Developmental and Comparative Immunology 33 (2009) 881–889

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

The diversity of C-type lectins in the genome of a basal metazoan, *Nematostella vectensis*

Elisha M. Wood-Charlson [a,b,*], Virginia M. Weis [b]

[a] Department of Oceanography, University of Hawai'i at Manoa, Honolulu, HI 96822, United States
[b] Department of Zoology, Oregon State University, Corvallis, OR 97331, United States

ARTICLE INFO

Article history:
Received 16 December 2008
Received in revised form 29 January 2009
Accepted 30 January 2009
Available online 21 February 2009

Keywords:
Cnidarian
C-type
Glycan
Innate immunity
Lectin
Nematostella
Symbiosis

ABSTRACT

C-type lectins (CTLs) are involved in cell–cell adhesion, recognition, and innate immunity in higher vertebrates, but little is known about CTLs in basal metazoans. The recent sequencing of the cnidarian *Nematostella vectensis* genome allowed us to explore the CTL-like gene family at the base of metazoan evolution. Sixty-seven predicted CTLs, with a total of 92 putative C-type lectin domains (CTLDs), were classified according to number of CTLDs present and their association with other protein domains in the CTL. Conserved residues in the glycan-binding pocket suggest that approximately half of the CTLDs retain glycan-binding function. Phylogenetic analysis of *N. vectensis* CTLDs with respect to other model invertebrates and humans indicates *N. vectensis* CTLD sequences more closely resemble vertebrate CTLDs. This study provides a *N. vectensis* CTL database that can be used for further research on the evolution of cnidarian CTLs and the role of CTLs in cnidarian innate immunity.

1. Introduction

The recent expansion of genome sequencing and the generation of large EST datasets is beginning to allow for studies on the evolution of gene families. Many of these studies have focused on gene families that have been conserved throughout metazoan evolution [1–5]. Most of the major gene families involved in conserved, developmentally regulated, vertebrate signaling pathways have also been identified in cnidarians, a basal metazoan group [4,6] (Fig. 1). Recently, similar comparisons have begun for less well-conserved gene families, such as receptors that function in innate immunity. These immunity gene families, which are also present in cnidarians [7], are likely involved in "trench warfare"—continuous, diversifying selection pressure imposed by host–microbe interactions [8,9]. One example of an innate immune receptor family is the C-type lectin (CTL) gene family, which has been described for model vertebrate and invertebrate organisms [10–13], but has not been investigated for basal metazoans.

CTLs make up a very diverse gene family that is identified by a C-type lectin-like domain (CTLD), which bind glycans in a Ca^{2+}-dependent manner. The CTLD ranges from 115 to 130 amino acids in length and can be identified by 14 invariant and 18 highly conserved amino acids [14]. Most of the conserved residues occur within two functional sites: the first binds a single Ca^{2+} ion and the second binds another Ca^{2+} and the glycan ligand [15] (Fig. 2). Outside of these motifs, CTLD sequences are not well conserved, and CTLs often contain more than one CTLD in addition to numerous other domains that determine many of the CTL functions [16].

CTLDs play a role in a variety of biological events that require recognition of specific glycans, such as cell–cell adhesion, recognition, and phagocytosis of potential pathogens [17–19]. Since cell–cell adhesion and immunity have been major contributors to metazoan evolution [20,21], understanding the diversity within the CTL gene family will provide valuable insight into ancestral metazoan complexity. Analyses of CTLDs from two invertebrate model organisms, *Drosophila* and *Caenorhabditis*, suggested that CTLs diverged dramatically since the split between vertebrate and invertebrate lineages [22,23]. However, no study has examined the diversity of the CTL gene family in basal metazoans, such as cnidarians, which harbor a surprising amount

* Corresponding author at: 1000 Pope Road, Marine Sciences Building, Honolulu, HI 96822, United States. Tel.: +1 808 956 7633; fax: +1 808 956 9225.
E-mail address: woodchae@lifetime.oregonstate.edu (E.M. Wood-Charlson).

Abbreviations: CTL, C-type lectin; CTLD, C-type lectin-like domain; EGF, epidermal growth factor; F58C, coagulation factor 5/8 type C; Gal, galactose; Gal-lectin, D-galactoside-binding lectin; Glc, glucose; GPS, G-protein coupled receptor (GPCR) proteolytic site; Ig, immunoglobulin-like; LDLRA, low-density lipoprotein receptor A; Man, mannose; SCP, sperm-coating glycoprotein; 7tm, 7-pass transmembrane GPCR; Signal, signal sequence; SRCR, scavenger receptor cysteine rich; tm, transmembrane domain; TSP1, thrombospondin type I repeat; vWA, von Willebrand factor A.

0145-305X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2009.01.008

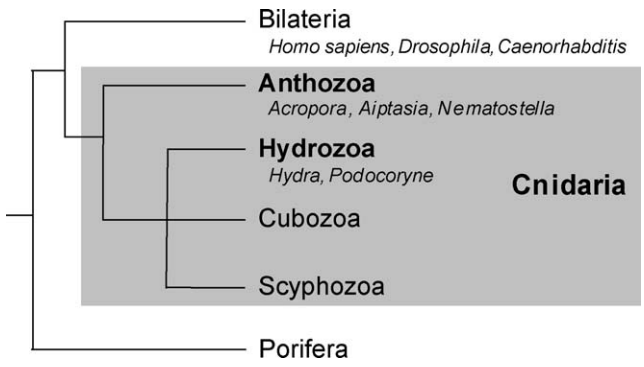

Fig. 1. Cnidarians as a basal group in metazoan evolution. Within the Cnidaria, the Anthozoa are basal and the Hydrozoa are derived (modified from [7]).

of genomic complexity that, in many cases, more closely reflects vertebrate-level complexity than the other invertebrate model organisms [2–4,24,25]. Aside from preliminary genomics work, there are a few studies that have explored the functional role of lectins in cnidarians. Cnidarians, such as anemones and corals, are often found in a mutualistic relationship with a single-celled dinoflagellate. In most cases, these very selective relationships must be established anew for every host generation and requires a complex series of steps. These steps include recognition between the appropriate symbiotic partners and phagocytosis of the symbiont by the host (reviewed in [26,27])—both known functions of lectin/glycan interactions. During the onset of symbiosis, lectin/glycan interactions appear to provide a mechanism of recognition between host cnidarians and their algal symbionts [28–30]. There is also evidence that cnidarian lectins can harbor extensive sequence variation and be used to bind potential bacterial pathogens, as well as potential symbionts [31]. The identity, localization, and function of cnidarian lectins during other stages in the symbiont selection process or other aspects of innate immunity remain a mystery.

