MOLECULAR MECHANISMS OF HOST - SYMBIONT RECOGNITION IN A HIGHLY SPECIFIC SPONGE - ARCHAEAL SYMBIOSIS

by

Elena Zaikova

B.Sc., The University of British Columbia, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Microbiology and Immunology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2014

© Elena Zaikova, 2014

Abstract

Evolution of multicellular eukaryotes is intimately associated with microbial interactions resulting in diversification and niche expansion. This long history of co-evolution is evident in metabolic interdependence, and reliance of animal (i.e. metazoan) ecosystems on their microbiota for healthy development and function. Specific recognition between interacting partners is essential for establishing and successfully maintaining interspecies associations, and involves host immunity and symbiont-encoded factors. Sponges represent the most deeply branching animal phylum with the potential to shed new light on the evolution of innate immunity and host-microbe interactions within the metazoa. Marine sponges harbour diverse microbial communities that contribute to higher order ecosystem functions including primary production and nutrient cycling. However, molecular mechanisms mediating symbiont recognition and host immune signalling in sponge symbioses are unknown. This knowledge gap stems from the fact that most sponge-associated microbes remain uncultivated and no sponge host/symbiont culture systems exist. In this thesis, I used cultivation-independent approaches including environmental genomics, transcriptomics, and proteomics in combination with homology modeling and community composition profiling to identify molecular determinants of sponge symbiosis in the sponge Dragmacidon mexicanum. Community composition profiling indicated that D. mexicanum is a high microbial abundance sponge harbouring a specific microbial community dominated by the Thaumarchaeaote Cenarchaeum symbiosum. Comparative genomics and gene expression profiling identified potential symbiont-encoded proteins including serine protease inhibitors (serpins) with the potential to mediate host-microbe interactions that were not found in closely related free-living Thaumarchaeaota, consistent with

C. symbiosum's adaptation to a symbiotic lifestyle. Biochemical assays were subsequently used to characterize serpin activity and infer function. Immunity determinants previously unreported in sponges were identified, enabling near-complete reconstruction of innate immune signalling pathways and partial adaptive immunity pathways. Thus, this work expanded the known complexity of sponge immune signalling and suggests a more ancient origin of certain pathways than previously recognized. The composition of sponge innate immunity may reflect the complex nature of sponge-associated microbiota, which likely acquired adaptive features to thrive in the host milieu. Taken together, this thesis provides novel insights into the evolution of host-microbial recognition, archaeal adaptations to a symbiotic lifestyle, and molecular interactions between archaea and eukaryotic cells.

Preface

The work presented in this dissertation is my original and unpublished work. As thesis research supervisor, Dr. Steven J. Hallam was involved in the identification, design and execution of the research program.

Collaborators at the Monterey Bay Aquarium and the Vancouver Aquarium collected sponge specimens used in this thesis. In Chapter 2, Gareth D. Mercer constructed archaeal small subunit rRNA libraries and obtained sponge small subunit rRNA sequences for 8 sponges. Undergraduate students under my training and supervision using protocols established by me conducted a number of DNA amplification experiments. I performed all other work and all analyses presented.

In Chapter 3, Dr. Marcus Taupp cloned serine protease inhibitor genes, and Dr. Jinshu Yang purified the protein used for antibody production. Dr. Yang also extracted sponge protein for proteomics experiments. Heather Brewer and Angela Norbeck conducted liquid chromotography-tandem mass spectrometry and peptide identification, respectively at the Pacific Northwest National Labratories. I performed cell stimulation experiments with the help of Nita Shah and Melissa Elliot from the Fernandez and Hancock labs at UBC, respectively. I performed the peptide pull-down experiment with Dr. John Cheng from the Jean lab at UBC, who performed the peptide searches and protein identification. The metallo-β-lactamase similarity network was done with the help of Florian Baier in the Tokuriki lab at UBC. I conducted all the work and data analysis in Chapter 4.

