expression screening and annotation of a zebrafish myoblast cDNA library. *Gene Expr. Patterns* 9:73–82.

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.

Beldade, P., Rudd, S., Gruber, J. D. & Long, A. D. 2006. A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* 7:16.

Bhattacharya, D. & Medlin, L. 1998. Algal phylogeny and the origin of land plants. *Plant Physiol.* 116:9–15.

Boetius, A. & Joye, S. 2009. Thriving in salt. *Science* 324:1523–5.

Chen, H., Jiang, J. G. & Wu, G. H. 2009. Effects of salinity changes on the growth of *Dunaliella salina* and its isozyme activities of glycerol-3-phosphate dehydrogenase. *J. Agric. Food. Chem.* 57:6178–82.

Cowan, A. K., Rose, P. D. & Horne, L. G. 1992. *Dunaliella salina*: a model system for studying the response of plant cells to stress. *J. Exp. Bot.* 43:1535–47.

Fisher, M., Gokhman, I., Pick, U. & Zamir, A. 1996. A salt-resistant plasma membrane carbonic anhydrase is induced by salt in *Dunaliella salina*. *J. Biol. Chem.* 271:17718–23.

Forster, B., Mathesius, U. & Pogson, B. J. 2006. Comparative proteomics of high light stress in the model alga *Chlamydomonas reinhardtii*. *Proteomics* 6:4309–20.

Huang, X. & Madan, A. 1999. CAP3: a DNA sequence assembly program. *Genome Res.* 9:868–77.

Jain, M., Shrager, J., Harris, E. H., Halbrook, R., Grossman, A. R., Hauser, C. & Vallon, O. 2007. EST assembly supported by a draft genome sequence: an analysis of the *Chlamydomonas reinhardtii* transcriptome. *Nucleic Acids Res.* 35:2074–83.

Jin, E. S. & Melis, A. 2003. Microalgal biotechnology: carotenoid production by the green algae *Dunaliella salina*. *Biotechnol. Bioprocess Eng.* 8:331–7.

Kampers, T., Friedhoff, P., Biernat, J. & Mandelkow, E. M. 1996. RNA stimulates aggregation of microtubule-associated protein tau into Alzheimer-like paired helical filaments. *FEBS Lett.* 399:344–9.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38:D355–60.

Katz, A., Waridel, P., Shevchenko, A. & Pick, U. 2007. Salt-induced changes in the plasma membrane proteome of the halotolerant alga *Dunaliella salina* as revealed by blue native gel electrophoresis and nano-LC-MS/MS analysis. *Mol. Cell Proteomics* 6:1459–72.

Kijimoto, T., Costello, J., Tang, Z. J., Moczek, A. P. & Andrews, J. 2009. EST and microarray analysis of horn development in *Onthophagus* beetles. *BMC Genomics* 10:13.

Kim, M., Park, S., Polle, J. E. & Jin, E. 2010. Gene expression profiling of *Dunaliella* sp. acclimated to different salinities. *Phycol. Res.* 58:17–28.

Lamers, P. P., Janssen, M., De Vos, R. C. H., Bino, R. J. & Wijffels, R. H. 2008. Exploring and exploiting carotenoid accumulation in *Dunaliella salina* for cell-factory applications. *Trends Biotechnol.* 26:631–8.

Lewis, L. A. & McCourt, R. M. 2004. Green algae and the origin of land plants. *Am. J. Bot.* 91:1535–56.

Li, G., Liu, M., Jiang, Y., Qiao, D. R. & Cao, Y. 2004. Construction and functional gene screening of cDNA library of *Dunaliella salina*. *J. Trop. Subtrop. Bot.* 12:5.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. & Quackenbush, J. 2000. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* 28:3657–65.

May, P., Christian, J. O., Kempa, S. & Walther, D. 2009. ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10:11.

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318:245–51.

Murthy, K. N. C., Vanitha, A., Rajesha, J., Swamy, M. M., Sowmya, P. R. & Ravishankar, G. A. 2005. In vivo antioxidant activity of carotenoids from *Dunaliella salina* – a green microalga. *Life Sci.* 76:1381–90.

Nematollahi, G., Kianianmomeni, A. & Hallmann, A. 2006. Quantitative analysis of cell-type specific gene expression in the green alga *Volvox carteri*. *BMC Genomics* 7:19.

Oren, A. 2005. A hundred years of *Dunaliella* research: 1905–2005. *Saline Syst.* 1:2.

Park, S., Polle, J. E. W., Melis, A., Lee, T. K. & Jin, E. S. 2006. Up-regulation of photoprotection and PSII-repair gene expression by irradiance in the unicellular green alga *Dunaliella salina*. *Mar. Biotechnol.* 8:120–8.

Shrager, J., Hauser, C., Chang, C. W., Harris, E. H., Davies, J., McDermott, J., Tamse, R., Zhang, Z. D. & Grossman, A. R. 2003. *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 131:401–8.

Smith, D. R., Lee, R. W., Cushman, J. C., Magnuson, J. K., Tran, D. & Polle, J. E. W. 2010. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biol.* 10:14.

Soderlund, C., Johnson, E., Bomhoff, M. & Descour, A. 2009. PAVE: program for assembling and viewing ESTs. *BMC Genomics* 10:10.

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.

Stauber, E. J. & Hippler, M. 2004. *Chlamydomonas reinhardtii* proteomics. *Plant Physiol. Biochem.* 42:989–1001.

Tafreshi, A. H. & Shariati, M. 2009. *Dunaliella* biotechnology: methods and applications. *J. Appl. Microbiol.* 107:14–35.

Tian, J. Y. & Yu, J. 2009. Changes in ultrastructure and responses of antioxidant systems of algae (*Dunaliella salina*) during acclimation to enhanced ultraviolet-B radiation. *J. Photochem. Photobiol. B Biol.* 97:152–60.

Turmel, M., Otis, C. & Lemieux, C. 2002. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol. Biol. Evol.* 19:24–38.

Vizcaino, J. A., Gonzalez, F. J., Suarez, M. B., Redondo, J., Heinrich, J., Delgado-Jarana, J., Hermosa, R. et al. 2006. Generation, annotation and analysis of ESTs from *Trichoderma harzianum* CECT 2413. *BMC Genomics* 7:14.

Wang, Y. & Spalding, M. H. 2006. An inorganic carbon transport system responsible for acclimation specific to air levels of CO_2 in *Chlamydomonas reinhardtii*. *Proc. Natl. Acad. Sci. U. S. A.* 103:10110–5.

Wu, J. M., Mao, X. Z., Cai, T., Luo, J. C. & Wei, L. P. 2006. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34:W720–4.

Wu, T. D. & Watanabe, C. K. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–75.

Ye, Z. W., Jiang, J. G. & Wu, G. H. 2008. Biosynthesis and regulation of carotenoids in *Dunaliella*: progresses and prospects. *Biotechnol. Adv.* 26:352–60.

Zdobnov, E. M. & Apweiler, R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–8.

Supplementary Material

The following supplementary material is available for this article:

Table S1. Expressed sequence tag (EST) libraries that were used in our analysis. All data are available from dbEST at NCBI.

Table S2. Assembled unique sequences were assigned UniProt Protein Ids and InterPro Numbers.

Table S3. Significantly different distribution of gene pathways in *Dunaliella salina* and *Arabidopsis thaliana*.

This material is available as part of the online article.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.