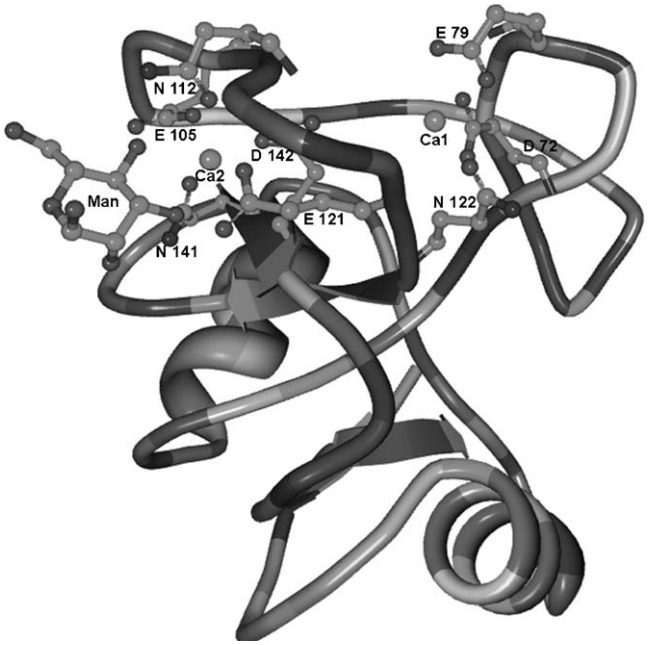

Fig. 2. Structure of a typical CTLD in complex with a glycan ligand (generated from Protein Data Bank: 1kzc, www.rcsb.org/pdb). The residues that create the two Ca^{2+}-binding pockets are identified by their single-letter amino acid abbreviation and sequence position that corresponds to the alignment in Fig. 3. The second Ca^{2+}-binding pocket also binds the glycan ligand (shown here in association with Man).

Cnidarians, such as the non-symbiotic sea anemone *Nematostella vectensis*, represent a sister-group to the Bilateria (Fig. 1). *N. vectensis* has become a cnidarian model for developmental studies, and the recent sequencing of its genome allows for analyses of gene families in a basal metazoan [2]. This study identifies and describes members of the CTL gene family in *N. vectensis* for future use in evolutionary and functional innate immunity studies.

2. Materials and methods

2.1. Database searching

Using the Joint Genome Institute (JGI) interactive *N. vectensis* genome browser (v1.0) (http://genome.jgi-psf.org/Nemve1/Nemve1.home.html), searches were conducted to retrieve potential CTL-like sequences from the genome. A tblastn search ($e < 10^{-5}$) was performed using CTLD sequences from human, mouse, and 13 invertebrates (query sequences are available by request). In addition, the annotated *N. vectensis* genome was searched for key terms: lectin (280 hits), C-type (201 hits), mannose (18 hits), galactose (16 hits), and InterPro C-type lectin (IPR001304, 89 hits). CTLD-containing sequences from other cnidarians were identified by a tblastn search ($e < 10^{-5}$) of the cnidarian sequences (taxonomy ID: 6073) at the National Center for Biotechnology Information (NCBI), the *Hydra magnipapillata* genome browser (v1.64, http://hydrazome.metazome.net/cgibin/gbrowse/hydra/), and an EST database from *Aiptasia pallida* (http://aiptasia.cs.vassar.edu/AiptasiaBase/index.php) [32].

2.2. Assessing gene prediction models

For all CTLD-containing sequences, the coding region was identified by the *Ab initio* model using Fgenesh gene prediction (available on the JGI genome browser) and only modified when EST data were available. In some cases, the *Ab initio* model may have missed a potential CTLD coding sequence and/or not predicted the full-length CTLD; however, the *Ab initio* model was selected for consistency because it predicted a gene model for all putative CTLD sequences retrieved from the database searches. Using a combination of methods, the predicted CTL sequences were screened to confirm the presence of at least one CTLD: (1) blast search against the NCBI protein database resulted in a significant hit to a CTLD-containing sequence ($e < 10^{-5}$) and (2) protein domain search against PfamA [33], ScanProsite [34], and InterProScan [35] resulted in a CTLD profile hit from more than one database. The complete list of CTLD-containing gene predictions from the *N. vectensis* genome can be found in Table 1 and viewed on the JGI genome browser by searching Gene Models with the gene model name as listed.

2.3. Identification of domains within gene predictions

Using the compiled dataset of *N. vectensis* CTL-like sequences, additional protein domains within the predicted coding regions were identified by PfamA, ScanProsite, and InterProScan. Domains were included if they were predicted by more than one annotation method or if a single annotation method had a significant hit to the domain ($e < 10^{-5}$). Putative signal sequences and transmembrane domains were identified by InterProScan and confirmed by hydropathy plots.

2.4. Sequence analysis

CTLDs were trimmed at the longest disulfide bond and aligned using ClustalX [36]. CTLDs have been predicted to have functional glycan-binding if they contain at least three out of the five conserved residues known to function in the Ca^{2+}/glycan-binding

Table 1
Nematostella vectensis CTL-like sequences retrieved from the JGI genome browser (v1.0).