Table of Contents

Abstra	act	ii
Prefac	ce	iv
Table	of Contents	V
List of	f Tables	X
List of	f Figures	xi
List of	f Abbreviations	. xiii
Ackno	owledgements	XV
Dedica	ation	. xvi
Chapt	er 1: Introduction	1
1.1	Co-evolution of animals and their microbiota	1
1.2	Sponge-microbial associations	2
1.	2.1 Sponge biology	2
1.	2.2 Sponge microbiota	4
1.3	Host-symbiont recognition	6
1.4	Sponge innate immunity	8
1.5	A specific symbiosis: Dragmacidon mexicanum and Cenarchaeum symbiosum	10
1.6	Research question and thesis objectives	12
Chapt	er 2: Microbial community structure in <i>Dragmacidon mexicanum</i> and other spong	ge
species	s from northeast Pacific coastal waters	14
2.1	Introduction	14
2.2	Materials and methods	16
		v

2.2.1	Sample collection and processing	16
2.2.2	DNA isolation and purification	17
2.2.3	Quantification of archaeal and bacterial taxonomic marker genes	18
2.2.4	Preparation of PCR amplicons for pyrosequencing	19
2.2.5	Three-domain variable region sequence data analysis	19
2.2.6	Archaeal SSU rRNA gene amplification, library production and screening	21
2.2.7	Eukaryotic SSU rRNA gene PCR amplification	21
2.2.8	Phylogenetic analysis of SSU rRNA sequences	22
2.3 Re	sults	24
2.3.1	Composition of sponge sample set	24
2.3.2	Quantification of archaea and bacteria in sponges	26
2.3.3	Community structure relationships between sponges	27
2.3.4	Taxonomic composition of sponge microbiota	31
2.3.5	Archaeal phylogeny in SB sponges	34
2.3.6	Indicator species analysis	35
2.3.7	Dragmacidon mexicanum core microbiota	36
2.4 Dis	scussion	40
2.4.1	Core and indicator sponge microbiota	40
2.4.2	Taxonomic composition of sponge microbiota	41
2.4.3	Intraspecific variation between Dragmacidon mexicanum microbiota	43
Chapter 3	Genomic and functional characterization of C. symbiosum-encoded genes.	46
3.1 Int	roduction	46
3.2 Ma	terials and methods	48
		vi

3.2.1	Comparative genomics of C. symbiosum and other Thaumarchaeota	. 48
3.2.2	Protein extraction and peptide mapping	. 50
3.2.3	Sequence homology and structure prediction	. 51
3.2.4	Biological activity of C. symbiosum serpins	. 52
3.2	4.1 Cloning, expression and purification of <i>C. symbiosum</i> serpins	. 52
3.2	4.2 Protease inhibition assay	. 53
3.2	4.3 Serpin-sponge lysate pull-down	. 54
3.2	4.4 Serpin protein and C-terminal peptide activity in NFκB reporter cell line	. 55
3.2	4.5 C-terminal peptide activity in human peripheral blood mononuclear cells	. 57
3.2	4.6 C-terminal peptide pull-down in sponge and mammalian cells	. 58
3.3 Re	sults	. 59
3.3.1	Comparative genomics between C. symbiosum and free-living Thaumarchaeota	. 59
3.3.2	C. symbiosum proteins implicated in host-microbe interactions	. 63
3.3.3	Biological activity of C. symbiosum serpins	. 66
3.4 Dis	scussion	. 73
3.4.1	Putative C. symbiosum symbiosis factors	. 73
3.4.2	Biological activity of C. symbiosum serpins	. 76
Chapter 4	: Microbial recognition and host defense systems in marine sponges	80
4.1 Int	roduction	. 80
4.2 Ma	aterials and methods	. 83
4.2.1	RNA isolation and purification	. 83
4.2.2	Sponge housekeeping genes PCR protocol	. 85
4.2.3	cDNA library production, sequencing and assembly	. 86
		vii

4.2.4 Transcriptome annotation
4.2.5 Identification of putative pathways and interactions between expressed genes 88
4.2.6 Identification of protein-coding genes from existing Oscarella carmela EST data. 89
4.3 Results
4.3.1 Composition of sponge transcriptome datasets
4.3.1.1 Taxonomic composition of transcriptomes
4.3.1.2 Functional annotation of transcriptomes
4.3.2 Diversity and distribution of pathways and predicted ORFs in sponge transcriptomes
93
4.3.3 Innate immunity genes and pathways in <i>D. mexicanum</i> and <i>T. californiana</i>
4.3.3.1 Toll-like receptor signalling
4.3.3.2 Nod-like receptor signalling
4.3.3.3 Phagocytosis and autophagy
4.3.3.4 Lectins, complement and coagulation 104
4.3.3.5 Viral recognition mechanisms
4.3.3.6 Apoptosis, transendothelial migration and adaptive immunity pathways 108
4.4 Discussion
4.4.1 Distribution of <i>D. mexicanum</i> and <i>T. californiana</i> transcripts
4.4.2 Microbial recognition and response mechanisms in <i>D. mexicanum</i> and <i>T.</i>
californiana112
Chapter 5: Conclusion118
5.1 <i>Dragmacidon mexicanum</i> hosts a specific microbial community

