A partial expressed sequence tag (EST) library of the economically important red alga *Eucheuma denticulatum* (N. L. Burham) F. C. Collins and Hervey

Paulina S. Aspilla[1,2], Anna Angela Camille B. Antonio[1], Giuseppe C. Zuccarello[3], Nina Rosario L. Rojas[1,*]

[1] Department of Chemistry, School of Science and Engineering, Ateneo de Manila University, Loyola Heights, Quezon City, Philippines 1108

[2] Department of Chemistry, Silliman University, Dumaguete City, Philippines

[3] School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington, 6140 New Zealand

A library of expressed sequence tags (ESTs) was derived from the economically important Rhodophyta species *Eucheuma denticulatum*. This small scale EST library represents the first look at the set of genes expressed in any *Eucheuma* species. A total of 311 clones were analyzed. These 311 sequences clustered into 143 unigenes. Thirty-two of the 143 unigenes showed sufficient similarity to known genes to allow annotation. Of these 32 unigenes, 22 were found in at least one other EST collection from Rhodophyta species. Sixteen of the unigenes showed strong similarity to known genes, with e-values smaller than $1x10^{-14}$. Another 16 unigenes could be annotated by relaxing the e-value cut-off to $5x10^{-4}$. The identified genes covered general metabolic processes (10); translation (2); protein folding, transport, and degradation, including heat shock proteins (6); DNA and RNA binding (2); transcription factors (2); signal transduction, and protein binding (2); cell cycle and apoptosis (2); ion transport (1); and cell membrane, cell wall or extracellular matrix-associated proteins, and other structural proteins (5). Most of the unigenes (111 of them, representing 77.6% of the unigenes), showed no significant similarity to known genes, although 15 of the 111 showed similarity to hypothetical or predicted proteins or raw genomic sequences. Majority of the unannotated unigenes (93 or 64.6% of the library) have not been found in the other Rhodophyta EST libraries, suggesting that many novel genes may still be uncovered by small scale EST collections.

KEYWORDS

cDNA library; Rhodophyta; carrageenophyte; expressed sequence tag; gene expression; unigene

*Corresponding author
Email Address: nrojas@ateneo.edu
Submitted: February 28, 2010
Revised: May 11, 2010
Accepted: May 11, 2010
Published: May 28, 2010
Editor-in-charge: Eduardo A. Padlan

INTRODUCTION

The red alga *Eucheuma denticulatum* (N. L. Burham) F. C. Collins and Hervey (Gigartinales, Rhodophyta), which is a source of iota-carrageenan, is one of the most economically important aquatic plants in the world. The bulk (93%) of the total world demand for aquatic plants is supplied by aquaculture (FAO Fisheries and Aquaculture Department, 2008). While China is the world leader in aquatic plant production, supplying 10.9 billion tons in 2006, the Philippines was 2nd, producing 1.5 million tons, followed by Indonesia (0.91 tons), South Korea (0.77 million tons) and Japan (0.49 million tons) (FAO Fisheries and Aquaculture Department, 2008). Exports of farmed seaweeds and other algae products contributed US$ 72.3 million to the Philippine economy in 2005, with US$ 43.3 million coming from carrageenan exports (DA-AMAS, 2005).

The polysaccharide content of *E. denticulatum* and its ecology and farming practices have been well studied (Doty, 1987). Its cell-wall biochemistry has also been well studied over the years (Kloareg & Quatrano, 1988) but few gene sequences are known of *E. denticulatum*. The only nucleotide entries that are publicly available are for regions commonly used for phylogenetic analysis: cytochrome oxidase genes (*cox*1 and *cox*2); ribulose-1,5-bisphosphate carboxylase/oxygenase genes (*rbc*L and *rbc*S); ribosomal RNA genes (23S, 18S); and several partial plastid genes and spacers (URP1, RuBisCo spacer) (Benson et al. 2008). Beyond studies of the evolutionary relationships of *E. denticulatum* (Lluisma and Ragan, 1995, Fredericq et al. 1999, Zuccarello et al. 2006), DNA data can help in our understanding of this alga, which may prove to be useful in aquaculture practices. Highly variable markers (microsatellites, single nucleotide polymorphisms or SNPs) may make artificial selection on improved cultivar characters more tractable. Identification of genes important in physiological process and biochemical outcomes may help our understanding of the process involved in the polysaccharide chemistry of *E. denticulatum* and other commercially important algal species.

In this study, we used the expressed sequence tag approach to exploring *E. denticulatum* genes. The genes of an organism can be probed directly at the genomic level or at the level of expression, through RNA or protein. Despite major advances in DNA sequencing technology, full genome sequencing is still not available for *Eucheuma* or its relatives. On the other end, direct protein sequencing, such as via mass spectrometric techniques, has gained much ground but still cannot compare with the high throughput of gene sequencing. A more accessible approach to exploring an organism's sequence space is using expressed sequence tags, or ESTs. These ESTs are short sequences derived from RNA molecules expressed in the tissues, and thus represent the working repertoire of genetic information (Boguski et al., 1993). Aside from accessing gene sequences, analysis of the mRNAs present in the organism provides a window into its physiology, by observing the set of genes transcribed and translated into protein.