Group	Domain structure[a]	Gene model name[b]	Description[a,c]
A	Single CTLD	e_gw.1517.2.1	CTLD alone
		e_gw.2934.3.1	CTLD alone
		e_gw.7241.1.1	CTLD alone
		estExt_fgenesh1_pg.C 110046	CTLD alone
		estExt_fgenesh1_pg.C 170085	CTLD alone
		fgenesh1_pg.scaffold 12000058	CTLD alone
		fgenesh1_pg.scaffold 17000085	CTLD alone
		fgenesh1_pg.scaffold 3000113	CTLD alone
		fgenesh1_pg.scaffold 3035000002	CTLD alone
		fgenesh1_pg.scaffold 327000008	CTLD alone
		fgenesh1_pg.scaffold_4914000001	CTLD alone
		fgenesh1_pg.scaffold 748000003	CTLD alone
		fgenesh1_pg.scaffold 156000026	signal/tm, CTLD
		fgenesh1_pg.scaffold 19000102	signal/tm, CTLD
		fgenesh1_pg.scaffold 3283000001	signal/tm, CTLD
		fgenesh1_pg.scaffold 50000070	signal/tm, CTLD
		fgenesh1_pg.scaffold 542000003	signal/tm, CTLD
		fgenesh1_pg.scaffold 7000235	signal/tm, CTLD
		fgenesh1_pg.scaffold 76000071	signal/tm, CTLD
		fgsh_est.C_scaffold_15000009	signal/tm, tm, CTLD
		estExt_fgenesh1_pg.C 5290004	signal, CTLD, tm
		fgenesh1_pg.scaffold_121000039	signal, CTLD
		fgenesh1_pg.scaffold 4000026	signal, CTLD
		fgenesh1_pg.scaffold 66000008	signal, CTLD
		estExt_fgenesh1_pg.C 17400001	tm, CTLD
		fgenesh1_pg.scaffold 117000056	tm, CTLD
		fgenesh1_pg.scaffold 76000072	tm, CTLD
		fgenesh1_pg.scaffold 685000004	CTLD, tm (2)
		fgenesh1_pg.scaffold 256000002	tm, CTLD, tm (2)
B	2 CTLDs	fgenesh1_pg.scaffold 150000031	signal, CTLD (2), tm
C	vWA	fgenesh1_pg.scaffold 7000087	signal, CTLD (2), vWA
		fgenesh1_pg.scaffold 161000025	signal, CTLD (2), vWA
		fgenesh1_pg.scaffold 161000026	CTLD (2), vWA
D	SCP	estExt_fgenesh1_pg.C 1010028	signal, SCP, CTLD
		estExt_fgenesh1_pg.C 4250015	signal, SCP, CTLD
		fgenesh1_pg.scaffold 425000013	signal, SCP, CTLD
E	EGF	fgenesh1_ pg.scaffold 121000042	EGF, CTLD
		fgenesh1_pg.scaffold 18000071	signal, EGF, CTLD
		fqenesh1_pg.scaffold 748000002	signal, EGF (3), CTLD
F	EGF + Kazal_2	fgenesh1_pg.scaffold 7191000001	kazal 2, CTLD
		fgenesh1_pg.scaffold 266000022	signal, kazal 2, CTLD, EGF (2)
		fgenesh1_pg.scaffold 367000012	signal, kazal 2, CTLD, EGF (3)
		fgenesh1_pg.scaffold 413000002	signal, kazal 2, tm (4), CTLD, EGF (4)
		fgenesh1_pg.scaffold 655000001	kazal 2, CTLD, EGF (3)
		fgenesh1_ pg.scaffold 89000004	signal, kazal 2, CTLD, EGF (4)
G	Link	fgenesh1_ pg.scaffold 121000041	Link, CTLD
		fqenesh1_pg.scaffold 3000110	signal/tm, Link, CTLD
H	GPS/7tm	fgenesh1_ pg.scaffold 31000058	signal/tm, CTLD, GPS + 7tm
		fgenesh1_ pg.scaffold 54000011	signal, CTLD, GPS + 7tm
		fqenesh1_pg.scaffold 48000036	signal, LDLRA, CTLD, GPS + 7tm
I	F58C	fgenesh1_pg.scaffold 12000147	signal/tm (2), F58C, CTLD (2)
	F58C + Ig	fgenesh1_pg.scaffold 120000043	F58C (2), CTLD, Ig
	F58C + Sushi	fqenesh1_pg.scaffold 15000042	signal, F58C, Sushi (12), CTLD
J	10 CTLDs + F58C	estExt_fgenesh1_pg.C 30109	signal, CTLD (3), F58C, CTLD (7), F58C (3), SEA, tm
		fgenesh1_pg.scaffold 3000106	signal, CTLD (3), F58C, CTLD (2), Gal Lectin, CTLD (5), F58C (3), SEA
K	TSP1 + Astacin	fgenesh1_pg.scaffold 146000014	Astacin, TSP1, CTLD, TSP1 (2), tm
	TSP1	fgenesh1_pg.scaffold 425000001	signal, TSP1 (2), CTLD
L	CUB	fgenesh1_pg.scaffold 1000051	CUB (2), CTLD
M	MAM + SRCR	fgenesh1_pg.scaffold 37000121	SRCR, MAM, CTLD
	MAM + LDLRA/CUB	fgenesh1_pg.scaffold 28000132	signal, MAM, CTLD (2), MAM, LDLRA (3), MAM (2) CUB
N	Ig	estExt fgeneshi pg.C 490082	signal, Ig (3), CTLD
	Ig + Collagen	estExt fgeneshi pg.C 270065	signal/tm, collagen, Ig (3), CTLD
	Ig + PAN	fgenesh1_pg.scaffold 42000068	CTLD, Ig, PAN
	PAN	fgenesh1_pg.scaffold 156000038	signal/tm, PAN, CTLD
		fqenesh1_pg.scaffold 42000073	CTLD, PAN

Table 1 (*Continued*)

Group	Domain structure[a]	Gene model name[b]	Description[a,c]
O	ShK	fgenesh1_pg.scaffold 68000060	signal, ShK, tm (6), CTLD (2)
P	Cadherin	fgenesh1_pg.scaffold 7000012	tm, Cadherin (3), CTLD

[a] *Abbreviations*: CTL: C-type lectin, CTLD: C-type lectin-like domain, EGF: epidermal growth factor, F58C: coagulation factor 5/8 type C, Gal: galactose, Gal-lectin: D-galactoside-binding lectin, GPS: G-protein coupled receptor (GPCR) proteolytic site, Ig: immunoglobulin-like, LDLRA: low-density lipoprotein receptor A, Man: mannose, SCP: sperm-coating glycoprotein, 7tm: 7-pass transmembrane GPCR, Signal: signal sequence, SRCR: scavenger receptor cysteine rich, tm: transmembrane domain, TSP1: thrombospondin type I repeat, vWA-von Willebrand factor A.

[b] *Ab initio* gene sequence predictions are available by searching "Gene Model name equals" at http://genome.jgi-psf.org/Nemve1/Nemve1.home.html.

[c] Presented in order of appearance in the gene prediction model. Additional domains were identified using PfamA, ScanProsite, and InterProScan. Domains unique to *N. vectensis* CTLs are underlined. The number in parentheses following a domain indicates number of consecutive repeats of that domain in the gene prediction.

pocket [15,37]. Using these criteria, which were previously used to describe CTLDs in the model invertebrates [12,22], a subset of *N. vectensis* CTLDs were identified as sequences likely to bind glycans. Alignments were finalized with BioEdit [38] to maintain the position of conserved residues [14]. Bayesian analyses were performed using MrBayes 3.1 [39] with a mixed model of protein evolution: (1) all CTLDs recovered from the *N. vectensis* genome in addition to unique CTLDs found in other cnidarians and (2) *N. vectensis*, *H. sapiens*, *C. elegans*, and *D. melanogaster* CTLDs containing at least three of the five conserved Ca^{2+}/glycan-binding pocket residues (alignments are available as Supplementary files). The analysis of *N. vectensis* and other cnidarian CTLDs was run on four chains for 5,000,000 generations, and after a burnin of 2500 generations, every 500th tree was sampled for 50% majority consensus. A similar analysis (10,000,000 generations, 2500 generation burnin, and sampled every 1000th tree) was performed for *N. vectensis*, *H. sapiens, C. elegans*, and *D. melanogaster* CTLDs.

3. Results

3.1. Description of CTL-like sequences in the N. vectensis genome

The *N. vectensis* genome was found to contain 92 putative CTLDs within a total of 67 gene predictions. Additional CTLDs may be present in the genome and could be discovered using less restrictive methods; however, the purpose of this study was to provide a conservative list of potentially functional CTLDs, not to identify all possible sequences.