5.2	Cenarchaeum symbiosum acquired genes not found in other thaumarchaea to interac	t
with	the host	. 120
5.3	Sponges have the potential to recognize a variety of microbes	. 122
5.4	Emerging themes	. 123
5.5	Future directions	. 125
5.6	Closing remarks	. 127
Bibliog	graphy	128
Appen	dices	146
Appe	endix A	. 146
Appe	endix B	. 148
B.	1 Comparison of "A" and "B" type genomic sequences	. 148
В.	2 Quantification of <i>C. symbiosum</i> -encoded serpin genes by qPCR	. 149
В.	3 Identities of proteins pulled down by a synthetic peptide based on <i>C. symbiosum</i>	
CH	ENSYa_1605 serpin C-terminus	. 150
Appe	endix C	. 152
C.	1 Unique KEGG orthologs in <i>D. mexicanum</i> and <i>T. caiifornaian</i>	. 152
C	2 Unique protein families expressed by <i>D. mexicanum</i> and <i>T. californiana</i>	. 153
C	3 Proteases expressed at the mRNA level by <i>D. mexicanum</i>	. 157
(B	LASTp, E-value <1E-6, bsr >0.4)	. 157

List of Tables

Table 2.1 Sponge sample collection, taxonomy summary and quantification of bacteria and	
archaea in sponges.	23
Table 2.2 Amplicon pyrosequencing data summary and community diversity.	29
Table 2.3 Statistical comparison of gamma-proteobacteria composition in sponges.	39
Table 3.1 Thaumarchaeal genomes used in comparative analysis	49
Table 3.2 C. symbiosum serpin C-terminal peptide sequences.	56
Table 3.3 Proteases encoded by C. symbiosum as per original and updated annotation	64
Table 4.1 Summary of sponge sequence data	82
Table 4.2 Summary of sponge transcriptome sequencing	91
Table 4.3 Taxonomic distribution of sponge transcripts.	91
Table 4.4 Distribution of sponge transcriptome BLASTp hits (Evalue < 1E-6, Bitscore > 50)	and
Pfam matches (Evalue <0.1)	92
Table 4.5 D. mexicanum genes with roles in coagulation.	105

List of Figures

Figure 1.1 Sponge anatomy schematic.	3
Figure 2.1 Sponges in sample set	17
Figure 2.2 Distance tree of sponge SSU rRNA gene sequences.	.25
Figure 2.3 Observed species richness rarefaction curve of mean OTU diversity	. 27
Figure 2.4 Relationships between sponge-associated communities.	. 28
Figure 2.5 Core microbial communities of sponge groups as designated by HCA	31
Figure 2.6 Relative abundance and distribution of archaeal, bacterial and eukaryotic taxa	33
Figure 2.7 Distance tree of archaeal SSU rRNA gene sequences recovered from SB sponges	35
Figure 2.8 Indicator and core community members.	38
Figure 2.9 Richness and abundance of OTUs detected in <i>D. mexicanum</i> sponges	. 39
Figure 3.1 Comparative analysis between thaumarchaeal genomes	61
Figure 3.2 Position of proteins of interest identified in this study on the C. symbiosum genome	.62
Figure 3.3 Genomic arrangement of <i>C. symbiosum</i> serpins.	63
Figure 3.4 Putative cell surface modification operon encoded by C. symbiosum.	64
Figure 3.5 Protein sequence similarity network between C. symbiosum and reference metallo-f	3-
lactamases.	65
Figure 3.6 Multiple sequence alignment of predicted C. symbiosum serpins and reference	
homologs	. 67
Figure 3.7 Serpin inhibiton of protease activity assays.	68
Figure 3.8 Effect on TLR signalling in NFkB reporter cells stimulated by C. symbiosum serpin	1
protein and C-terminal peptides	.70 xi