The marine red alga *Gracilaria gracilis* (Stackhouse) Steentoft, Irvine & Farnham was the first alga to be studied using the EST approach (Lluisma and Ragan, 1997). The authors reported that out of 200 ESTs in their library, they identified genes for carbohydrate metabolism (7), amino acid metabolism (3), photosynthesis (5), nucleic acid synthesis, repair and processing (3), protein synthesis (14), protein degradation (6), cellular maintenance and stress response (3), and other identifiable protein-coding genes (13). A majority of their ESTs, 146 sequences, did not have significant matches in the sequence database at that time, suggesting the large number of novel genes that can be discovered through this approach. Since then, several thousand ESTs for red alga have been deposited in the dbEST database (www.ncbi.nlm.nih.gov/dbEST/) and described in the literature.

A summary of EST libraries from Rhodophyta species in the dbEST database as of June 2009 is shown in Table 1: the carrageenophyte *Chondrus crispus* Stackhouse (Collén et al., 2006); the agarophyte Gracilaria: *Gracilaria changii* (B.M. Xia & I. Abbott) I. Abbott, J. Zhang & B.M. Xia (Teo et al., 2007), *Gracilaria gracilis* (Lluisma and Ragan, 1997), *Gracilaria lemaneiformis* (Bory de Saint Vincente) Greville (Sun et al., 2002); *Griffithsia okiensis* Kajimura (Lee et al., 2007); *Porphyra haitanensis* T. J. Chang & B.F. Zheng Baofu (Fan et al., 2007), and *Porphyra yezoensis* Ueda (Nikaido et al., 2000). In addition to the ESTs described in the publications, additional sequences have been added to dbEST for these species, from the same research teams as well as from other laboratories.

These EST libraries represent different degrees of relation to *E. denticulatum* based on Algaebase (Guiry & Guiry, 2009), as noted in Table 1.Both *E. denticulatum* and *Chondrus crispus* are under order Gigartinales. Together with *Gracilaria* (order Gracilariales) and *Griffithsia* (order Ceremiales), these fall under Class Florideophyceae: Subclass Rhodymeniophycidae. *Porphyra* is the least related to *E. denticulatum*, belonging to Class Bangiophyceae.

In this study, we present the first analysis of an EST library from *E. denticulatum.* Despite the thousands of EST entries from Rhodophyta species in dbEST, we show that this small scale library can yield new information on genes expressed in red algae. Aside from annotating the sequences to glimpse which genes are actively expressed in the organism, we also compare our sample of expressed genes with the sequence space covered by the larger-scale red algal EST collections available to see if we could uncover additional novel algal genes.

MATERIALS AND METHODS

To generate the library, *E. denticulatum (*green variety) was collected from a seaweed farm in Negros Oriental, Philippines. The thalli (5-cm pieces taken from the tips) were washed and frozen in liquid nitrogen.

Table 1. *Eucheuma denticulatum* vs. other Rhodophyta species with sequence data stored at dbEST as of June 23, 2009 (http://www.ncbi.nlm.nih.gov/dbEST/) (Boguski et al, 1993). Taxonomy from http://www.algaebase.org (Guiry & Guiry, 2009).

Class	Order	Family	Species name	Number of EST sequences	Main reference
Florideophyceae	Gigartinales	Areschougiaceae[a]	*E. denticulatum*	311[b]	This work
		Gigartinaceae	*Chondrus crispus*	4,114	Collén et al., 2006
	Gracilariales	Gracilariaceae	*Gracilaria changii*	8,147	Teo et al., 2007
			Gracilaria gracilis	200	Lluisma and Ragan, 1997
			Gracilaria lemaneiformis	178	Sun et al., 2002
	Ceremiales	Wrangeliaceae	*Griffithsia okiensis*	1,274	Lee et al., 2007
Bangiophyceae	Bangiales	Bangiaceae	*Porphyra haitanensis*	6,035	Fan et al., 2007
			Porphyra yezoensis	22,069	Nikaido et al., 2000

[a] Solieriaceae in NCBI Taxonomy

[b] The sequences are stored in dbEST under ID numbers 69655260-69655569 and 69657818. The corresponding GenBank accession numbers are GW915368-GW915677 and GW917926. The sequence clearly identified as a segment of 18S rRNA is also deposited as HM235648 in GenBank.

The cDNA library construction was performed by American Gene C.T. LLC (Cranston, Rhode Island USA) as follows: total RNA was extracted and enriched for mRNA via polyA-tail selective binding; cDNA was synthesized via RT-PCR and then cloned into pDNR-LIB (Clontech Laboratories, Inc., Madison, WI, USA).

The cDNA collection that was obtained was transformed into DH10BT electrocompetent cells (Invitrogen, Carlsbad, California, USA) by electroporation in a 0.2 mm cuvette according to manufacturer's specifications (Life Technologies Cell Porator, Carlsbad, California, USA). A total of 400 clones were randomly picked from the transformed cells. Bacterial plasmid DNA was extracted by lysis with alkali according to the procedure of Sambrook and Russell (2006). Single-pass sequencing of the plasmids using M13 forward primers was performed by Macrogen DNA Sequencing Service (Seoul, Korea).

