Mar. Biotechnol. 6, 278–290, 2004
DOI: 10.1007/s10126-003-0035-3

MARINE BIOTECHNOLOGY

Analysis of Expressed Sequence Tags from Calcifying Cells of Marine Coccolithophorid (*Emiliania huxleyi*)

Thomas M. Wahlund,[1] Ahmad R. Hadaegh,[2] Robin Clark,[1] Binh Nguyen, Michael Fanelli,[1] and Betsy A. Read[1]

[1]*Department of Biological Sciences, California State University San Marcos, 333 S. Twin Oaks Valley Road, San Marcos, California 92096-0001, USA*
[2]*Department of Computer Science, California State University San Marcos, 333 S. Twin Oaks Valley Road, San Marcos, California 92096-0001, USA*

Abstract: An expressed sequence tag (EST) approach was used to investigate gene expression in the unicelluar marine alga *Emiliania huxleyi*. We randomly selected 3000 EST sequences from a cDNA library of transcripts expressed under conditions promoting coccolithogenesis. Cluster analysis and contig assembly resulted in a unigene set of approximately 1523 ESTs. Only 36% of the unique sequences exhibited significant homology to sequences in GenBank. Of particular interest were the numerous transcripts with homology to sequences associated with sexual reproduction and calcium homeostasis in other unicellular and multicellular organisms. The majority of ESTs (64%) had little or no significant sequence homology to entries in GenBank, suggesting a potential for further novel gene discovery. The catalog of ESTs reported herein represents a significant increase in the limited sequence information currently available for *E. huxleyi* and should make the coccolithophorid more accessible to powerful genomics and postgenomics technologies.

Key words: coccolithophorid, *Emiliania huxleyi*, EST sequencing, algae genomics.

INTRODUCTION

Coccolithophorids are an extremely important calcite-producing group of unicellular algae in the marine environment. The most abundant coccolithophorid, *Emiliania huxleyi*, is distributed throughout the world's oceans and coastal waters. *E. huxleyi* is unique among the marine phytoplankton in that it is capable of fixing atmospheric carbon into both photosynthetic and biomineralized product. This alga has a significant impact on the flux of CO_2 across the air-sea interface, and also on the removal of CO_2 as calcium carbonate at the deep water–sediment interface (Westbroek et al., 1993). These data indicate that *E. huxleyi* plays an important role in the ocean carbon cycle and may even influence the global climate system by decreasing the oceanic draw of CO_2. *E. huxleyi* is also recognized as a major sink for calcium carbonate in the ocean (Hide, 1990; Samtleben and Bickert, 1990). Ecophysiologists and climatologists are interested in *E. huxleyi*'s involvement in sulfur biotransformations in the ocean and its ability to synthesize long-chain alkenones and alkyl alkenoates. The production of dimethylsulfide (DMS) in

Received July 14, 2003; accepted October 27, 2003; online publication May 13, 2004.

Corresponding author: Betsy A. Read; e-mail: bread@csusm.edu

E. huxleyi blooms may affect production and regional weather patterns (Bates et al., 1987; Charlson et al., 1987), while the long-chain polyunsaturated ketones have proved to be accurate paleotemperature proxies for estimating surface water temperature distributions to determine patterns in ocean circulation and paleoclimate (Prahl et al., 1988; Sikes et al., 1991; Conte et al., 1992).

In addition to its ecologic importance, *E. huxleyi* has attracted the attention of materials scientists interested in using these porous shells of calcium carbonate to develop novel materials. Potential applications include the design of new lightweight ceramics, catalyst supports, robust membranes for high-temperature separation technology, and biomedical devices (Walsh and Mann, 1995). Despite its use in biogeochemistry, climatology, and materials science, little is known about the molecular genetics of this important marine alga. Molecular approaches aimed at elucidating the complex life cycle of *E. huxleyi*, and tools for analyzing genes that express the protein machinery responsible for calcium carbonate biomineralization and DMS production, are lacking (Paasche, 2002). The size of the *E. huxleyi* genome is not known, and there is little information that describes the content and organizational structure of the genome. At the time of this study, a search of databases for protein-encoding genes in *E. huxleyi* yielded only 5 to 10 entries; this situation has restricted our understanding of the biochemical and physiologic pathways that govern the biology of this alga.

Therefore, to accelerate the genetic and molecular characterization of the biology of *E. huxleyi*, we present results obtained from the identification of 3000 *E. huxleyi* expressed sequence tags (ESTs) based on cDNA sequencing. The analysis of ESTs generated by systematic partial sequencing of randomly picked cDNA clones is an effective means of rapidly gaining information about an organism at its most fundamental level. Analyses of ESTs have been published for several model plants, including *Arabidopsis*, rice, maize, and wheat (DeRisi and Iyer, 1999), but this approach has not been extensively employed with algae. We have identified transcripts that are expressed under conditions that promote calcification and coccolithogenesis, which include those encoding proteins that are likely to be involved in calcium homeostasis and transport. In addition, many apparently novel genes have been identified. These genes include transcripts that are present in *Volvox*, yeast, and other organisms and that are known to be involved in gametogenesis and sexual reproduction. The EST sequence information presented herein will complement the large set of physiological information already available and enable new technologies to be rapidly exploited to advance our understanding of the global significance of *E. huxleyi*.

Materials and Methods

Media and Growth Conditions

E. huxleyi strain 1516 was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton and grown as described previously (Laguna et al., 2001). RNA was extracted from cultures obtained by inoculating cells into 1 L of f/50 medium (Guillard, 1975) in 4-L flasks. Cultures were incubated photoautotrophically at 17° to 18°C under cool white fluorescent light (660 μmol $\cdot$ $m^{-2} \cdot s^{-2}$) under a discontinuous-light (12-hour dark, 12-hour light) cycle.

RNA Extraction

RNA was isolated from 3 L of cultures in mid to late log phase. Prior to RNA extraction cells were decalcified by lowering the pH of the culture with HCl to a pH of 5.0 for 2 minutes, followed by rapid readjustment with NaOH to pH 8.0. Total RNA was extracted from cells using a standard guanidinium isothiocyanate procedure (Strommer et al., 1993). Briefly, cells were lysed by grinding in liquid nitrogen with a mortar and pestle. Cell material was resuspended in extraction buffer (4 M guanidinium isothiocyanate, 25 mM sodium citrate, 0.5% sarkosyl, 0.1 M β-mercaptoethanol) to inhibit the activity of ribonucelases and disrupt membranes. Total RNA was separated from other cellular components by phenol extraction followed by isopropanol precipitation with sodium acetate. A final lithium chloride precipitation was performed to further purify the RNA. The concentration of RNA was determined from its absorbance at 260 nm, and the integrity was assessed using denaturing agarose gel electrophoresis.