rigure 5.7 Cytokine and elemokine production in ribiwes stimulated with C-terminal peptides.
Figure 3.10 Proteins pulled down by a synthetic CENSYa_1605 peptide in human cells lines 72
Figure 4.1 KEGG pathway composition and distribution in sponge transcriptomes
Figure 4.2 Comparison of sponge transcriptomes based on pathway, gene and Pfam distribution.
Figure 4.3 Shared and species-specific expressed genes and protein families
Figure 4.4 Toll-like receptor signalling in sponges 100
Figure 4.5 NOD-like receptor signalling in sponges
Figure 4.6 Autophagy-associated genes in <i>D.mexicanum</i> and <i>T. californiana</i>
Figure 4.7 Complement signalling genes in <i>D.mexicanum</i> and <i>T. californiana</i>
Figure 4.8 RIG-I-like receptor signalling in sponges 107
Figure 4.9 Apoptosis-associated genes expressed in <i>D. mexicanum</i> and <i>T. californiana</i>
Figure 4.10 Evolutionary conservation of key innate immune signalling molecules
Figure 5.1 Emerging patterns of sponge-microbial interactions in <i>D. mexicanum</i>

Figure 3.9 Cytokine and chemokine production in PBMCs stimulated with C-terminal peptides.

List of Abbreviations

 $A2M - \alpha$ -2-macroglobulin BC – Sponges from British Columbia, Canada BLAST - Basic local alignment search tool Caspase - Cysteine-dependent aspartate-directed proteases CMFASW - Calcium- and magnesium-free artificial seawater ECM – Extracellular matrix EGFR – Epidermal growth factor receptor ELISA - Enzyme-linked immunosorbent assay EPS – Exopolysaccharide HCA – Hierarchical cluster analysis HMA – High microbial abundance IFN – Interferon Ig – Immunoglobulin IL – Interleukin ISA – Indicator species analysis LBP – Lipopolysaccharide-binding protein LMA – Low microbial abundance LPS - Lipopolysaccharide LRR – Leucine-rich repeats KEGG - Kyoto Encyclopedia of Genes and Genomes $m-\beta L - metallo-\beta-lactamase$ MAMP - Microbe associated molecular pattern MAPK - Mitogen-activated protein kinase MBL - Mannose binding lectin MCP-1 - Monocyte chemotactic protein-1 MEGAN - Metagenome Analyzer MRPP – Multi-response permutation procedures MyD88 – Myeloid differentiation primary response gene 88 NFκB – Nuclear factor κ-light-chain-enhancer of activated B cells NOD - Nucleotide-binding oligomerization domain receptor NLR - NOD-like receptor NMS – Non-metric multidimensional scaling NSAF - Normalized spectral abundance factor ORF – Open reading frame OTU – Operational taxonomic unit PBMC – Peripheral blood mononuclear cell Pfam - database of Protein Families PRR – Pattern recognition receptor QIIME – Quantitative insights into microbial ecology RFLP – Restriction fragment length polymorphism RIG-I – Retinoic acid-inducible gene 1

RLR – RIG-I-like receptor

SB – Sponges from Santa Barbara, USA SDS-PAGE – Sodium dodecyl sulfate polyacrylamide gel electrophoresis

Serpin – Serine protease inhibitor

SRCR – Scavenger receptor cysteine-rich

SSU rRNA – Small subunit ribosomal RNA

TEP – Thioester-containing protein

TIR - Toll/interleukin-1 receptor

TLR – Toll-like receptor

 $TNF\alpha$ – Tumor necrosis factor alpha

TPR – Tetratricopeptide repeat

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Steven Hallam, for letting me take on this fascinating and multifaceted project. Dr. Hallam's tutelage has afforded me the opportunity to develop scientific skills in diverse disciplines and grow as an independent researcher, for which I am very grateful. I am also very appreciative to have gained invaluable leadership experience and the freedom to pursue skills that contributed to my academic development.

I am deeply grateful to my thesis committee members, Dr. François Jean, Dr. Michael Gold, and Dr. Robert Hancock for their guidance and encouragement. Their advice, feedback and sincere support were critical in shaping this thesis and helped me become a better scientist.

I offer my heartfelt thanks to members of the Hallam lab, past and present, and colleagues at UBC for their friendship, support and stimulating scientific exchange. I owe particular thanks to Dr. Nicole Sukdeo, who has been an immense source of personal, academic and scientific support and advice. I would also like to thank Darlene Birkenhead and Dr. Michael Murphy for ensuring that the graduate student experience in the Department of Microbiology and Immunology is smooth and positive.