Of the 67 CTLD-containing gene predictions in *N. vectensis*, 43% contained a single CTLD with no additional domains. Further classification of the remaining CTLD-containing sequences was made based on the number of CTLDs and the organization of other domains in the predicted coding sequence. This method of CTLD classification is similar to an analyses performed on the *C. elegans* genome [12,13]. In *N. vectensis*, this classification method resulted in sixteen groupings: A. single CTLD; B. two CTLDs; C. two CTLDs and von Willebrand factor A domains (vWA); D. CTLD and sperm-coating glycoprotein (SCP); E. CTLD and epidermal growth factor (EGF); F. CTLD, EGF, and Kazal_2; G. CTLD and Link; H. CTLD, G-protein coupled receptor (GPCR) proteolytic site (GPS), and a 7-pass transmembrane GPCR (7-tm); I. CTLD and coagulation factor 5/8 type C (F58C); J. ten CTLDs; K. CTLD and thrombospondin type I repeat (TSP1); L. CTLD and CUB; M. CTLD and MAM; N. CTLD, Immunoglobulin-like (Ig) and/or PAN domains; O. two CTLDs and ShK; P. CTLD and cadherin (Table 1). Additional subgroups were created for sequences containing a signal sequence and/or transmembrane domains, or if the orientation, identity, and number of domains within the gene prediction changed. Some of these subgroups may be artifacts of the gene prediction model, especially subgroups created by the presence/absence of a signal sequence or transmembrane domain. Full-length sequencing of the *N. vectensis* CTL-like coding sequences may identify additional localization domains, which could further condense these subgroups.

3.2. Analysis of potential carbohydrate-binding sites

The 92 *N. vectensis* CTLDs were isolated from the rest of the coding sequence by trimming the CTLD at cysteine residues that created the longest disulfide bond. The CTLD domains were aligned, and any CTLDs that lacked the outer cysteine residues were trimmed to the alignment (Supplementary Fig. 1). To identify CTLDs that most likely complex with a glycan ligand, a subset of *N. vectensis* CTLD sequences was selected because they contained at least three of the five residues that form the Ca^{2+}/glycan-binding pocket [15] (Fig. 3). The binding pocket was completely conserved in 26% of CTLDs in the *N. vectensis* genome and 48% contained most of the conserved residues. This subset of cnidarian CTLDs was compared to CTLDs from the *C. elegans* and *D. melanogaster* genomes. A recent re-examination of the *C. elegans* genome uncovered 278 CTLDs, but only 25% contained three of the five conserved glycan-binding pocket residues [13]. In *D. melanogaster*, 19% of the 32 total CTLDs contained most of the residues that form the Ca^{2+}/glycan-binding site [22].

In addition to predicting functional glycan-binding, the binding pocket also predicts general glycan specificity for Man/glucose (Glc) or Gal derivatives. Of the 92 CTLDs found in the *N. vectensis* genome, thirteen can be designated as Man/Glc-binding and four as Gal-binding. CTLDs in these general binding groups have been shown, both structurally and experimentally, to rely on a triplet motif within the Ca^{2+}/glycan-binding pocket [37,40]. Man/Glc and Gal-derivatives differ in the structural orientation of their 3- and 4-OH groups. Man/Glc ligands, with equatorial 3- and 4-OH, are recognized by CTLDs that contain a Glu-Pro-Asn (EPN) triplet motif, and Gal ligands, with equatorial 3-OH and axial 4-OH, are recognized by a Gln-Pro-Asp (QPD) triplet motif [14,41].

3.3. Organization of CTLs in the genome

The *N. vectensis* genome browser revealed that 72% of CTLDs were contained within a single exon, and most CTLDs were located near the C-terminus of the predicted gene. In addition, 18% of CTLDs may have been cut short by the *Ab initio* model for gene prediction; therefore, predictions that did not cover the full CTLD were allowed to remain in the dataset (Supplementary Fig. 1).

3.4. Phylogenetic analysis of CTLDs in N. vectensis

In addition to the 92 CTLDs from *N. vectensis*, many unique CTLDs were identified from other cnidarian sequencing projects, including two anthozoans—*Acropora millepora* and *A. pallida*, and several hydrozoans—*Hydra* spp. and *Podocoryne carnea* (Fig. 1). A Bayesian analysis was conducted on an alignment of cnidarian CTLDs (Supplementary alignment for Fig. 4). The resulting tree showed tight clusters for ~80% of CTLDs but very little resolution between clusters (complete tree available as Supplementary Fig. 2). Twenty-three out of the 28 *H. magnipapillata* sequences grouped together and were combined into a single consensus sequence (RTK consensus). These CTLDs occur in the extracellular