Construction of cDNA Library and EST Sequencing

Total RNA was used for the construction of a cDNA library prepared by ResGen (Invitrogen Corp.). First-strand synthesis was performed using a *Not*I primer-adapter (GAC TAG TTC TAG ATC GCG AGC GGC CGC $CC(T)_{15}$) and Superscript II reverse transcriptase. Following second-strand synthesis

using *Escherichia coli* DNA polymerase, *Not*I/blunt end products were directionally cloned into the *Not*I/*Eco*RV sites of the Gateway cloning vector pMAB58. Plasmids were used to transform ElectroMax DH10B-TON cells via electroporation, and random clones were picked for quality control analysis.

Plasmid DNA was prepared from recombinant clones using a standard alkaline lysis procedure, and unidirectional sequencing was accomplished using the pMAB58 forward primer (TAT AAC CGC TTT GGA ATC ACT), providing sequence from the 5′ end of cDNA clones. Sequencing was performed by Integrated Genomics of Chicago, Illinois.

Data Analysis

ESTs were trimmed to remove the vector and ambiguous sequences, and high-quality sequences with a minimum of 400 bp of continuous sequence with at least 98% accuracy were retained for further analysis. High-quality sequences were compared with sequences in GenBank (National Center of Biotechnology and Information, NCBI) using BLASTX. A sequence was considered to be a significant match when the BLAST probability value (*e* value) was less than 1×10^{-2}. High-quality ESTs were assembled into contigs using the phrap/cross_match/swat package version 0.990329 (available at pg@umpqua.genome.washinton.edu). A final unique set of 1523 sequences has been deposited into GenBank (accession numbers CF753162–CF754684; dbEST_Id 20096956–20098478) and archived in our *E. huxleyi* database (*Ehux Express*). A Web interface is currently being constructed to allow keyword or sequence homology searches to be performed.

BLASTCLUST was used to group the initial ESTs into consensus sequences using match reward of 1, mismatch penalty of −3, non-affine gapping cost, and a word size of 28 with an *e*-value threshold set at 1*e*-6. Pairwise comparisons across the initial sequences were also used to determine the total redundancy in the library. Random subsets of ESTs (500, 1000, 1500, 2000, 2500, and 3000) were sampled, and the number of unique sequences in each subset was determined (Figure 1).

Results and Discussion

EST Library Sequence Analysis

The cDNA library employed in this study consisted of 6×10^5 clones, from which the 5′ ends of 3000 cDNAs were sequenced. After editing to eliminate vector and other problematic sequences, high-quality ESTs with an average length of 559 nucleotides were used in database searches. As shown in Table 1, 1836 (approx. 61%) of the ESTs exhibited an *e* value greater than or equal to 10^{-2}, and 78 (approx. 3%) of the sequences had no GenBank match. For the remaining 1086 (approx. 36%) of ESTs returning an *e* value less than 10^{-2}, matches were found to genes from a wide diversity of organisms. Highly significant matches were most frequently obtained with sequences from animals and plants and fungi. However, significant matches to sequences from prokaryotes and unicellular eukaryotes were also observed. Table 1 also lists significant *E. huxleyi* EST matches assigned into groups or domains based on GenBank search data. The GenBank search results appear to reflect the current bias in the databases for animal sequences relative to eukaryotic photosynthetic organisms, plants, or algae, as one would expect sequences from *E. huxleyi* to be most closely related to plants or algae.

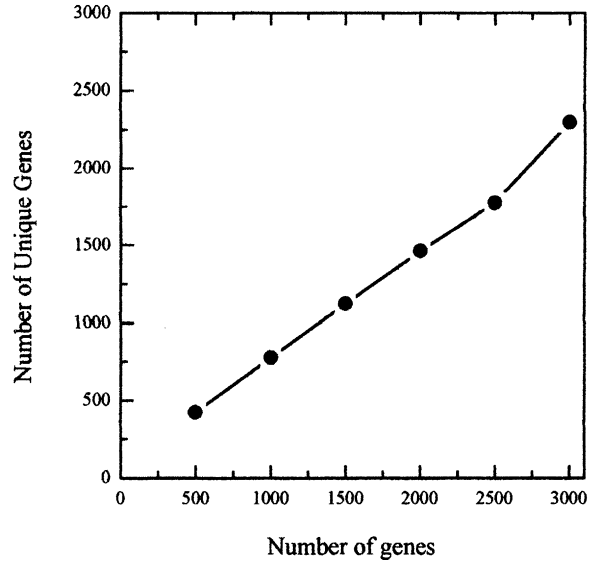

Figure 1. Characterization of the rate of new gene discovery expressed as the number of unique sequences obtained versus the total number of clones sequenced.

Analysis of rates of gene discovery indicated that our library prepared from RNA extracted from calcifying *E. huxleyi* cells contains more information than we had mined from this initial sequence screen. The number of sequences that can be processed and the potential new information that can be gleaned from that effort is represented graphically in Figure 1. After sequencing 3000 ESTs, 2298 different transcripts were predicted using BLASTCLUST and the rate of new sequence discovery was still at 76.6%. At this

Table 1. BLAST Search Analysis of cDNA Library Clones

Descriptive category	No. of ESTs
Total cDNAs sequenced	3000
EST matches with *e* value $\geq 1 \times 10^{-2}$	1836
ESTs with no GenBank match[a]	78
EST matches with *e* value $< 1 \times 10^{-2}$	1086
Eukaryotes	
Animal	341
Plant/fungi	294
Unicellular, photosynthetic	76
Unicellular, nonphotosynthetic	116
Prokaryotes (Bacteria/Archaea)	242
Viral-related sequences	17

[a]Represents searches with no *e* values >10.