The sequences were assessed for quality. Ambiguous nucleotide readings, sequences with low complexity regions, and vector contamination were edited out. Short sequences (generally those with fewer than 200 bp, unless of good quality sequence readings) were also excluded. Sequences that passed quality control were compared to sequences available at NCBI and the UniProt databases on several levels (UniProt Consortium, 2009; Jain et al, 2009).

Comparison with the other Rhodophyta EST libraries was performed using BLASTn, for nucleotide-nucleotide matches, and tBLASTx, for more distant putative protein-putative protein matches (Altschul et al., 1997). To compare the possible translated proteins from the EST nucleotide sequences against known protein sequences, BLASTx was used against the UniProt databases and the RefSeq protein database of the NCBI, filtered to exclude putative or hypothetical sequences. More distant similarity with translations of other nucleotide sequences was assessed using tBLASTx against sequences in the NCBI's subset of GenBank+EMBL+DDBJ+PDB sequences, excluding ESTs, short genomic landmark sequences or sequence tagged sites (STS), genome survey sequences (GSS), environmental samples, or phase 0, 1 or 2 high throughput genome sequencing (HTGS) output.

For the similarity searches described above, a match was considered good if the e-value was less than $1x10^{-14}$, and considered weak if the e-value was greater than $1x10^{-14}$ but less than $5x10^{-4}$. In this paper, the term similar, unless specified, is used for both good and weak matches. Matches with e-values greater than $5x10^{-4}$ were considered too distant to be useful.

Since a single gene may be expressed as several copies of RNA sequences and cDNA synthesis generally captures only fragments of sequences, the ESTs were sorted and assembled into so-called unigenes, which represent single genes. This was performed with the help of Contig Express of the Vector NTI suite using default parameters (Invitrogen, Carlsbad, California, USA) and visually verified using ClustalX (Larkin et al, 2007). ESTs that clustered into unigenes were essentially identical sequences in the regions where they overlapped, although occasional single nucleotide differences were allowed in order to account for limitations in fidelity of the PCR amplification process and the sequencing process. Overall, about 90% identity, and often over 95% identity, was observed in the aligned regions.

Gene ontology (GO) terms for members of the EST library were inferred from GO annotations of UniProt entries with highest sequence similarity (Camon et al., 2004; Gene Ontology Consortium, 2004). Hypothetical or predicted proteins were excluded as noted above. These GO annotations were compared with automated GO annotation by ESTPiper, which included hypothetical and predicted proteins (Tang et al., 2009).

Table 2. Summary of *Eucheuma denticulatum* EST analysis results.

Description	Number	Percentage
Cluster summary		
Total unigenes	143	100%
Contigs	72	50.3%
Singletons	71	49.7%
Overlap with other Rhodophyta ESTs		
Unigenes with matches within Rhodophyta ESTs	40	28.0%
Nucleotide matches (BLASTn)	19	13.3%
Translated protein-translated protein (tBLASTx)	40	28.0%
Annotated genes with similarity to other Rhodophyta ESTs	22	15.4%
Unannotated genes with similarity to other Rhodophyta ESTs	18	12.6%
Annotation of unigenes		
Unigenes with annotation	32	22.4%
High probability annotation (e-value $<1x10^{-14}$)	16	11.2%
Loose annotation ($1x10^{-14}$ < e-value < $5x10^{-4}$)	16	11.2%
Unigenes with no annotation or with hits to only hypothetical proteins or genomic sequences	111	77.6%

RESULTS

Starting with sequences from 400 random clones (designated C001 to C400), those with ambiguous nucleotide readings, sequences with low complexity regions, and vector contamination were edited out. Studies indicate that overall sequence quality tends to be high at the middle of a sequence and of lower quality at approximately 50-100 nucleotides at the beginning and end, so short sequences below 200 were also excluded from the final collection despite their possible information content (Nagaraj et al., 2006). However, two clones that showed unambiguous sequence traces were included in the analysis (C161 and C162). The final cDNA library contains a total of 311 sequences which were further analyzed for annotation and for similarity to other Rhodophyta EST libraries.

For the 311 sequences in the collection, the average length was 544 bp, with a median of 610 bp. The longest sequence was 704 bp (C032). The shortest sequences (C161 and C162) were 169 bp, as noted previously. All the rest were longer than 213 bp. A large fraction of the sequences ended with runs of adenylates, with 118 clones ending with 20 or more A's.

Since ESTs represent mRNA in the organism, multiple transcripts of a single gene may occur. This redundancy was resolved by assembling the 311 sequences into unique clusters or unigenes. The EST collection was coded as follows: single clone unigenes were designated C followed by the clone number, e.g., C186. Unigenes spanning two or more clones were designated U followed by the lowest-numbered clone in the cluster, e.g. U223 is the unigene derived from C223, C357, and C365 in the collection.

Table 2 summarizes the statistics for the unigene clusters. Out of the 311 sequences, 71 or about 23% of them are single, unique sequences, while the rest fall into 72 unique clusters for a total of 143 unigenes. One unigene (U005) was represented 47 times in the EST library. This highly expressed unigene has neither nucleotide nor translated polypeptide sequence matches with the Rhodophyta EST database. This unigene also showed no translated peptide matches with the full non-redundant protein database, suggesting that it is a novel gene. U005 is worth investigating further for taxon-specific information. The rest of the unigenes are comprised of seven or fewer individual EST sequences, with a median of two sequences per unigene.