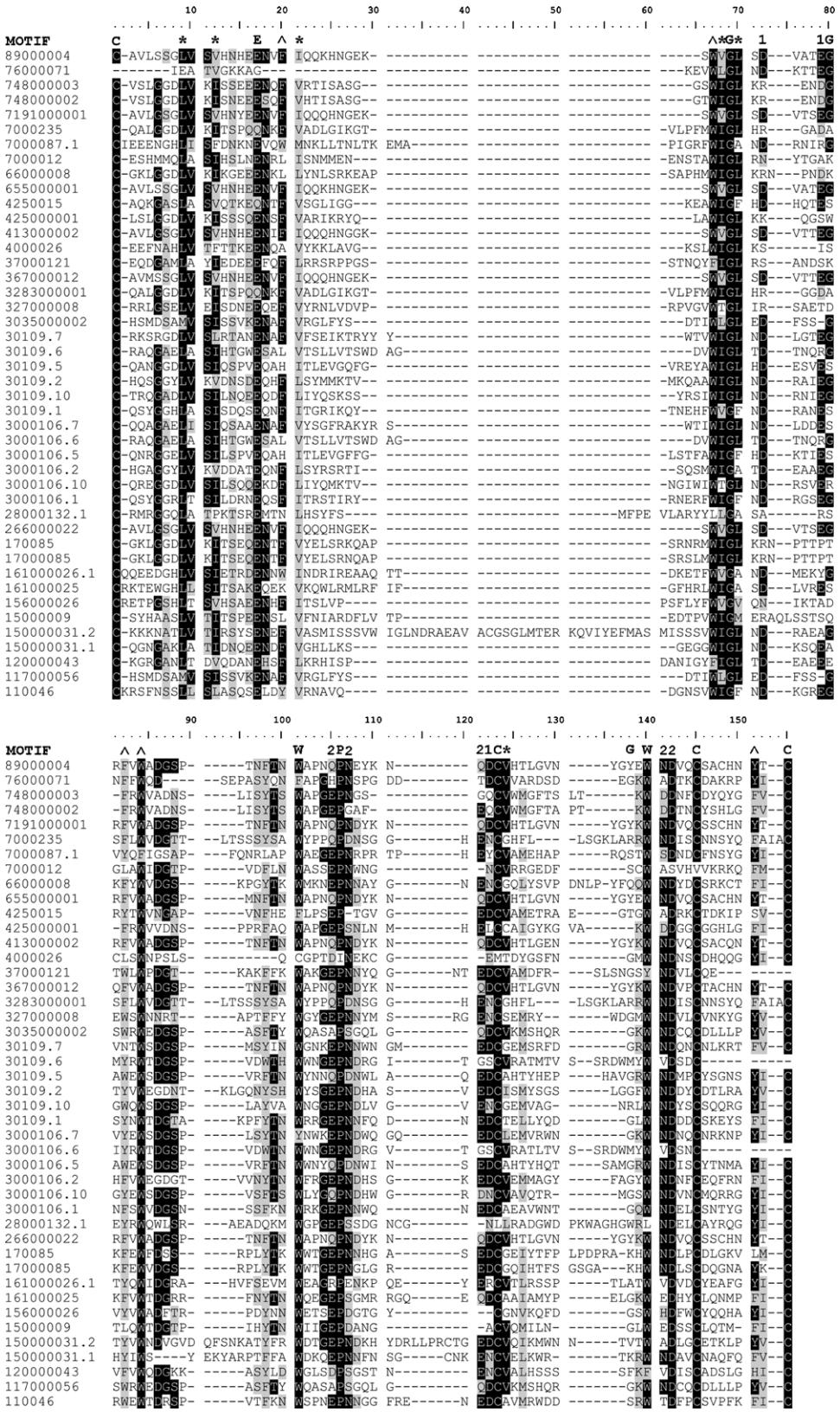

Fig. 3. Alignment of CTLDs found in the *Nematostella vectensis* genome that contain three out of the five residues shown to structurally form the Ca^{2+}/glycan-binding pocket (Fig. 2). Potential ligands for binding the first or second Ca^{2+} are denoted 1 and 2, respectively. The conserved cysteine residues, which form disulfide bonds are denoted by C. Conserved aliphatic and aromatic residues are represented by * and ^, respectively. Other conserved residues are identified by their single letter amino acid abbreviation. Black highlighting indicates at least 50% identity and grey indicates similar chemistry at a given position. Sequences were named by the numerical identifier from their full gene model name as assigned on the *N. vectensis* JGI genome browser (v1.0). Full-length gene names are listed in Table 1. Sequence names followed by a decimal after the numerical identifier indicate that multiple CTLDs were present in a single gene prediction, and the decimal number corresponds to CTLD in order of appearance in the sequence.

domain of *H. magnipapillata* tyrosine kinase receptors (e.g. [42]), and have likely diverged within the cnidarian CTL gene family since the split between Anthozoa and Hydrozoa (Fig. 1).

Even with little resolution, many *N. vectensis* CTLDs clustered with other CTLDs in their classification group. For example, five of the single CTLDs (group A) clustered together along with single CTLDs from *A. millepora*, including a consensus sequence for the Millectin sequence variants [31] (Fig. 4). *N. vectensis* CTLs containing EGF + Kazal_2 sequences (group F) made up a single cluster. Sequences containing 2 CTLDs and vWA (group C) clustered such that the CTLDs grouped in order of appearance between sequences rather than with the other CTLD in the same sequence. This was also seen for other sequences with more than one CTLD (e.g. group J with 10 CTLDs and group B with 2 CTLDs); the CTLDs typically clustered with CTLDs from other sequences in their group rather than other CTLDs in the same sequence (group J is highlighted by stars in Fig. 4 and Supplementary Fig. 2). However, in some cases, CTLDs associated with different domains, but found adjacent to each other on the genome browser, clustered tightly together (highlighted by the ellipse in Supplementary Fig. 2), which suggests that CTLDs may undergo domain shuffling between nearby sequences and maybe a reason for the diversity of domains found in CTLs throughout the *N. vectensis* genome.

Phylogenetic analyses of cnidarian CTLDs may help predict the structure of full-length coding sequences from other cnidarians. For example, we compared *A. pallida* EST sequences that clustered with EGF + Kazal_2 (group F) against the PfamA database. Several of the sequences (Contig 973, 1204; Fig. 4) contained a single CTLD with additional EGF domains, similar to the *N. vectensis* group F (Table 1). Our Bayesian analysis predicts that full-length *A. pallida* sequences will contain a N-terminal Kazal_2 domain and 2–4 total EGF domains. Finally, two sequences, one from *A. pallida* and one from *H. magnipapillata* (highlighted by an arrows in Fig. 4) clustered with *N. vectensis* group J. The *A. pallida* sequence clustered with the tenth CTLD in *N. vectensis*, which immediately precedes a F58C domain, and PfamA analysis of the *A. pallida* sequence identified a CTLD followed by a F58C domain. Full-length sequencing may provide *A. pallida* and *H. magnipapillata* sequences with structure similar to group J from the *N. vectensis* genome.

Since, the Ca^{2+}/glycan-binding pocket was conserved in many of the *N. vectensis* CTLD sequences, we wanted to determine where they fit in an evolutionary context. *N. vectensis* CTLDs containing at least 3 out of the 5 conserved binding pocket residues were analyzed against conserved CTLDs from *H. sapiens*, *C. elegans*, and *D. melanogaster*. Even for less-well conserved gene families, such as CTLs, the resulting tree demonstrates that the cnidarian CTL gene family more closely resembles the CTL gene family in higher vertebrates than in other model invertebrates (Fig. 5). *D. melanogaster* and *C. elegans* CTLD sequences grouped independent of each other and of *N. vectensis* and human CTLDs. In contrast, *N. vectensis* and human CTLDs showed very little separation between their sequences except for terminal human sequences (labeled on Fig. 5) that likely underwent gene duplication after metazoans diverged from the cnidarian lineage.