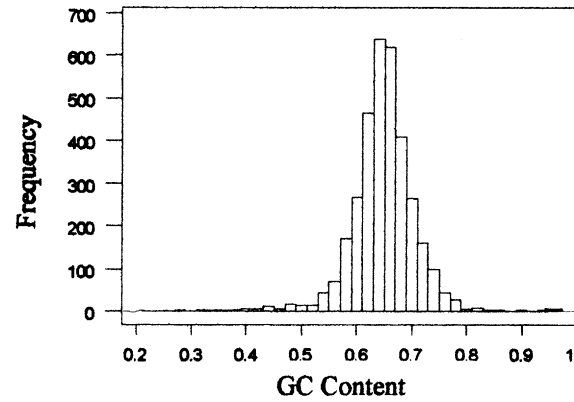

Figure 2. G + C content of 3000 EST sequences from cDNA library of *E. huxleyi* strain 1516. The frequency distribution mean of these EST data (approx. 65%) reflects the high GC content previously described for *E. huxleyi*.

point there is no indication of a plateau effect, suggesting the sequencing of more library clones is warranted. Assembly of individual ESTs into groups of tentative consensus sequences yielded 1523 unique transcripts, a 200-fold increase in what was previously contained in GenBank. Our unigene set is composed of 1054 singletons and 459 contigs.

The average G + C content ratio from this library sampling was 0.65, with 68% of the sequences having a G + C content between 0.59 and 0.70 (Figure 2). The leptokurtic distribution suggests that the G + C content is constant across the coding region of the genome and indicates that the presence of contaminating sequences is minimal.

Given its high G + C content, *E. huxleyi* might be expected to use a GTG initiation codon in addition to the preferred ATG codon, as is the case with *Mycobacterium tuberculosis*, which has a similar G + C content (Lowery and Ludden, 1988). Analysis of the predicted start codon of a small subset of matched ESTs reported herein ($n = 70$) revealed that a GTG start codon was used to define the start of translation at least 14% of the time, and possibly as much as 44% of the time.

Preliminary data we have collected using open reading frames from 85 ESTs (those with the lowest *e* values) and 15 full-length cDNA sequences suggest that *E. huxleyi* exhibits a codon bias consistent with its high G + C content (Table 2). These results are in agreement with previous findings that suggested a codon bias based on the G + C composition of codon positions in cDNA clones from the actin multigene family in *E. huxleyi* (Bhattacharya et al., 1993). Information pertaining to the alga's preferred codon usage is of practical importance in terms of designing degenerate primers for polymerase chain reaction and performing in vivo genetic manipulation experiments. Our preliminary data also suggest the high G + C content may reflect a biased amino acid content of the *E. huxleyi* proteome (Table 2). In *E. huxleyi*, as in *M. tuberculosis* and other organisms harboring genomes with a high G + C content, there appears to be a distinct preference for amino acids encoded by the GC-rich codons of Ala, Gly, Pro, Arg, and Trp, as compared with those encoded by the A + T-rich codons of Asn, Ile, Lys, Phe, and Tyr (Collins and Jukes, 1993; Foster et al., 1997; Lobry, 1997; Gu et al., 1998). Whether this preference is characteristic of the entire *E. huxleyi* proteome and influences the structure and chemistry of its proteins is not known and beckons further analysis.

ESTs were grouped according to putative cellular function (Table 3) as described previously (Adams et al., 1995). The ESTs with putatively identified functions encompassed a wide variety of biological processes including ribosomal proteins, cell division, gene or protein expression, cell signaling, cell structure, defense, and metabolism. Table 3 is not an inclusive list of all ESTs with *e* values less than 1×10^{-2}, but rather a representation of a select set of ESTs from each functional class to demonstrate the apparent diversity of the library. Figure 3 shows the percentage distribution of sequences falling into each of the functional categories. Data from the 1086 ESTs with significant matches indicate that 35% of those sequences encode proteins involved in metabolism (Figure 3, A). Interestingly, 15% of the represented sequences encoded proteins involved in cell defense supporting the hypothesis

Table 2. Codon and Amino Acid Usage in *Emiliania huxleyi* from Analysis of 85 ESTs and 15 Full-length cDNA Clones[a]

AA	Codon	N[b]	RSCU[c]	AA	Codon	N[b]	RSCU[c]
Phe	UUU	45	0.46		UCA	32	0.29
	UUC	153	1.55		UCG	124	1.23
Leu	UUA	7	0.06		AGU	23	0.25
	UUG	35	0.31		AGC	105	1.32
	CUU	82	0.72	Cys	UGU	17	0.25
	CUC	348	2.92		UGC	116	1.75
	CUA	45	0.40	Trp	UGG	65	0.50
	CUG	192	1.61	Pro	CCU	85	0.73
Tyr	UAU	16	0.21		CCC	237	2.03
	UAC	138	1.80		CCA	66	0.54
His	CAU	108	0.77		CCG	184	1.50
	CAC	177	1.23	Arg	CGU	129	0.93
Gln	CAA	171	0.98		CGC	234	1.65
	CAG	179	1.03		CGA	194	1.19
Ile	AUU	28	0.28		CGG	140	0.88
	AUC	243	0.65		AGA	39	0.29
	AUA	6	0.07		AGG	137	1.07
Met	AUG	153	1.00	Thr	ACU	29	0.31
Asn	AAU	13	0.25		ACC	181	2.22
	AAC	164	1.85		ACA	21	0.20
Lys	AAA	60	0.34		ACG	145	1.27
	AAG	277	1.66	Ala	GCU	86	0.50
Val	GUU	47	0.41		GCC	236	1.61
	GUC	213	1.88		GCA	64	0.36
	GUA	30	0.27		GCG	239	1.54
	GUG	162	1.45	Gly	GGU	87	0.43
Asp	GAU	110	0.50		GGC	380	1.88
	GAC	337	1.50		GGA	81	0.46
Glu	GAA	79	0.40		GGG	205	1.24
	GAG	328	1.61	ter	UAA	8	0.00
Ser	UCU	53	0.56		UAG	5	0.00
	UCC	144	0.58		UGA	19	0.00

Amino acid usage[c]

Phe	Leu	Ser	Tyr	Cys	Trp	Pro	His
2.15	7.71	8.95	1.68	1.84	1.39	6.99	3.1
Gln	Arg	UGA	Ile	Met	Thr	Asn	Lys
3.81	12.46	0.76	3.01	1.67	5.81	1.93	3.65
Val	Ala	Asp	Glu	Gly	UAA	UAG	
4.92	10.51	4.87	4.45	8	0.09	0.06	

[a]Analysis was performed using the General Codon Usage Analysis Version 1.1 program of J. McInerney.