The pie chart in Figure 1 gives an overview of the 143 unigenes. Majority (93 or 65%) of the unigenes could not be matched with known genes or with existing Rhodophyta EST libraries, suggesting they are novel. Another 18 (13%) could not be annotated by similarity with known genes but could be matched with ESTs from other Rhodophyta species. This brings a total of 78% of genes that could not be annotated. Among the 32 genes (22%) annotated by comparison with known genes, 22 (15%) had corresponding entries in other Rhodophyta species while 10 (7%) did not.

Looking closer at the similarities to other Rhodophyta genes, only 19 unigenes show similarity at the nucleotide level with the Rhodophyta ESTs (Table 2). Expanding the similarity

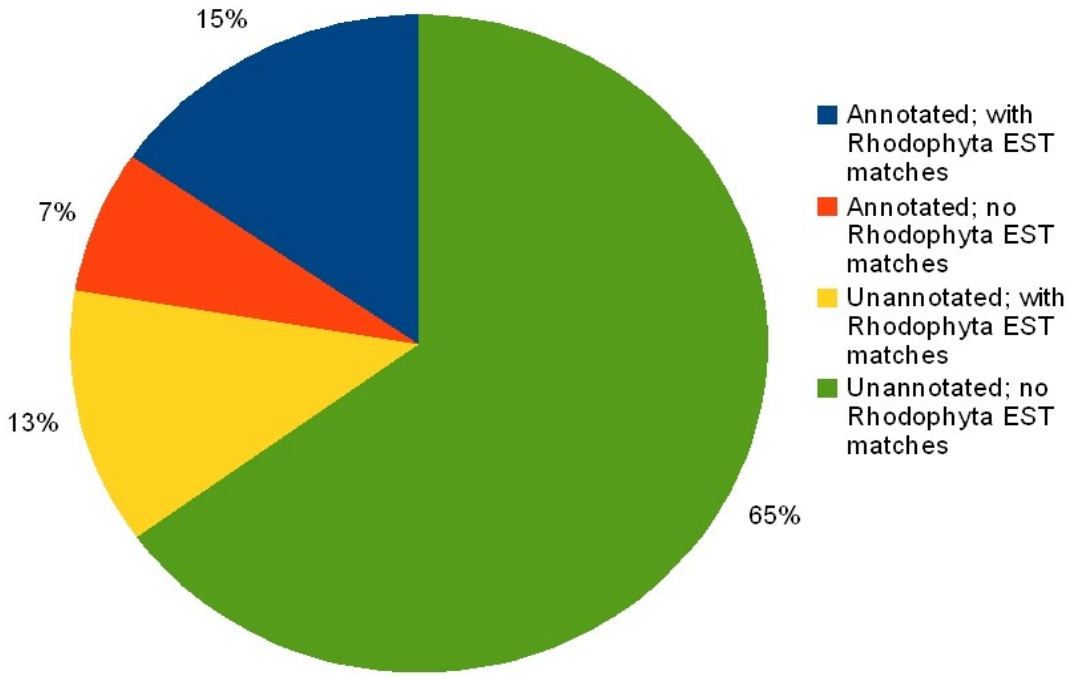

Figure 1. Annotation breakdown of the EST library of 143 unigenes.

search to putative translations into protein using tBLASTx produced 40 hypothetical matches at the protein sequence level (Table 2). Among the unigenes that showed similarity to Rhodophyta ESTs, three (C037, U024, and U108) had matches to only *Chondrus crispus* ESTs, suggesting that these unigenes may be specific to the carrageenophytes. None of the three could be annotated by similarity to other known proteins.

Annotation by similarity searching was done on two levels. Of the 143 unigenes, only 16 had high probability matches with known proteins (excluding putative or hypothetical proteins), with e-values smaller than 10^{-14}. Relaxing the match to include those with matches with e-values larger than 10^{-14} but smaller than 5 x 10^{-4} yielded an additional 16 matches, for a total of 32 unigenes with some functional annotation by similarity with known proteins. As the e-value increases, the reliability of the annotation generally decreases. However, given the relative paucity of known genes from related organisms, use of weak similarities may be necessary to uncover homologous genes from more distantly-related organisms. Table 3 lists these annotated unigenes sorted into functional categories. Those with corresponding matches in the other Rhodophyta EST libraries are noted as well.

The results indicate that only 22.4% of our small-scale EST library could be annotated by similarity to known proteins. Of the 32 annotated genes, 22 are similar to sequences already in the EST databases of other Rhodophyta species. A single clone in our collection, C186, is the 18S rRNA gene of *E. denticulatum*. It matches with positions 1086-1774 of the 1777 nucleotides of GenBank entry U25439 from *E. denticulatum* with identity of 685/689 nucleotides (99.4%). It is also highly similar to the homolog from *Eucheuma isiforme* (GenBank U25438), as well as to other 18S rRNA sequences from other red algal species. This 18S rRNA sequence was the only match in our collection with known *Eucheuma* genes.