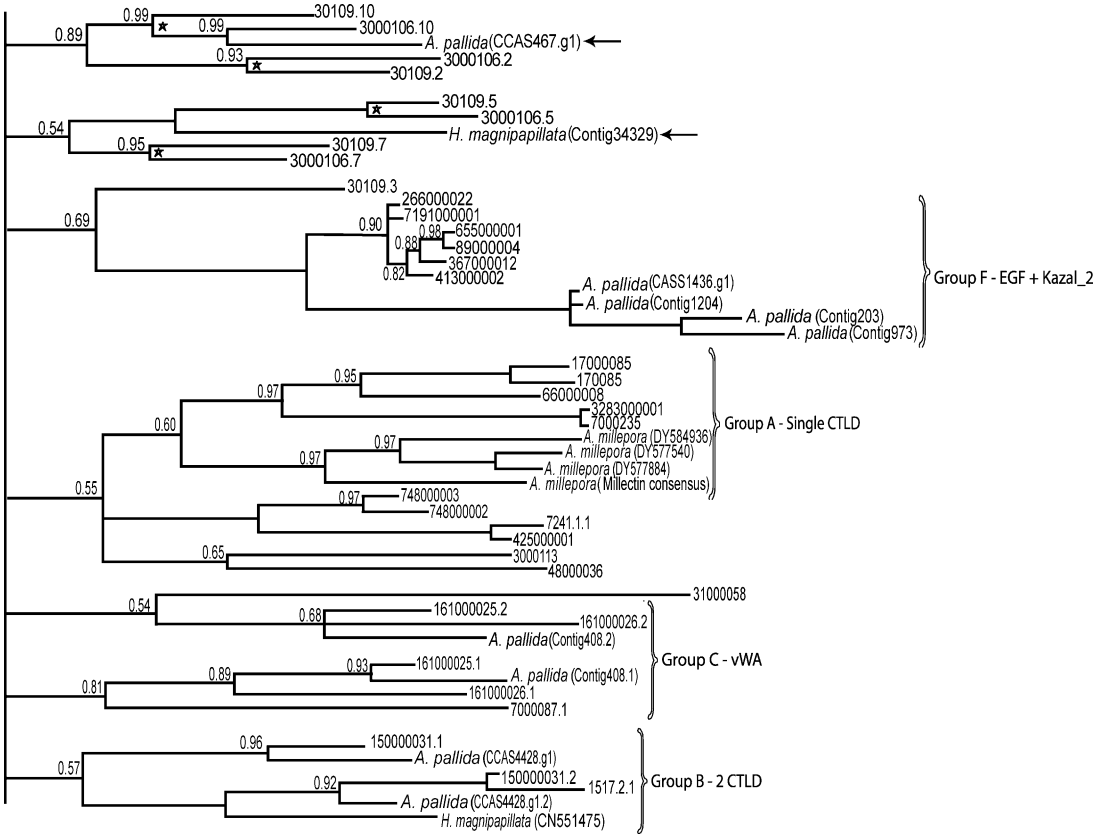

Fig. 4. Selection of branches from Bayesian analysis tree of the CTLD gene predictions from the *Nematostella vectensis* genome and other cnidarian sequencing projects (complete tree available as Supplementary Fig. 2). Numbers at each branch represent posterior probabilities, with unlabeled branches indicating a posterior probability of 1.0. *N. vectensis* sequences were named as in Fig. 3. CTLD groups (referenced in Table 1) have been bracketed or highlighted with an arrow, and starred branches identify *N. vectensis* CTLD group J. Sequences from other cnidarians were named by species and either a Genbank accession number or current identifier from the sequencing project. Sequence names followed by a decimal after the numerical identifier indicate that multiple CTLDs were present in a single gene prediction, and the decimal number corresponds to CTLD in order of appearance in the sequence. Abbreviations: *A. millepora*: *Acropora millepora*, *A. pallida*: *Aiptasia pallida*, *H. magnipapillata*: *Hydra magnipapillata*.

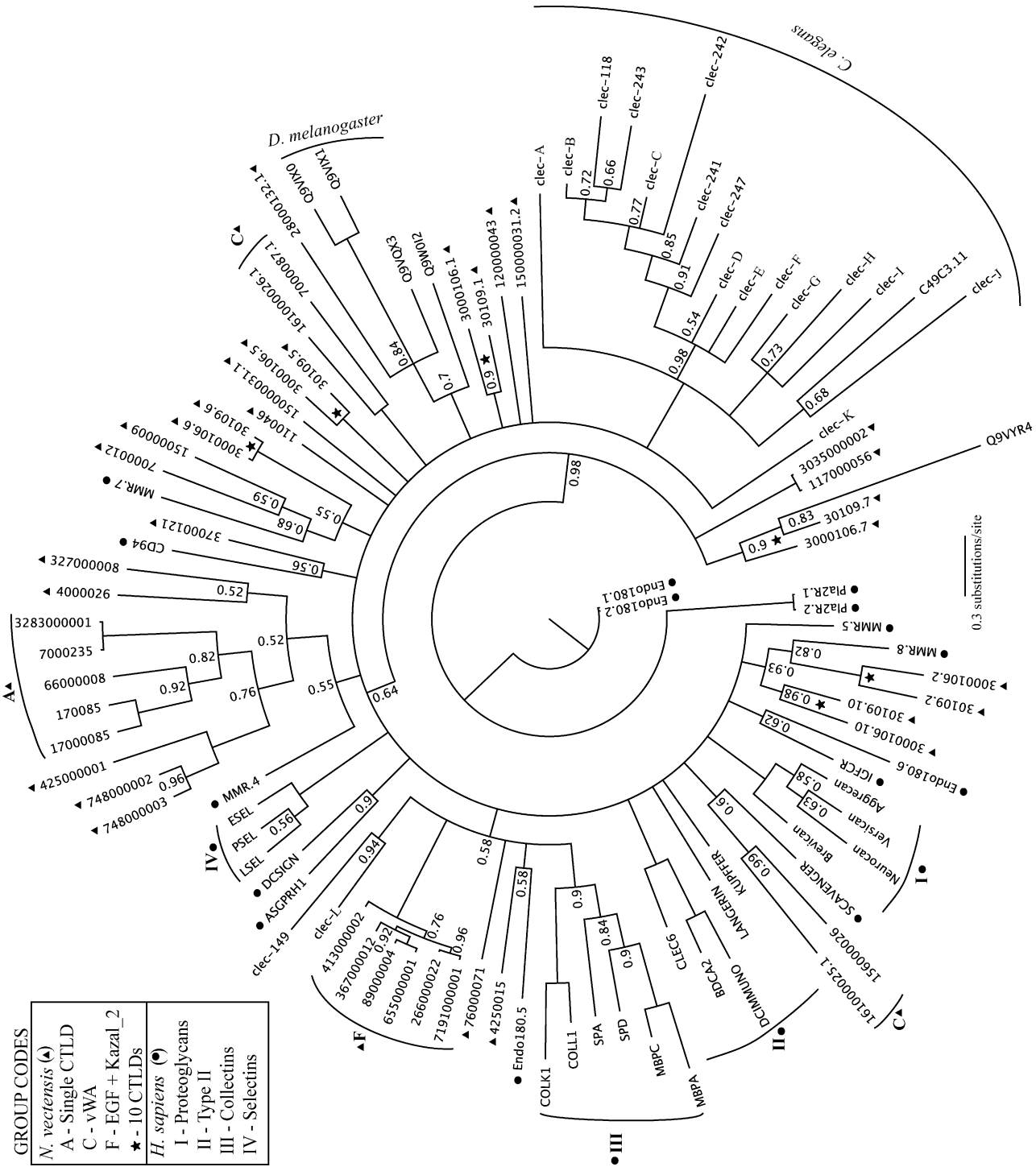

Fig. 5. Bayesian analysis tree of CTLDs containing at least three of the five conserved residues that structurally form the Ca^{2+}/glycan-binding pocket from *Nematostella vectensis*, *Homo sapiens*, *Caenorhabditis elegans*, and *Drosophila melanogaster* (available in color as Supplementary Fig. 3). Numbers at each branch represent posterior probabilities, with unlabeled branches indicating a posterior probability of 1.0. *N. vectensis* sequences were named as in Fig. 3, and group codes are identified in Table 1. Human CTLDs groups were identified by a numerical identifier described by Drickamer and Taylor [16]. (▲) *N. vectensis*; (●) *H. sapiens*. Accession numbers: *H. sapiens*—(I) Aggrecan (P16112), Versican (P13611), Neurocan (O14594), Brevican (Q96GW7); (II) ASGPRH1 (P07306), SCAVENGER (Q8WZA4), LANGERIN (Q9UJ71), KUPFFER (Q8N1N0), DCSIGN (Q9NNX6), BDCA2 (Q8WTT0), CLEC6 (Q6EIG7), DCIMMUNO (Q9UMR7), IGFCR (P06734); (III) MBPA (P39039), MBPC (P11226), SPA (Q8IWL1), SPD (P35247), COLK1 (Q7Z6N1), COLL1 (Q9Y6Z7); (IV) LSEL (P14151), PSEL (P16581), ESEL (P16109); NK receptor CD94 (Q13241); MMR family MMR (P22897), Endo180 (Q9UBG0), Pla2R (Q13018). *C. elegans* accession numbers included in consensus sequences: clec-A (79, 151, 152, 153, 155), clec-B (245, 246), clec-C (238, 239), clec-D (207, 208), clec-E (126, 133), clec-F (93, 94, 95, 96, 97, 98, 99, 103, 121.1, 121.2, 122, 125, 127, 128.1, 128.2, 129, 130, 131, 132, 181), clec-G (100, 108, 232, 233), clec-H (92, 104, 105, 106, 109, 203, 248, 249, 260), clec-I (111, 112, 179, 194, 213, 214, 215), clec-J (141, 147, 158, 159), clec-K (51, 52, 53), clec-L (48, 49, 50).