[b]Number of times a particular codon is observed.

[c]The relative synonymous codon usage (RSCU) values represent the number of times a particular codon is observed relative to the number of times that the codon would be observed in the absence of any codon bias. The RSCU value is 1 in the absence of codon bias.

that coccolithogenesis may be a response to environmental or physiologic stress (Paasche, 2002). Genes with hypothetical or putative function represented 8.2% (Figure 3, B, groups 8 and 9), whereas novel sequences represented the vast majority of the total sequences, at (group 10).

The most prevalent transcripts in the cDNA library generated from *E. huxleyi* cells grown under conditions promoting calcification as determined by BLASTCLUST are listed in Table 4. The fact that we have constructed a nonnormalized primary library suggests that the abundance or cluster size is more likely to be indicative of the relative messenger RNA population. Of the 3000 ESTs, a total of 25 clusters contained 10 or more sequences, together constituting 19% of the sequenced clones. Sequences in the 3 largest clusters contained 131, 52, and 51 members, respectively. These transcripts, which are presumably the most abundant in the library, showed no significant similarity to sequences in GenBank. The most prevalent identifiable transcripts in the library were actin and polyubiquitin, clusters of which contained 51 and 37 members, respectively.

Gene Content Analysis

Most known transcripts are considered housekeeping genes, such as those involved in metabolism (e.g., photosynthesis and carbon fixation, amino acid and carbohydrate metabolism, nitrogen and sulfur assimilation, and the synthesis of isoprenoids and phenylpropanoids). One metabolic transcript of particular interest is phosphoenolpyruvate (PEP) carboxykinase (5 copies), which plays a key role in C_4 metabolism in plants. In many algae and vascular plants, the fixation of CO_2 by PEP carboxylase works in concert with a C_4-C_1 decarboxylase (e.g., an $NADP^+$- or NAD^+-dependent malic enzyme) to provide CO_2 to RubisCO (Raven, 1997). The presence of multiple PEP carboxykinase transcripts in the library suggests that *E. huxyleyi* may be CO_2 limited in seawater, and that C_4 photosynthesis may support carbon assimilation in *E. huxleyi*, as described in the marine diatom *Thalassiosira weissflogii* (Reinfelder et al., 2000). Alternatively, PEP carboxykinase may function as another carbon-concentrating mechanism (CCM) in this alga. Many contend that *E. huxleyi* does not require a CCM because calcification (which shifts the DIC equilibrium toward CO_2) is an efficient alternative in coccolithophorids and may even be more efficient than a traditional CCM (Steeman, 1966; Brownlee et al., 1994). Data obtained from recent studies, however, did not show a significant correlation between

Table 3. Representative ESTs Showing Significant GenBank Match and Grouped into Functional Classes[a]

Reference number	Putative/known function[b] (total ESTs per class)	*e* value
	Cell Division (17)	
AF079404	Cell cycle switch protein	10^{-65}
NC_000911	Cell division protein; FtsH	10^{-8}
NC_002751	Cell division cycle protein 48	10^{-42}
NC_002932	DNA helicase (*Chlorobium tepidum*)	10^{-27}
NM_001255	Cell division cycle 20	10^{-9}
NM_113414	Cdc45-like protein	10^{-6}
AF480497	Putative apospory-associated protein	10^{-13}
AF421549	CDH1-D (*Gallus gallus*)	10^{-7}
NC_003888	DNA polymerase III γ subunit	10^{-3}
D14489	PRIB protein	10^{-29}
	Cell signaling/communication (86)	
AF216527	Calcium-dependent protein kinase	10^{-62}
AJ294903	Cyclin-dependent kinase C	10^{-61}
AY062449	Cdc2-like protein kinase	10^{-11}
AF055079	Inositol 1,4,5-trisphosphate receptor	10^{-24}
NM_077143	Calmodulin	10^{-8}
NM_079883	Calcium/calmodulin protein kinase	10^{-32}
AF386797	Serine-threonine protein kinase PK2	10^{-12}
NC_004317	Serine/threonine protein phosphatase	10^{-36}
AB035141	Mitogen-activated protein kinase	10^{-50}
AF302112	CBL-interacting protein kinase 1	10^{-15}
AK011258	Checkpoint kinase 1 homologue	10^{-10}
AF121198	14-3-3 protein	10^{-43}
AB070345	Matrix metalloproteinase	10^{-5}
AF086823	Rho/rac-interacting citron kinase	10^{-7}
A56492	Protein kinase ERK2	10^{-68}
NM_102380	Pto kinase interactor	10^{-3}
	Cell structure/motility (101)	
S64188	Type 1 actin (*Emiliania huxleyi*)	10^{-96}
AB092418	Calponin (*Branchiostoma belcheri*)	10^{-29}
M87526	Flagellar radial spoke protein	10^{-23}
NM_101279	Mitochondrial carrier protein	10^{-8}
AC115608	Spore coat protein SP96	10^{-3}
AF502577	ϵ-tubulin	10^{-19}
NM_100360	Tubulin α-2/α-4 chain	10^{-78}
NM_138958	Autocrine motility factor receptor	10^{-4}
NM_115477	a-soluble NSF attachment protein	10^{-23}
AF303112	Actin-binding protein fragmin 60	10^{-33}
AF317890	Paxillin	10^{-8}
AJ311549	VMP3 protein (*Volvox carteri*)	10^{-3}
M87526	Clathrin coat assembly protein AP50	10^{-46}
L36202	Fimbrin	10^{-21}
NM_033161	Troponin C	10^{-6}
NP_035642	Surfeit 4	10^{-10}
NM_080306	Pecanex	10^{-3}
	Cell defense (160)	

(Continued)