Translation and general metabolic processes are well-represented in the unigene collection. Besides the 18S rRNA, the 60S ribosomal protein (e-value=4x10^{-21}) and an asparaginyl-tRNA synthetase (e-value=4x10^{-53}) were also identified as genes expressed in relation to translation. Metabolic proteins such as glyceraldehyde 3-phosphate dehydrogenase (e-value=1x10^{-37}) and pyrophosphate-dependent phosphofructokinase (e-value=2x10^{-10}), ATP-citrate lyase (e-value=4x10^{-10}), cysteine desulfurase (e-value=9x10^{-28}), adenylate kinase (e-value=6x10^{-14}), aldo/keto reductase (e-value=3x10^{-13}) and peptide-methionine sulfoxide reductase (e-value=2 x 10^{-67}) were also represented in the collection.

Some genes with possible regulatory roles were also found in the collection. These include DNA and RNA binding proteins, such as zinc finger protein (e-value=3x10^{-18}) and a reverse-transcriptase-like protein fragment (e-value=3x10^{-18}). Other possible

Table 3. Tentative gene annotation for *Eucheuma denticulatum* ESTs. All matched entries are from UniProt unless otherwise noted.

Unigene ID	Length (bp)	No. of clones	dbEST ID	GenBank Accn No.	Gene, organism, and entry ID of best match	E-value	GO term	Chondrus EST	Gracilaria EST	Griffithsia EST	Porphyra EST
General metabolic processes											
U345	592	2	69655525, 69655531	GW915633, GW915639	Glyceraldehyde-3-phosphate dehydrogenase, phosphorylating (*Chondrus crispus*, P34920)	Good: $1x10^{-37}$	Glycolysis	Y	Y	Y	Y
U146	472	5	69655366, 69655371, 69655372, 69655402, 69655403	GW915474, GW915479, GW915480, GW915510, GW915511	Pyrophosphate-dependent phosphofructo-kinase (*Medicago truncatula*, Q2HTG9)	Weak: $2x10^{-10}$	Glycolysis		Y		
C051	565	1	69655294	GW915402	Aldo/keto reductase (*Frankia* sp. Strain EAN1pec, A8KXW6)	Weak: $3x10^{-13}$	Oxidation-reducation		Y		
U204	524	2	69655415	GW915523	Asparaginyl-tRNA synthetase, cytoplasmic (Atlantic salmon, *Salmo salar*, B5X340)	Good: $4x10^{-33}$	Amino acid metabolism				Y
U056	774	2	69655298, 69655312	GW915406, GW915420	Cysteine desulfurase (*Ajellomyces capsulata* (strain ATCC 26029), C0NZZ7)	Good: $9x10^{-28}$	Mo-molybdopterin cofactor biosynthetic process				
U147	701	2	69655367, 69655368	GW915475, GW915476	Hemolysin-type calcium-binding region protein (*Lyngbya sp.* PCC 8106, A0YRU7)	Good: $2x10^{-26}$	Calcium ion binding				
C393	613	1	69655561	GW915669	Peptide methionine sulfoxide reductase (*Gracilaria verrucosas*, Q9XHG1)	Good: $2x10^{-67}$	Cell redox homestasis; protein metabolic process	Y	Y	Y	
U086	854	6	69655319, 69655362, 69655363, 69655434, 69655442, 69655518	GW915427, GW915470, GW915471, GW915542, GW915550, GW915626	Serpin peptidase inhibitor (*Clostridium thermocellum* strain ATCC 27405 / DSM 1237, A3DBV3)	Good: $3x10^{-18}$	Serine-type endopeptidase inhibitor activity				
C311	673	1	69655498	GW915606	Adenylate kinase (chloroplast) (*Zea mays*, P43188)	Weak: $6x10^{-14}$	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	Y	Y		

Table 3. continued.

Unigene ID	Length (bp)	No. of clones	dbEST ID	GenBank Accn No.	Gene, organism, and entry ID of best match	E-value	GO term	Chondrus EST	Gracilaria EST	Griffithsia EST	Porphyra EST
U161	169	2	69655377, 69655378	GW915485, GW915486	ATP-citrate lyase (*Griffithsia japonica*, Q7XY52)	Weak: $4x10^{-10}$	Lyase activity			Y	
Translation											
C081	340	1	69655314	GW915422	60S ribosomal protein (*Oryza sativa* Japonica, Q0DK10)	Good: $4x10^{-21}$	Translation	Y	Y		Y
C186	692	1	69655401	Raw sequence as GW915509; with annotation as HM235648	18S ribosomal RNA (*E. denticulatum*, GenBank U25439.1)	Good: $8x10^{-151}$	Translation	Y	Y	Y	Y
Protein folding, transport, and degradation											
C085	508	1	69655318	GW915426	DnaJ chaperone protein (*Bos taurus*, Q5BIP8.1)	Weak: $6x10^{-07}$	Protein folding		Y		Y
C376	250	1	69655550	GW915658	Heat shock protein HSP90 (*Griffithsia japonica*, Q9ZT55)	Weak: $7x10^{-06}$	Protein folding	Y	Y	Y	
U015	540	4	69655269, 69655359, 69655445, 69655485	GW915377, GW915467, GW915553, GW915593	Heat shock protein 70 (*Fucus serratus*, B3VMZ7)	Good: $4x10^{-16}$	Stress response	Y	Y	Y	
C064	698	1	69655302	GW915410	Vacuolar protein sorting associated protein (*Rattus norvegicus*, GenBank NP_001101476.1)	Weak: $5x10^{-04}$	Protein localization; protein transport				Y

Table 3. continued.