I would like to gratefully acknowledge the Canadian Institute for Health Research and the University of British Columbia for funding to support my doctoral work.

I extend my deepest gratitude to my wonderful family and family-in-law for believing in me and being so understanding and accommodating, and a special thanks to my mother, Liubov, for being a pillar of support. I am grateful to my amazing husband, Scott Pennykid, for encouraging me to follow my dreams, helping me find the humour in every situation, and bringing joy and balance to my life. For my grandmothers, Lidia and Elena, who inspired me to pursue this degree

and

For Scott, whose love and support made it possible

Chapter 1: Introduction

1.1 Co-evolution of animals and their microbiota

Many animals are colonized by symbiotic microbes, which directly impact host fitness and ecology. Given that prokaryotic evolution predates the emergence of eukaryotes, the evolution of multicellular organisms has always involved interactions with microbes, resulting in diversification and niche expansion (1, 2). Furthermore, the reciprocal dependence between host and microbe implies that co-evolution is a dynamic and iterative process, whereby the host and symbiont exert selective pressure, resulting in genomic changes in both partners. This coevolutionary history manifests in strong physiological, developmental and genomic dependencies between eukaryotic hosts and their symbiont communities (1-5). Note that in this thesis, no distinction is made between mutualistic and commensal associations, thus the more general term "symbiosis" is used. In fact, many animal (i.e. metazoan) hosts require specific symbiotic communities for healthy development of organ and immune systems (4, 6-12). The extent to which animal hosts recognize and selectively retain symbiotic microbes among hundreds or thousands of different microbial species encountered over space and time is an important question. The implied necessity for recognition between microbial and animals partners led to the hypothesis that adaptive immunity evolved to manage complex and diverse resident microbial communities (13, 14). However, adaptive immunity in invertebrates, which represent more than 95% of animal species, has not been described yet many invertebrate taxa, such as sponges and corals harbor complex and host-specific microbial communities (15-18). Therefore, understanding host-microbe interactions in these systems will provide new insights into the evolution of host-microbe interactions in complex symbioses.

1.2 Sponge-microbial associations

1.2.1 Sponge biology

Sponges (phylum Porifera) are sessile benthic filter feeding animals that represent the deepest-branching extant metazoan phylum. Collectively, this diverse phylum contains at least 8,500 recognized species from diverse saltwater and freshwater habitats, ranging from high latitude to temperate and tropical regions in shallow tidal zones to the deep ocean. Almost 85%, of all sponge species belong to the class Demospongiae and inhabit marine environments (19). Marine sponges are important members of benthic communities where they contribute to biodiversity, provide habitat and protection from predation, and are a nutrient source for many species including endangered fish (20-24). As highly efficient suspension feeders that can filter 24L per gram sponge per day with a 96% filtration efficiency, marine sponges are an important link between benthic and pelagic ecosystems, with discernable impact on marine nutrient cycling (22, 25-35).

Sponges have a simple anatomy, optimized for water flow and nutrient uptake, and relatively few cell types with the capacity to re-differentiate into other cell types (36). Sponge differentiated epithelia, the choanoderm, endopinacoderm and exopinacoderm, are the most ancient metazoan tissues (37). Sponge epithelial cells, particularly the choanocytes have a high turnover rate, exceeding that of mammalian intestinal epithelial cells (38). Like epithelia in more recently evolved animals, sponge epithelial cells function to detect pathogens, control secretion and absorption, provide support and create barriers for compartmentalization and osmoregulation within the extracellular matrix, termed the mesohyl (37, 39). In addition to the non-epithelial sponge cells, including the motile phagocytic amoebocytes, collagen-producing collenocytes and lophocytes, archaeocytes and spicule-producing sclerocytes, the mesohyl contains microbial cells

(Figure 1.1). Food uptake in sponges occurs by phagocytosis, primarily by the choanocytes, as well as pinacocytes, in choanocyte chambers. Following capture, microbes are transferred from the small (\sim 1µm) choanocyte cells to the larger amoebocytes for digestion, with the potential for subsequent transfer to other cell types (40-42). Whether the feeding process is selective, and whether it differs from phagocytosis associated with innate immunity, is currently unknown.



Figure 1.1 Sponge anatomy schematic.