Table 3. Continued

Reference number	Putative/known function[b] (total ESTs per class)	*e* value
AB003732	Polyubiquitin	10^{-94}
AJ416499	Putative ubiquitin	10^{-91}
AF043518	20S proteasome subunit PAA1	10^{-47}
NC_002752	Heat shock protein 70	10^{-75}
NM_061169	DnaJ, prokaryotic heat shock protein	10^{-17}
NM_104252	26S proteasome, ATPase subunit4	10^{-62}
X99730	Cathepsin	10^{-29}
U67931	Ubiquitin/ribosomal protein	10^{-40}
AB024993	DnaK-type molecular chaperone	10^{-16}
AC027038	Hypersensitive response protein	10^{-34}
AC091774	Heat shock protein 90	10^{-43}
AF083890	19S proteosome subunit 9	10^{-28}
AF221856	Heat shock protein 80	10^{-57}
AF397903	AAA-metalloprotease FtsH	10^{-45}
BT000717	Putative heat shock protein 81-2	10^{-53}
NC_001147	Metacaspase; Mca1p	10^{-15}
NM_124642	Heat shock protein 81-1	10^{-68}
	Gene/protein expression (72)	
AC009322	Putative splicing factor Prp8	10^{-74}
AJ490820	c-*myb* like protein	10^{-17}
NP_062518	Histone deacetylase 1	10^{-45}
AL360314	Dead Box RNA helicase RH15-like	10^{-49}
NM_004597	Small nuclear ribonucleoprotein D2	10^{-35}
AF037460	GF14 protein (*Fritillaria agrestis*)	10^{-30}
AF139989	rRNA intron-homing endonuclease	10^{-8}
NC_000917	Transcriptional regulatory protein	10^{-17}
NC_002516	DnaJ protein	10^{-13}
NC_003423	ATP-depen RNA helicase, putative	10^{-33}
NM_012245	SKI-interacting protein	10^{-11}
AF232676	Prophet of pit-1	10^{-8}
NM_123501	TFIIH basal transcription factor	10^{-43}
	Metabolism (382)	
NM_007591	Calreticulin	10^{-29}
NM_018946	N-acetylneuraminic acid-P-synthase	10^{-48}
NM_076383	Ammonium transporter	10^{-19}
NC_001264	Sulfite oxidase, putative	10^{-24}
NC_002696	Thiolase family protein	10^{-48}
NC_002753	CbbX protein homologue	10^{-14}
NC_002932	Aldehyde DH (*C. tepidum*)	10^{-15}
AY049067	Phosphoenolpyruvate carboxykinase	10^{-50}
AF012542	Calcium binding protein (*E. huxleyi*)	10^{-47}
AF265362	3GPA dehydrogenase	10^{-39}
AF302496	NADPH-cyt P450 oxydoreductase	10^{-21}
BC010570	HMG-CoA lyase	10^{-37}
AF521254	Fructose-bisphosphate aldolase	10^{-25}
NM_004458	Long-chain fatty-acid-CoA ligase 4	10^{-14}
NM_126074	Succinate dehydrogenase flavoprotein	10^{-47}

(Continued)

Table 3. Continued

Reference number	Putative/known function[b] (total ESTs per class)	*e* value
AY059637	Malate synthase	10^{-38}
NC_000918	Thioredoxin reductase	10^{-16}
NC_003424	Diphosphomevalonate decarboxylase	10^{-22}
NM_054070	Mitochondrial Zn metalloprotease	10^{-30}
NM_060801	ATP synthase α and β subunits	10^{-72}
NM_068639	Vacuolar ATPase E-like subunit	10^{-20}
NM_126074	Succinate DH flavoprotein α subunit	10^{-46}
U58680	Light-harvesting complex I polypep.	10^{-8}
AJ000670	Fucoxanthin-chl a/c protein	10^{-16}
U73686	Cytosolic glycoprotein FP21	10^{-29}
Y16748	Malate dehydrogenase	10^{-33}
AF110782	Phosphoglycerate kinase, chloroplast	10^{-31}
D47019	RubisCO-expression protein CfxX	10^{-22}
NM_069175	Mitochondrial processing peptidase	10^{-21}
AC018727	Urea active transport protein, putative	10^{-13}
P45699	Endoglucanase type K precursor, put	10^{-7}
AB045172	Family 45 cellulase homologue	10^{-8}
P49307	Rhizopine catabolism protein *mocA*	10^{-4}
NM_018779	Phosphodiesterase 3A, cGMP-inhib	10^{-4}
NC_001263	Acetyl-CoA acetyltransferase	10^{-33}
	Ribosomal proteins (15)	
NM_106066	Putative 60S ribosomal protein	10^{-31}
AJ011717	40S ribosomal protein S12	10^{-40}
NM_001023	40S ribosomal protein S20	10^{-41}
M76762	Ribosomal protein (*Mus musculus*)	10^{-47}
AF400191	Ribosomal protein L27	10^{-35}
	Other (168)	
AF466203	Putative gag-pol precursor (*Zea mays*)	10^{-3}
NM_000462	Human papilloma E6-protein	10^{-15}
NC_002642	11L protein (Yaba-like disease virus)	10^{-5}
AY090452	Polymerase (hepatitis B virus)	10^{-4}
AB074880	Putative Pi-transporter homologue B1	10^{-27}
AC004145	L5801.5 (*Leishmania major*)	10^{-23}
NM_101441	PRLI-interacting factor L	10^{-46}
NM_113195	ADP-ribosylation factor, putative	10^{-17}

[a]Data shown represent a selected set of ESTs in each functional class from the total ESTs with *e* value less than 1×10^{-2}.
[b]Representative ESTs with significant *e* value ($<1 \times 10^{-2}$) and a match in GenBank with a known or putative function. The number in parentheses is the total number of ESTs in each class in the library, and the reference number refers to the accession number corresponding to the most significant *e*-value search result. Please refer to these numbers for the relevant journal/database reference.
[c]Shows the number of ESTs from the *E. huxleyi* cDNA library that are similar to each listed member of given functional class.

increased calcification rates under low CO_2 concentrations—the results of which would presumably generate more CO_2 for photosynthesis (Clark and Flynn, 2000). In *E. huxleyi*, carboxylases other than RubisCO that have been shown to be involved C_4 photosynthesis in other organisms have not been investigated (Raven, 1997).