4. Discussion

The diversity of predicted CTL-like sequences in *N. vectensis* suggests that this gene family fills a variety of functional roles, many of which occur extracellularly and may function in recognition and immunity at the host/microbe interface. Given the current knowledge of the structure–function relationship within the Ca^{2+}/glycan-binding pocket, almost half of the CTLDs in

N. vectensis are expected to have the ability to bind glycan ligands. Of those, general ligand specificity can be predicted for sequences that contain experimentally tested triplet motifs (EPN or QPD). However, there was also a QPN motif present in six *N. vectensis* CTLDs (group F). Functional studies on the *N. vectensis* sequences are needed to confirm glycan specificity for sequences carrying the EPN/QPD motifs and to determine if the QPN motif represents a novel binding pocket with unique glycan specificity.

Full-length sequencing of *N. vectensis* CTLs will also be necessary to confirm the *Ab initio* gene model predictions. For example, one sequence (7191000001), listed in Table 1 as group F, was only shown to contain a Kazal_2 and a CTLD domain. However, the CTLD sequence clustered tightly with other *N. vectensis* sequences that contain Kazal_2 along with several EGF domains (Fig. 4). On the JGI genome browser, closer examination of the gene prediction for 7191000001 revealed that it was located at the end of a short scaffold. Therefore, the sequence prediction was likely incomplete, and we predict that the full-length sequence will contain C-terminal EGF domains.

Most of the sixteen major CTL groups contain one or more subgroups with a signal sequence suggesting that those sequences are likely to be extracellular, either secreted or localized to the plasma membrane. Even though cnidarians arose early in metazoan evolution, their immune systems had already developed many of the elaborate defense mechanisms thought to be found only in higher metazoans [43,44]. Extracellular CTLs serve as recognition receptors in other organisms, and they appear to serve a similar function in cnidarians. For example, lectin/glycan interactions have been shown to play a role in recognition during infection of host cnidarians by symbiotic algae [29–31].

Similar to CTLDs in vertebrates, cnidarian CTLDs were found in combination with a diverse array of other protein domains, and surprisingly, many of these domains were shared between cnidarians and vertebrates. So far, vertebrate CTLs have been found to contain 17 additional protein domains [23], and of those, 10 were also present in CTLs from *N. vectensis*. For comparison, *C. elegans* CTLs share only 5 of the additional domains (all found in *N. vectensis*) and *D. melanogaster* CTLs share 3 (two found in *N. vectensis*) [12,22]. In addition to sharing most of the vertebrate CTL domains, *N. vectensis* had several additional protein domains that were unique to CTL sequences (underlined in Table 1). The overlap in CTL domain composition and CTLD sequence similarity between cnidarians and vertebrates, but not with the model invertebrates, support the conclusions that basal metazoans are more complex than previously thought and the classical invertebrate model organisms have undergone extensive gene loss and divergence. The level of divergence in the CTL gene family is highlighted by the almost ten-fold difference in size—*C. elegans* with 278 CTLDs and *D. melanogaster* with 32 CTLDs.

Identification of CTL-like sequences from the *N. vectensis* genome will be useful in future studies that focus on the functional role of CTLs in basal metazoans. These studies may identify CTL-like sequences with conserved functions in cell–cell adhesion and immunity. For example, the *N. vectensis* CTLs with 10 CTLDs (group J) show a similar CTLD architecture to the vertebrate mannose receptor (MR) family. One principle function of the MR family is phagocytosis of microorganisms, but they also act as cell–cell adhesion and/or recognition receptors (reviewed in [45]). Another interesting group of CTLs contain TSP1 domains (group K), which bind CD36, a host receptor involved in cell–cell adhesion that is partially responsible for infection of erythrocytes by *Plasmodium falciparum* [46]. A recent microarray study comparing a symbiotic and non-symbiotic state of a host cnidarian found a CD36 family member had increased expression in the symbiotic host [24]. The authors discuss that the cnidarian CD36 family member may facilitate the symbiosis, similar to host infection by *P. falciparum*, a member of apicomplexa (sister taxa to the dinoflagellates) [47]. In addition, TSP1 repeats were uncovered in a family of putative immune recognition proteins from the hydroid *Hydractinia symbiolongicarpus* [48].

Cnidarian lectins have already been shown to play a role during initial contact and recognition of intracellular symbionts [28–30]. After contact, these algal symbionts are phagocytosed by the host's gastrodermal cells [49], but it is unknown if lectins also participate in the phagocytic process. In addition to mutualistic partnerships, cnidarian lectins also bind potential pathogens [31], and it will be interesting to explore the function of lectins in cnidarian innate immunity. This question is becoming increasingly important as increasing ocean temperatures, agricultural runoff, acidification, and other anthropogenic disturbances continue to stress coral reef ecosystems and instances of coral disease become more prevalent [50,51]. Unfortunately, almost nothing is known about the immune capabilities of cnidarians.

Acknowledgements

This work was supported by an NSF grant (IOB0542452) to V.M.W. The US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) produced the *N. vectensis* sequence data. We would like to thank Jodi Schwarz for her help during the development of this project, Christy Schnitzler and Olivier Detourney for their insightful comments, and Kristin Motz at Linfield College for her help retrieving sequences. Contribution No. 109 of the Center for Microbial Oceanography: Research and Education at the University of Hawaii.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dci.2009.01.008.

References

[1] Bosch TCG, Khalturin K. Patterning and cell differentiation in *Hydra*: novel genes and the limits to conservation. Can J Zool 2002;80(10):1670–7.
[2] Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 2007;317(5834):86–94.
[3] Kortschak RD, Samuel G, Saint R, Miller DJ. EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. Curr Biol 2003;13(24):2190–5.
[4] Technau U, Rudd S, Maxwell P, Gordon PMK, Saina M, Grasso LC, et al. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet 2005;21(12):633–9.
[5] Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, et al. Unexpected complexity of the Wnt gene family in a sea anemone. Nature 2005;433(7022):156–60.
[6] Miller DJ, Ball EE, Technau U. Cnidarians and ancestral genetic complexity in the animal kingdom. Trends Genet 2005;21(10):536–9.
[7] Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, et al. The innate immune repertoire in Cnidaria—ancestral complexity and stochastic gene loss. Genome Biol 2007;8(4):59.
[8] Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. Nature 1999; 400(6745):667–71.
[9] Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet 2002;32(4):569–77.
[10] Theopold U, Rissler M, Fabbri M, Schmidt O, Natori S. Insect glycobiology: a lectin multigene family in *Drosophila melanogaster*. Biochem Biophys Res Commun 1999;261(3):923–7.
[11] Zelensky AN, Gready JE. C-type lectin-like domains in *Fugu rubripes*. BMC Genome 2004;5(1):51.
[12] Drickamer K, Dodd RB. C-Type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence. Glycobiology 1999;9(12): 1357–69.
[13] Schulenburg H, Hoeppner MP, Weiner Iii J, Bornberg-Bauer E. Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode *Caenorhabditis elegans*. Immunobiology 2008;213(3–4):237–50.