Unigene ID	Length (bp)	No. of clones	dbEST ID	GenBank Accn No.	Gene, organism, and entry ID of best match	E-value	GO term	Chondrus EST	Gracilaria EST	Griffithsia EST	Porphyra EST
U096	508	6	69655326, 69655331, 69655332, 69655334, 69655335, 69655385	GW915434, GW915439, GW915440, GW915442, GW915443, GW915493	Sonic hedgehog protein (*Griffithsia japonica*, Q7XZ18)	Good: $2x10^{-17}$	Proteolysis	Y	Y	Y	Y
U003	700	6	69655262, 69655311, 69655407, 69655412, 69655429, 69655469	GW915370, GW915419, GW915515, GW915520, GW915537, GW915577	Ubiquitin-conjugating enzyme family protein (*Oryza sativa* Japonica, Q2QLM1)	Good: $1x10^{-30}$	Ubiquitin-dependent protein catabolic process	Y	Y	Y	Y
DNA and RNA binding											
U014	596	2	69655268, 69655560	GW915376, GW915668	Zinc finger protein (*Culex quinquefasciatus*, B0WC55)	Good: $3x10^{-18}$	DNA binding motif	Y	Y	Y	Y
C007	672	1	69655264	GW915372	Reverse transcriptase-like protein fragment (*Ectropia obliqua*, B1PN78)	Weak: $5x10^{-07}$	RNA binding; RNA-directed DNA polymerization				
Transcription factors											
C033	698	1	69655281	GW915389	Transcription factor IWS1 (*Ustilago maydis*, Q4P7X6)	Good: $9x10^{-17}$	Regulation of DNA-dependent transcription		Y		Y
C198	369	1	69655410	GW915518	Histone-specific transcription factor (*Triticum aestivum*, P23922)	Weak: $1x10^{-04}$	Regulation of DNA-dependent transcription	Y	Y	Y	Y
Signal transduction and protein binding											
U059	701	6	69655300, 69655409, 69655414, 69655436, 69655459, 69655466	GW915408, GW915517, GW915522, GW915544, GW915567, GW915574	SNF1 family protein kinase (*Arabidopsis thaliana*, Q9ZRA0)	Good: $1x10^{-32}$	Protein serine/ threonine kinase activity	Y	Y	Y	Y

Table 3. continued.

Unigene ID	Length (bp)	No. of clones	dbEST ID	GenBank Accn No.	Gene, organism, and entry ID of best match	E-value	GO term	Chondrus EST	Gracilaria EST	Griffithsia EST	Porphyra EST
U170	699	2	69655386, 69655387	GW915494, GW915495	BRCA-1 associated protein (*Micromonas sp. RCC299*, C1FD36)	Weak: $5x10^{-5}$	Protein binding; zinc ion binding				
Cell cycle and apoptosis											
U387	695	2	69655556, 69655563	GW915664, GW915671	Rootletin (Ciliary rootlet coiled-coil protein) (*Homo sapiens*, Q5TZA2)	Weak: $7x10^{-08}$	Cell cycle; cell projection organization				
C083	555	1	69655316	GW915424	Transmembrane BAX inhibitor motif-containing protein (*Rattus norvegicus*, P55062)	Good: $5x10^{-20}$	Apoptosis				Y
Ion transport											
C344	631	1	69655524	GW915632	Voltage gated chloride channel domain-containing protein (*Toxoplasma gondii* ME49, B6KLS9)	Weak: $1x10^{-13}$	Chloride transport				
Cell wall, extracellular matrix, and other structural proteins											
U018	562	3	69655272, 69655273, 69655276	GW915380, GW915381, GW915384	Cell wall protein DAN4/ delayed anaerobic protein 4 precursor (*Saccharomyces cerevisiae*, P47179)	Weak: $6x10^{-05}$	Anchored to membrane; extracellular region; cell wall				
U375	698	2	69655549, 69655554	GW915657, GW915662	Cell wall-plasma membrane linker protein (*Brassica napus*, Q39353)	Weak: $9x10^{-09}$	Lipid transport				
C050	701	1	69655293	GW915401	Proline-rich glycoprotein (*Chlamydomonas reinhardtii*, Q9FPQ6)	Weak: $6x10^{-07}$	Structural constituent of cell wall	Y	Y		
C178	540	1	69655393	GW915501	Spondin-2 extracellular matrix protein (*Dario rerio*, B3DGW3)	Weak: $8x10^{-05}$	Cell adhesion	Y	Y		
U340	544	2	69655522, 69655528	GW915630, GW915636	Flagellar associated protein (*Chlamydomonas reinhardtii*, A8JG73)	Good: $4x10^{-16}$	Flagellum				

regulatory proteins include sequences similar to the apoptosis inhibitor BAX inhibitor-motif-containing protein (e-value=5 x 10^{-20}), SNF family protein kinase (e-value 1x 10^{-32}), ubiquitin-conjugating enzyme (e-value=51 x 10^{-30}), and a few transcription factors or possible transcription factors such as IWS1 (e-value=9 x 10^{-17}) and histone-specific transcription factor (e-value=1x10^{-4}).