Our cDNA library was constructed from phosphate-stressed cells (f/50 medium), and thus it is not surprising that a number of cell stress or defense-related transcripts were present, including various heat shock proteins (HSP 70, HSP 80, HSP 81, HSP 82, and HSP 90) and the co-chaperonins Dna J and Dna K. A number of different

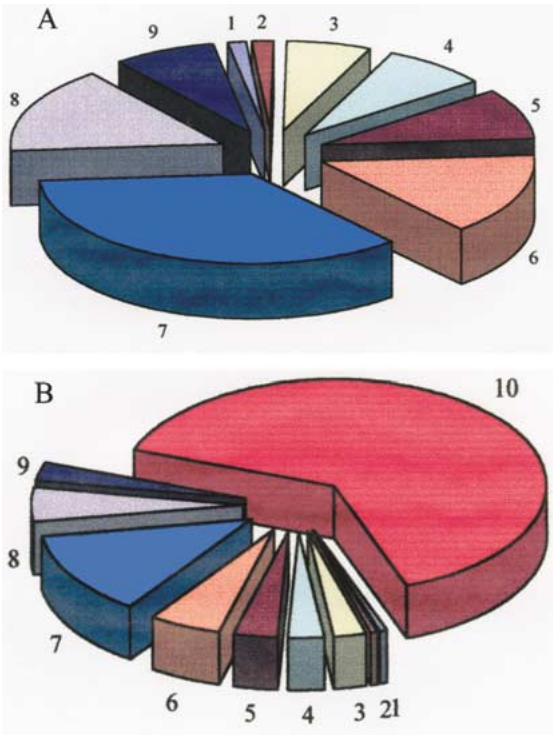

Figure 3. Percentage distribution of ESTs by functional classes. **A:** ESTs with significant (*e* value $< 10^{-2}$) matches. (1) ribosomal proteins, 1.35%; (2) cell division, 1.6%; (3) gene/protein expression, 6.6%; (4) cell signaling, 7.9%; (5) cell structure, 9.3%; (6) cell defense, 14.7%; (7) metabolism, 35.2%; (8) other matches, 15.5%; and (9) hypothetical proteins, 7.9%. **B:** Total ESTs sequenced, with class numbering the same as in A. (1) 0.5%, (2) 0.6%, (3) 2.4%, (4) 2.9%, (5) 3.4%, (6) 5.4 %, (7) 12.8%, (8) 5.4%, (9) 2.8%, (10) nonsignificant (*e* value $\geq 10^{-2}$) matches, 63.8%.

transcripts related to programmed cell death and apoptosis were also noted. Several copies of a metalloproteinase sequence and a hypersensitive response element were identified along with cathespin, caspase, metacaspase, and other members of the cysteine protease family. The collective presence and prevalence of these transcripts suggests that programmed cell death is an active process in *E. huxleyi*, and may be an adaptation to adverse environmental conditions, such as nutrient deprivation, that can trigger the rapid dissolution of algal blooms.

A number of different transcription factors and nucleic acid binding proteins were predicted from *E. huxleyi* ESTs by their similarity to known proteins. Although several general transcription factors are present, *cmyb* is the most abundantly represented transcription factor in the library, with 3 ESTs in the data set. Three other different *myb* transcription factors are also present. The Myb proteins are a family of transcription factors that occur in both animal and plant lineages but have been dramatically amplified in the plants. In *Arabidopsis* this large family of more than 100 gene regulatory proteins plays a fundamental role in regulation of metabolism. In both *Arabidopsis* and *Chlamydomonas reinhardtii*, one of the Myb transcription factors has been shown to be involved in signaling during phosphate starvation (Rubio et al., 2001). In *E. huxleyi* phosphate starvation is linked to calcification (Riegman et al., 2000); hence, it is reasonable to hypothesize that one of the Myb transcription factors could be involved in the regulation of genes involved in calcification and coccolithogenesis.

We were also able to identify proteins with zinc finger motifs as well as sequences with significant homology to several known homeodomain transcription factors. Although homeobox-containing genes play developmentally important roles in a wide variety of plants, animals, and fungi, few homeodomain proteins have been described in algae. A gamete-specific, sex-limited homeodomain protein has been identified in *Chlamydomonas* (Kurvari et al., 1998), and a homeodomain protein that appears to play a role during early reproductive development has been identified in *Acetabularia acetabulum* (Serikawa and Mandoli, 1999). Consequently, it is not unreasonable to envision a role for these homeobox transcription factors in the induction of phase variation events that lead to switching from the haploid (S-cell) to the diploid (C-cell) stage in the life cycle of *E. huxleyi*.

Another one of the more interesting nucleic acid binding proteins is a posttranscriptional regulator that belongs to the pumilio family of RNA binding proteins. Members of this family of proteins in *Drosophilia melanogaster* are responsible for maintaining germline stem cells (Forbes and Lehmann, 1998; Parisi and Lin, 1999); in *Caenorhabditis elegans* they promote the switch from sperm to egg production (Zhang et al., 1997; Tollervey and Caceres, 2000); and in *Dictyostelium discoideum* they control the development of reproductive structures (Souza et al., 1998, 1999). Pumilio-family proteins in *S. cerevisiae* regulate mRNA turnover by causing deadenylation and degradation of transcripts including the *HO* endonuclease involved in regulation of the mating-type switch (Tadauchi et al., 2001). In *E. huxleyi* the transition from one life cycle stage to another most likely affects the expression of a large number of transcripts, and it is easy again to imagine roles for posttranscriptional regulators such as a pumilio protein in maintaining one of the life cycle stages or in regulating mRNA turnover during phase transition. Given the fact that life cycle phase transition in *E. huxleyi* has only been inferred

Table 4. Most Prevalent mRNA Transcripts

Cluster	Best match	Organism	ESTs
1	No significant match		131
2	No significant match		52
3	No significant match		51
4	Actin	*Emiliania huxleyi*	51
5	Polyubiquitin	*Homo sapiens*	37
6	No significant match		37
7	No significant match		31
8	L5801.5	(*Leishmania major*)	29
9	No significant match		21
10	No significant match		20
11	Sulfite oxidase	*Deinococcus radiodurans*	19
12	No significant match		17
13	No significant match		17
14	Hypersensitive response element	*Hordeum vulgare*	16
15	No significant match		14
16	No significant match		14
17	K-family cellulase homologue	*Fusarium oxysporum*	12
18	α-Adrenergic receptor 2B	*Phoca vitulina*	12
19	Hypothetical protein	*Nostoc punctiforme*	11
20	Hypothetical protein	*Azotobacter vinelandii*	11
21	No significant match		11
22	No significant match		11
23	No significant match		10
24	Cathespin	*Litopenaeus vannamei*	10
25	No significant match		10

from observational (microscopic) data (Klaveness, 1972; Laguna et al., 2001), flow cytometric data (Green et al., 1996) and more recently molecular data (Laguna et al., 2001), this study may provide the means to begin molecular and genetic characterization of the life cycle of this organism.