Due to their filtering activities, sponges constantly engage in nutritional, pathogenic and beneficial interactions with billions of microbes. Consequently, the sponge immune system must be able to appropriately distinguish different microbes (43). Despite their simple anatomy and physiology, sponges have persisted for ~600 million years, and microbial symbioses are likely a key contributing factor to this evolutionary success (21). Given their basal phylogenetic position, the biology of sponges can provide insight into the evolution of cell adhesion, signalling and innate immunity systems (44, 45), presenting an opportunity to explore conserved patterns in microbial recognition and symbiont selection by animal hosts.

1.2.2 Sponge microbiota

Interactions with microbes and microbial cues are essential for all life stages of sponges, from larval settlement and development to adult sponge physiology, fitness and evolution (5, 46, 47). The symbiotic relationships established by the sponge with specific microbial groups may benefit the host sponge by increasing its fitness (48, 49). Most sponge-associated microbes have not been cultivated, thus genomic, transcriptomic and proteomic sequencing methods are necessary to infer potential modes of interaction between sponges and members of their microbiota. Sequencing efforts indicate that sponge-associated microbiota may be involved in the production of secondary metabolites, transformation of waste products, biosynthesis of vitamins, cofactors and amino acids, as well as roles in carbon and nitrogen metabolism (50-55). However, signalling and pattern recognition mediating sponge–microbial symbioses remain poorly understood.

Two types of sponges that can be discerned based on the size of associated microbial community size. The first, termed high microbial abundance (HMA) sponges, host microbial communities at concentrations 100-10,000 times greater than those in surrounding seawater, whereas the second type, the low microbial abundance (LMA) sponges, host communities at concentrations similar to those found in the surrounding seawater (35, 56). These two types of sponges differ not only in microbial abundance but community structure as well (57). HMA sponges tend to have slower pumping rates, higher respiration, and larger organic carbon requirements than LMA sponges (35). Furthermore, sponge extracellular matrix density and water vascular system morphologies differ between LMA and HMA sponges. However, it is still unclear whether this is due to differences in microbial abundance (58).

Sponge microbial communities are large, diverse, and complex, consisting of persistent and temporary associations with archaeal, bacterial and single-cell eukaryotes (21, 59, 60). While a few sponge-associated microbes are intracellular, most sponge symbionts colonize the mesohyl (21, 61, 62). Several bacterial and archaeal symbionts form sponge-specific clades that are distinct from their free-living relatives (21, 63). Indeed, a candidate monophyletic phylum of uncultured bacteria found almost exclusively in sponges called "Poribacteria" has been proposed (64). Overall, sponge microbial communities differ between species and manifest community structures that are markedly different from those found in surrounding seawater (16, 18, 21, 50, 63). These findings suggest not only microbial sponge-specific associations, but species-specific relationships as well. Given that co-occurring sponge species harbor distinct microbial communities, it is likely that sponge pattern recognition systems exist (65).

The origin of sponge-specific microbial lineages is an interesting and a foundational research question. A typical monophyletic sponge-specific clade may contain numerous symbiotic taxa found in multiple different sponge hosts in widely distributed geographical areas (21). This pattern of association may either arise from ancient associations resulting from very few colonization events, or may be due to colonization of every generation by these bacteria and archaea present in the surrounding seawater at very low abundances (16). This suggests that at least a portion of the symbiont pool is horizontally or environmentally, acquired from free-living microbial populations. Indeed, the low abundance of Poribacteria detected in the water column implies that sponge hosts are able to recognize potential symbionts and provide favorable conditions for growth (49). Alternatively, the source of these sponge-specific clades in seawater may be sponges themselves. Adult sponges may release symbionts when spawning to promote colonization of new sponges, consistent with vertical transmission from parent to offspring (16,

66). While some symbionts are exclusively vertically-transmitted, vertical transmission with some horizontal acquisition is a more common mode of parent-to-offspring symbiont transfer (67). With either mode of transmission, symbiont recognition is likely mediated through cell surface interactions necessary for successful colonization and appropriate spatial distribution to take place (67).

1.3 Host-symbiont recognition

Symbiotic and pathogenic microbes use symbiosis and virulence factors, respectively, to modify host systems to gain entry into the host. These factors can help the colonizing microbe attach to host cells or extracellular matrix, enter into host intra or intercellular compartments, acquire host micronutrients, and evade the host immune system and antimicrobial responses (68-71). The innate immune response depends on the recognition of conserved microbial structural features, termed microbial associated molecular patterns (MAMPs), by germline-encoded pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs), Nod-like receptors (NLRs), mannose receptors and C-type lectins (72-74). Bacterial and fungal MAMPs are well characterized and include essential, often highly expressed components such as flagellin, peptidoglycan, lipopolysaccharides, β -glucan, and lipoproteins (73, 75, 76). Cell surface characteristics are known for just two archaeal phyla, Euryarchaeata and Crenarchaeota, whereas Thaumarchaeal MAMPs are enigmatic (77).