A number of stress-response and protein folding-related proteins were also found such as Hsp70 (e-value=4x10^{-16}), Hsp90 (e-value=7x10^{-6}), DnaJ chaperone protein (e-value=6x10^{-7}), and cell wall protein DAN4/delayed anaerobic protein 4 precursor (e-value=6x10^{-5}).

Some genes associated with structures were also found. These include genes similar to spondin-2 extracellular matrix protein (e-value=8x10^{-5}), rootletin (e-value=7x10^{-08}), cell wall-plasma membrane linker protein (e-value=9x10^{-9}), proline-rich glycoprotein (e-value=6x10^{-7}), and flaggelar-associated protein (e-value=4x10^{-16}). Other sequences in the collection were similar to hemolysin-type calcium binding protein (e-value=2x10^{-26}), serpin peptidase inhibitor (e-value=3x10^{-18}), sonic hedgehog protein (e-value=2x10^{-17}), voltage-gated chloride channel domain-containing protein (e-value=1x10^{-13}), BRCA-1 associated protein (e-value=5x10^{-5}), and vacuolar protein sorting associated protein (e-value=5x10^{-4}).

Most (78%) of our collection *E. denticulatum* unigenes do not have any significant similarity to known proteins. However, 15 have matches to putative or hypothetical protein sequences or to genomic DNA sequences that do not have reliable annotations. These were not included in Table 3 due to the lower quality of the annotation.

DISCUSSION

Our small-scale EST library shows that new sequences can be discovered in this commercially important alga. Out of 143 unigenes, only 32 unigenes displayed similarity to known proteins. These annotated sequences cover a variety of functions, from structural to metabolic to regulatory. The translation machinery and most of the protein folding, transport, modification, and degradation genes seem to occur across the different Rhodophyta EST libraries. Other genes that occur across most, if not all of the Rhodophyta EST libraries include the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase, peptide methionine sulfoxide reductase, the heat shock proteins HSP90 and HSP70, sonic hedgehog protein, and ubiquitin-conjugating enzyme family protein, SNF family protein kinase, histone specific transcription factor, and zinc finger protein. The finding that these genes are also found in other Rhodophyta EST libraries suggests that the pathways where these processes occur are conserved in these organisms.

The economic importance of *E. denticulatum* and many other algal species is largely due to the uses of their cell wall polysaccharides. The EST sequences that are associated with the cell wall, extracellular matrix, and the plasma membrane could provide additional understanding of cell wall biochemistry and its relation to the cell-wall polysaccharides. It is interesting to note that all of the five unigenes in this category do not have strong similarity with known sequences, and are not well-represented in the other Rhodophyta EST collections. Two of them, the sequences similar to proline-rich glycoprotein and to spondin-2, are found only in the EST libraries of the more closely related *Chondrus* and *Gracilaria* species but not in the libraries of the more distantly related *Griffithsia* and *Porphyra*. These genes are worth investigating for their possible association with the specialized cell wall biochemistries of these algae.

Other *E. denticulatum* unigenes that are common with other Rhodophyta EST libraries provide additional opportunities for studying these genes within this group of organisms. Of the 40 unigenes that show similarity to other Rhodophyta ESTs, 22 are annotated while 18 are unannotated. Three of these unannotated unigenes, C037, U024, and U108, matched to *Chondrus crispus* ESTs but not to any other EST libraries, suggesting that these unigenes may be specific to the carrageenophytes. All these unigenes represent novel sequences that are expressed in Rhodophyta, and are waiting to be matched with their function.

Of the 111 unigenes with no annotation, 15 had matches to hypothetical or putative proteins or genomic sequences. Three of these were also matched by ESTPiper to hypothetical, predicted or putative uncharacterized proteins (C087, C301, U190). The majority (93) of the unigenes show no significant matches with the other Rhodophyta EST libraries, despite the thousands of ESTs already in the database.

In summary, our small scale EST library represents the first look at the set of genes expressed in any *Eucheuma* species. Thirty-two of the 143 unigenes showed sufficient similarity to known genes to allow annotation. Of these 22 were found in at least one other EST collection from Rhodophyta species. The rest of these genes may be used as starting points for exploring genes that may be specific to *Eucheuma* and closely-related species.

Almost 80% of the unigenes, or 111 them, could not be annotated. For discovery of novel genes, this compares favorably with the novel genes discovered by the other Rhodophyta EST libraries. One of the most intriguing unannotated genes is unigene U005, which had the highest number of copies in the collection, representing 47 clones. This gene is apparently highly expressed in the *E. denticulatum* thallus, but its function cannot be deduced by similarity with known genes and may require analysis of the full gene product through cloning and gene knockouts or heterologous expression.

Furthermore, 93 or 64.6% of the library are novel genes that have not been found yet in the other Rhodophyta EST libraries. This suggests that, despite the other algal large scale libraries be-

ing made available by high throughput methods, many novel genes may still be uncovered by small scale EST collections.

Finally, our results represent a first and limited look at genes from *E. denticulatum*. The library was derived from the thalli of mature, farmed samples of the green variety. Other genes may be found from EST libraries generated from other developmental stages and varieties of this alga and its relatives. Expansion of this study's small scale EST data set, plus comparative expression studies, may lead to novel insights into important cellular process and details of the physiology and biochemistry of this commercially important alga.