Several other cDNAs identified through this EST project should help to expand our knowledge of signal transduction pathways in *E. huxleyi*. Multiple copies of a calcium-dependent protein kinase showing significant homology to the green alga *Dunaliella* protein (Pinontoan et al., 2000) and a calcium/calmodulin-dependent protein kinase highly similar to the corresponding protein in *Drosophila* (Adams et al., 2000) were uncovered. Other signal transduction proteins related to the cell cycle and organelle inheritance included cyclin-dependent kinases (Cdks) and a cell cycle initiation mitogen-activated protein kinase with significant homology to the protein described in *Chlamydomonas reinhardtii*.

Knowledge of biomineralization and coccolithogenesis in *E. huxleyi* is in its infancy, and we have yet to unequivocally identify genes involved in these processes. In our library we have, however, found several genes encoding calcium binding proteins and proteins involved in calcium homeostasis. For example, the gene for the previously identified protein that is associated with intracellular precursors of coccolith polysaccharides (Corstjens et al., 1998) was present in our library, as was another acidic uncharacterized protein with a distinct calcium binding motif. The library was also found to contain multiple copies of the genes for both calnexin and calreticulin. Although calnexin and calreticulin reside predominately in the endoplasmic reticulum, the proteins affect many cellular functions both in the ER and outside of the ER environment. Calnexin and calreticulin are chaperones that also play a key role in calcium homeostasis and are known to affect a variety of cellular functions including lectin-like chaperoning, Ca^{2+}

storage and signaling, regulation of gene expression, protein trafficking, and cell adhesion (Michalak et al., 1999; Huang and Beck, 2003). Whether or not these proteins are involved in the regulation of calcium in biomineralization is not known, but preliminary data from Northern analysis in our laboratory indicate transcription of calreticulin is upregulated in *E. huxleyi* cells grown in low-phosphate medium that promotes calcification, as compared with levels in cells grown in rich medium that appears to inhibit calcification.

We expect genes encoding proteins involved in biomineralization and coccolithogenesis to be novel sequences unlikely to be found in GenBank. Hence efforts in our laboratory are also being directed toward the most prevalent uncharacterized genes in the library that are identified in Table 4.

Conclusions

Our initial EST analysis, presented herein, is informative and indicates that the calcifying *E. huxleyi* cells express a complex set of genes. To our knowledge this analysis is the only available genomic resource for *E. huxleyi* and, as such, represents a valuable resource for future work with this important alga. A complete description of the data set is beyond the scope of this work; however, the complete data set will be deposited in GenBank, and efforts to construct an *E. huxleyi* database are underway in our laboratory. We have putatively identified the function of 1086 sequences, but the incomplete nature of EST sequences dictates that any inferred function for a given sequence should be interpreted with caution. Nonetheless, we have provided a conceptual framework of ESTs from which clones may be identified for more complete functional analysis by gene expression profiling, gene silencing or RNA interference, or biochemical characterization. Further studies aimed at gene discovery and functional analysis in *E. huxleyi* will help resolve the underlying mechanisms defining calcification, DMS emissions, and the complex life cycle of this ubiquitous and ecologically important marine organism. These efforts will be greatly facilitated by the Department of Energy's recent selection of *E. huxleyi* for genome sequencing.

Acknowledgments

We thank Dr. Richard Bray for his help and advice in statistical analysis matters and Larry Anderson for his hard work on the organization of our preliminary EST data and identification of redundant sequences from the original raw data files. This work was supported by the National Institutes of Health (grant GM 059833).

References

Adams, M.D., Kerlavage, A.R., and Fleischmann, R.D., et al. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377(6547 Suppl):3–174.

Adams, M.D., Celniker, S.E., and Holt, R.A., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.

Bates, T.S., Charlson, R.J., and Gammon, R.H. (1987). Evidence for the climatic role of marine biogenic sulfur. *Nature* 329:319–321.

Bhattacharya, D., Stickel, S.K., and Sogin, M.L. (1993). Isolation and molecular phylogenetic analysis of actin-coding regions from *Emiliania huxleyi*, a Prymnesiophyta alga, by reverse transcriptase and PCR methods. *Mol Biol Evol* 10:689–703.

Brownlee, C., Nimer, N., Dong, L.F., and Merrett, M.J. (1994). Cellular regulation during calcification in *Emiliania huxleyi*. *Syst Assoc Spec Vol* 51:133–148.

Charlson, R.J., Lovelock, J.E., Andreae, M.O., and Warren, S.G. (1987). Oceanic phytoplankton, atmospheric sulfur, cloud albedo, and climate. *Nature* 326:665–661.

Clark, D.R., and Flynn, K.J. (2000). The relationship between the dissolved inorganic carbon concentration and growth rate in marine phytoplankton. *Proc R Soc Lond B* 267:953–959.

Collins, D.W., and Jukes, T.H. (1993). Relationship between G + C in silent sites of codons and amino acid composition of human proteins. *J Mol Evol* 36:201–213.

Conte, M.N., Eglinton, G., and Madureira, L.A.S. (1992). Long-chain alkenones and alkyl alkenoates as paleotemperature indicators: their production, flux, and early sedimentary diagenesis in the eastern North Atlantic. In: *Advances in Organic Chemistry*, Eckardt, C.B., and Larter, S.R. (eds.). pp 287–298.

Corstjens, P.L.A.M., van der Kooij, A., Linschooten, C., Brouwers, G.-J., Westbroek, P., and de Vrind-de Jong, E.W. (1998). GPA, a calcium -binding protein in the coccolithophorid *Emiliania huxleyi* (Prymnesiophyceae). *J Phycol* 34:622–630.

DeRisi, J.L., and Iyer, V.R. (1999). Genomics and array technology. *Curr Opin Oncol* 11:76–79.

Eide, L.D. (1990). Distribution of coccolithophorids in surface sediments in the Norwegian-Greenland Sea. *Mar Micropaleontol* 16:65–75.