[14] Drickamer K. Ca^{2+}-dependent carbohydrate-recognition domains in animal proteins. Curr Opin Struct Biol 1993;3(3):393–400.
[15] Weis WI, Drickamer K, Hendrickson WA. Structure of a C-type mannose-binding protein complexed with an oligosaccharide. Nature 1992;360(6400):127–34.
[16] Drickamer K, Taylor ME. Biology of animal lectins. Annu Rev Cell Biol 1993;9(1):237–64.
[17] Weis WI, Taylor ME, Drickamer K. The C-type lectin superfamily in the immune system. Immunol Rev 1998;163:19–34.
[18] Cambi A, Koopman M, Figdor CG. How C-type lectins detect pathogens. Cell Microbiol 2005;7(4):481–8.
[19] Lasky LA. Selectins: interpreters of cell-specific carbohydrate information during inflammation. Science 1992;258(5084):964–9.
[20] Hynes RO, Zhao Q. The evolution of cell adhesion. J Cell Biol 2000;150(2):89–96.
[21] Loker ES, Adema CM, Zhang SM, Kepler TB. Invertebrate immune systems—not homogeneous, not simple, not well understood. Immunol Rev 2004;198:10–24.
[22] Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. Glycobiology 2001;11(5):71–9.
[23] Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. FEBS J 2005;272(24):6179–217.
[24] Rodriguez-Lanetty M, Phillips W, Weis V. Transcriptome analysis of a cnidarian—dinoflagellate mutualism reveals complex modulation of host gene expression. BMC Genome 2006;7(1):23.
[25] Sullivan J, Finnerty J. A surprising abundance of human disease genes in a simple basal animal, the starlet sea anemone (*Nematostella vectensis*). Genome 2007;50(7):689–92.
[26] Rodriguez-Lanetty M, Wood-Charlson EM, Hollingsworth L, Krupp D, Weis V. Temporal and spatial infection dynamics indicate recognition events in the early hours of a dinoflagellate/coral symbiosis. Mar Biol 2006;149(4):713–9.
[27] Schwarz JA. Understanding the intracellular niche in cnidarian-*Symbiodinium* symbioses: parasites lead the way. Life Environ 2008;58(2):141–51.
[28] Koike K, Jimbo M, Sakai R, Kaeriyama M, Muramoto K, Ogata T, et al. Octocoral chemical signaling selects and controls dinoflagellate symbionts. Biol Bull 2004;207(2):80–6.
[29] Lin KL, Wang JT, Fang LS. Participation of glycoproteins on zooxanthellal cell walls in the establishment of a symbiotic relationship with the sea anemone, *Aiptasia pulchella*. Zool Stud 2000;39(3):172–8.
[30] Wood-Charlson EM, Hollingsworth LL, Krupp DA, Weis VM. Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. Cell Microbiol 2006;8(12):1985–93.
[31] Kvennefors EC, Leggat W, Hoegh-Guldberg O, Degnan BM, Barnes AC. An ancient and variable mannose-binding lectin from the coral *Acropora millepora* binds both pathogens and symbionts. Dev Comp Immunol 2008;32(12):1582–92.
[32] Sunagawa S, Wilson E, Thaler M, Smith M, Pringle J, Weis V, et al. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its algal endosymbiont; in press.
[33] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. Pfam: clans, web tools and services. Nucl Acids Res 2006;34(suppl_1):D247–51.
[34] de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucl Acids Res 2006;34:W362–5.
[35] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucl Acids Res 2005;33(suppl_2):116–20.
[36] Thompson J, Gibson T, Plewniak F, Jeanmougin F, Higgins D. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 1997;25(24):4876–82.
[37] Drickamer K. Engineering galactose-binding activity into a C-type mannose-binding protein. Nature 1992;360(6400):183–6.
[38] Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 1999;4:95–8.
[39] Ronquist F, Huelsenbeck JP. MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 2003;19(12):1572–4.
[40] Weis WI, Drickamer K. Structural basis of lectin-carbohydrate recognition. Annu Rev Biochem 1996;65:441–73.
[41] Takeuchi T, Sennari R, Sugiura K-I, Tateno H, Hirabayashi J, Kasai K-I. A C-type lectin of *Caenorhabditis elegans*: its sugar-binding property revealed by glycoconjugate microarray analysis. Biochem Biophys Res Commun 2008;377(1):303–6.
[42] Reidling JC, Miller MA, Steele RE. Sweet Tooth, a novel receptor protein-tyrosine kinase with C-type lectin-like extracellular domains. J Biol Chem 2000;275(14):10323–30.
[43] Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, et al. The innate immune repertoire in Cnidaria—ancestral complexity and stochastic gene loss. Genome Biol 2007;8(4):R59.
[44] Bosch TCG. The path less explored: innate immune reactions in cnidarians. In: Gross HJ, editor. Innate immunity of plants, animals, and humans. Berlin: Springer; 2008. p. 27–42.
[45] East L, Isacke CM. The mannose receptor family. Biochim Biophys Acta 2002;1572(2–3):364–86.
[46] Oquendo P, Hundt E, Lawler J, Seed B. CD36 directly mediates cytoadherence of *Plasmodium falciparum* parasitized erythrocytes. Cell 1989;58(1):95–101.
[47] Baldauf SL. The deep roots of eukaryotes. Science 2003;300(5626):1703–6.
[48] Schwarz R, Hodes-Villamar L, Fitzpatrick K, Fain M, Hughes A, Cadavid L. A gene family of putative immune recognition molecules in the hydroid *Hydractinia*. Immunogenetics 2007;59(3):233–46.
[49] Pardy RL, Muscatine L. Recognition of symbiotic algae by *Hydra viridis*. A quantitative study of the uptake of living algae by aposymbiotic *H. viridis*. Biol Bull 1973;145(3):565–79.
[50] Bruno JF, Selig ER, Casey KS, Page CA, Willis BL, Harvell CD, et al. Thermal stress and coral cover as drivers of coral disease outbreaks. PLoS Biol 2007;5(6):e124.
[51] Hoegh-Guldberg O, Mumby PJ, Hooten AJ, Steneck RS, Greenfield P, Gomez E, et al. Coral reefs under rapid climate change and ocean acidification. Science 2007;318(5857):1737–42.