ACKNOWLEDGEMENTS

We thank Dr. David Cheng of American Gene C.T. LLC (Cranston, RI, USA) for cDNA construction. We also thank Mr. Neil Tan Gana for his assistance and the Ateneo de Manila University Biology Department for the use of Vector NTI. This work was supported by the Commission on Higher Education (CHED) and the Department of Science and Technology-Philippine Council for Advanced Science and Technology Research and Development (DOST-PCASTRD) of the Republic of the Philippines, and by Ateneo de Manila University and Silliman University.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

PSA performed the bench experiments and annotation analysis for the unigenes. AACBA performed some of the bioinformatics analyses to annotate the major unigenes clusters. GCZ guided the development and sequencing of the cDNA library, provided insight on algal genetics, and hosted PSA's work in his laboratory. NRLR coordinated and assisted in the bioinformatics analysis of the unigenes and edited this manuscript.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25(17):3389-3402.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2009; 36(Database issue):D25-30. (http://www.ncbi.nlm.nih.gov/Genbank)

Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". Nature Genetics 1993; 4(4): 332-333.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N. Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Research 2004: 32 (Database issue):D262-D266.

Collén J, Roeder V, Rousvoal S, Collin O, Kloareg B, Boyen C. An expressed sequence tag analysis of thallus and regenerating protoplasts of *Chondrus crispus* (Gigartinales, Rhodophyceae). J Phycol 2006; 42:104-112.

DA-AMAS (2005) Seaweeds Factsheet. Manila: Department of Agriculture-Agribusiness and Marketing Assistance Service, Republic of the Philippines. 2005; http://www.da.gov.ph/agribiz/commodityfactsheet_seaweeds.html. Cited 11 November 2009.

Doty MS. The production and use of Eucheuma. In: Case study of seven commercial seaweeds resources. Doty MS, Caddy JF, Santelices B. FAO Fisheries Technical Paper-T281, 1987; http://www.fao.org/docrep/x5819e/x5819e06.htm. Cited 10 November 2009.

Fan X, Fang Y, Hu S, Wang G. Generation and analysis of 5,318 expressed sequence tags from the filamentous sporophyte of *Porphyra haitanensis* (Rhodophyta). J Phycol 2007; 43:1287-1294.

FAO Fisheries and Agriculture Department. The State of World Fisheries and Aquaculture 2008. Rome: Food and Agricultural Organization of the United Nations.

Fredericq, S., Freshwater, D. W. & Hommersand, M. H. Observations on the phylogenetic systematics and biogeography of the Solieriaceae (Gigartinales, Rhodophyta) inferred from *rbc*L sequences and morphological evidence. Hydrobiologia 1999; 399: 25-38.

Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 2004: 32, Database issue:D258-D261.

Guiry MD, Guiry GM. Algaebase. National University of Ireland, Galway. 2009; http://www.algaebase.org. Cited 23 June 2009.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 2009; 10:136.

Kloareg B, Quatrano RS. Structure of the cell walls of marine algae and ecophysiological functions of the matrix polysaccharides. Oceanogr Mar Biol Annu Rev 1988; 26:259-315.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.. Clustal W and Clustal X version 2.0. Bioinformatics 2007; 23:2947-2948.

Lee H, Lee HK, An G, Lee YK. Analysis of expressed sequence tags from the red alga *Griffithsia okiensis*. J Microbiol 2007; 45(6):541-546.

Lluisma AO, Ragan MA. Relationships among *Eucheuma denticulatum*, *Eucheuma isiforme* and *Kappaphycus alvarezii* (Gigartinales, Rhodophyta) based on nuclear ssu-rRNA gene sequences. J Appl Phycol 1995: 7:471-477.

Lluisma AO, Ragan MA. Expressed sequence tags (ESTs) from the marine red alga *Gracilaria gracilis*. J Appl Phycol

1997; 9:287-293.
Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics 2006: 8:6-21.
Nikaido I, Asamizu E, Nakajima M, Nakamura Y, Saga N, Tabata S. Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis*. DNA Res 2000; 7:223-227.
Sambrook J, Russel DW. The condensed protocols from Molecular Cloning: A Laboratory Manual. 2006. Cold Spring Harbor Press, Cold Spring Harbor New York.
Sun X, Yang G, Mao Y, Zhang X, Sui Z, Qin S. Analysis of expressed sequence tags of a marine red alga, *Gracilaria lemaneiformis*. Prog Nat Sci 2002; 12:518-523.
Tang Z, Choi J-H, Hemmerich C, Sarangi A, Colbourne JK, Dong Q. ESTPiper – a web-based analysis pipeline for expressed sequence tags. BMC Genomics 2009; 10:174.
Teo S-S, Ho C-L, Teoh S, Lee W-W, Tee J-M, Rahim RA, Phang S-M. Analyses of expressed sequence tags from an agarophyte, *Gracilaria changii* (Gracilariales, Rhodophyta). Eur J Phycol 2007; 42:41-46.
UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res 2009; 37:D169-D174.
Zuccarello G.C., Critchley A.T., Smith J., Sieber V., Bleicher Lhonneur G. & West J.A. Systematics and genetic variation in commercial *Kappaphycus* and *Eucheuma* (Solieriaceae, Rhodophyta). J Appl Phycol 2006; 18:643-651.