Forbes, A., and Lehmann, R. (1998). Nanos and Pumilio have critical roles in the development and function of *Drosophila* germline stem cells. *Development* 125:679–690.

Foster, P.G., Jermiin, L.S., and Hickey, D.A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44:282–288.

Green, J.C., Course, P.A., and Tarran, G.A. (1996). The life cycle of *Emiliania huxleyi*: a brief review and a study of ploidy levels analysed by flow cytometry. *J Mar Syst* 9:33–44.

Gu, X., Hewett-Emmett, D., and Li, W.H. (1998). Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102–102:383–391.

Guillard, R.R.L. (1975). Culture of phytoplankton for feeding marine invertebrates. In: *Culture of Marine Invertebrate Animals*, Smith, W.L., and Chanley, M.H. (eds.). New York, N.Y.: Plenum Press, pp 29–60.

Huang, K., and Beck, C.F. (2003). Phototropin is the blue-light receptor that controls multiple steps in the sexual life cycle of the green alga *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci USA* 100:6269–6274.

Klaveness, D. (1972). *Coccolithus huxleyi* (Lohm). Kamptn, II: the flagellate cell, aberrant cell types, vegetative propagation and life cycles. *Br Phycol J* 7:309–318.

Kurvari, V., Grishin, N.V., and Snell, W.J. (1998). A gamete-specific sex-limited homeodomain protein in *Chlamydomonas*. *J Cell Biol* 143:1971–1980.

Laguna, R., Romo, J., Read, B.A., and Wahlund, T.M. (2001). Induction of phase variation events in the life cycle of the marine coccolithophorid *Emiliania huxleyi*. *Appl Environ Microbiol* 67:3824–3831.

Lobry, J.R. (1997). Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205:309–316.

Lowery, R.G., and Ludden, P.W. (1988). Purification and properties of the dinitrogenase reductase ADP-ribosyltransferase from *Rhodospirillum rubrum*. *J Biol Chem* 263:16714–16719.

Michalak, M., Corbett, E.F., Mesaeli, N., Nakamura, K., and Opas, M. (1999). Calreticulin: one protein, one gene, many functions. *Biochem J* 344(Pt 2):281–292.

Paasche, E. (2002). A review of the coccolithophorid *Emiliania huxleyi* (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycol Rev* 20 40:503–529.

Parisi, M., and Lin, H. (1999). The *Drosophila pumilio* gene encodes two functional protein isoforms that play multiple roles in germline development, gonadogenesis, oogenesis and embryogenesis. *Genetics* 153:235–250.

Pinontoan, R., Yuasa, T., Anderca, M.I., Matsuoka, T., Uozumi, N., Mori, H., and Muto, S. (2000). Cloning of a cDNA encoding a 66-kDa Ca^{2+}-dependent protein kinase (CDPK) from *Dunaliella tertiolecta* (Chlorophyta). *J Phycol* 36:545–552.

Prahl, F.G., Muehlhavsen, L.A., and Zahnle, D.L. (1988). Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. *Geochim Cosmochim Acta* 52:2303–2310.

Raven, J.A. (1997). Putting the C in phycology. *Eur J Phycol* 32:319–333.

Reinfelder, J.R., Kraepiel, A.M., and Morel, F.M. (2000). Unicellular C_4 photosynthesis in a marine diatom. *Nature* 407:996–999.

Riegman, R., Stolte, W., Noordeloos, A.A.M., and Slezak, D. (2000). Nutrient uptake and alkaline phosphatase (EC 3:1:3:1) activity of *Emiliania huxleyi* (Prymnesiophyceae) during growth under N and P limitation in continuous cultures. *J Phycol* 36:87–96.

Rubio, V., Linhares, F., Solano, R., Martin, A.C., Iglesias, J., Leyva, A., and Paz-Ares, J. (2001). A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. *Genes Dev* 15:2122–2133.

Samtleben, C., and Bickert, T. (1990). Coccoliths in sediment traps from the Norwegian Sea. *Mar Micropaleontol* 16:39–64.

Serikawa, K.A., and Mandoli, D.F. (1999). *Aaknox1*, a *kn1*-like homeobox gene in *Acetabularia acetabulum*, undergoes developmentally regulated subcellular localization. *Plant Mol Biol* 41:785–793.

Sikes, E.L., Farrington, J.W., and Keigwin, L.D. (1991). Use of alkenone unsaturation ratios to determine past sea surface temperatures: core-top SST calibrations and methodology considerations. *Earth Planet Sci Lett* 104:36–47.

Souza, G., Lu, S., and Kuspa, A. (1998). YakA, a protein kinase required for the transition from growth to development in *Dictyostelium*. *Development* 125:2291–2302.

Souza, G.M., da Silva, A.M., and Kuspa, A. (1999). Starvation promotes Dictyostelium development by relieving PufA inhibition of PKA translation through the YakA kinase pathway. *Development* 126:3263–3274.

Steeman, N.E. (1966). The uptake of free CO_2 and HCO_3^- during photosynthesis of plankton algae with special reference to the coccolithophorid *Coccolithus huxleyi*. *Physiol Plantarum* 19:232–240.

Strommer, J., Gregerson, R., and Vayda, M. (1993). Isolation and characterization of plant mRNA. In: *Methods in Plant Molecular Biology and Biotechnology*, Glick, B.R., and Thompson, E. (eds.). Boca Raton, Fla.: CRC Press, pp 49–66.

Tadauchi, T., Matsumoto, K., Herskowitz, I., and Irie, K. (2001). Post-transcriptional regulation through the HO 3′-UTR by Mpt5, a yeast homolog of Pumilio and FBF. *EMBO J* 20:552–561.

Tollervey, D., and Caceres, J.F. (2000). RNA processing marches on. *Cell* 103:703–709.

Walsh, D., and Mann, S. (1995). Fabrication of hollow porous shells of calcium carbonate from self-organizing media. *Nature* 377:320–323.

Westbroek, P., Brown, C.W., and Van Bleuswijk, J, et al. (1993). A model system approach to biological climate forcing: the example of *Emiliania huxleyi*. *Global Planetary Change* 8:27–46.

Zhang, B., Gallegos, M., Puoti, A., Durkin, E., Fields, S., Kimble, J., and Wickens, M.P. (1997). A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390:477–484.