Bacterial constituents of the mammalian gut microbiota utilize extracellular features including pili, fimbriae, cell envelope as well as secreted proteins to adhere to host mucosa and cell surfaces and protect themselves from host immune responses (78, 79). The cell envelope contains exopolysaccharides (EPS) that are important for recognition and symbiosis maintenance

6

(78, 80). Host lectins bind sugars on symbiont cell surfaces for recognition (67). Further, modifications to sugar structure, conformation or composition leads to higher specificity of interaction with the host immune system (81). Therefore, like pathogen-derived capsules and EPS that are important in adherence and phagocytosis inhibition, symbiont EPS play roles in symbiotic communication and modulation of the host (78, 82). In addition to cell surface structures, serine protease inhibitors (serpins) help mediate host-microbial interactions and possibly modulate host immunity (78, 79). For example, multiple *Bifidobacterium* species encode serpins and several of these serpins have been implicated in symbiotic cell communication with potential to attenuate inflammation in the host milieu (83-85). Serpins also play important roles in host-microbial interactions in insects (86-90). Given these observations, it is plausible that serpin-mediated immunomodulation encoded in the genomes of both host or symbiont cells may be conserved across metazoan ecosystems.

Extensive symbiotic crosstalk must take place between host symbiotic microbes, as unwarranted immune responses to antigens from resident microbiota result in disease (reviewed in (81)). Furthermore, the presence of symbiotic microbiota prevents colonization of the host by pathogens, possibly through antagonistic interactions (91). Thus interactions between members of the microbiota with each other and the host help maintain a healthy community composition (78). Shifts in this composition can have detrimental affects on the host (92). Symbiotic microbes aid in the proper development of host immune system (93). Thus innate immunity, more specifically PRR signalling, is an important mechanism of host-symbiont communication (94-96). Other than interactions with immune signalling, symbiotic microbes affect host homeostasis and health by producing or extracting nutrients and metabolites, such as the anti-inflammatory bile acids in vertebrates (97).

Symbiosis evolution is associated with genomic changes and adaptations of both the host and microbial partner (78). Thus, genomic sequences of symbiotic microbes are essential for understanding the genetic adaptation of symbionts to the host environment (50, 82). Recent genomic characterization of the sponge-specific Poribacteria identified putative enzymes that may be involved in the degradation of proteoglycans, which are important components of the sponge mesohyl (98). Archaea also form important symbioses with metazoans, yet how they colonize, interact with, and succeed in the host is unknown (99-101). Interestingly, no archaeal pathogens have been identified (102-104). The apparent lack of archaeal pathogens has important implications for host-microbe interactions and indicates differences in virulence evolution between the prokaryotic domains (102). This could be due to differences in cell envelope organization between bacteria and archaea and the extent to which component parts are recognized by animal hosts. Although multiple functions of pili that include adhesion, and Slayer N-glycosylation in Euryarchaeaota have been described, the composition and regulation of archaeal surface structures is poorly understood (77, 105).

1.4 Sponge innate immunity

To be able to distinguish and mount appropriate responses to symbiotic and pathogenic microbes, sponges likely utilize a variety of innate immunity pathways. Indeed, functionally distinct interactions between the demosponge *Suberites domuncula* and potentially pathogenic and symbiotic bacteria have been observed (106). Several lines of molecular evidence from multiple sponge species have been used to reconstruct component parts of innate immune signalling pathways. These include: (i) biochemical studies and expressed sequence tag libraries from *S. domucula* (107-109), (ii) draft genome and transcriptome data for *Amphimedon*

8

queenslandica (45, 110-112), (iii) expressed sequence tag libraries from *Oscarella carmela* (44), and most recently (iv) Illumnia transcriptomes from *Crella elegans*, *Petrosia ficiformis* and seven other sponge species (113-115). Targeted molecular characterization and sequencing of specific genes involved in allorecognition and histocompatibility are also available for the demosponge *Geodia cydonium* (116-120). Despite the increased availability of sponge transcriptome sequences, profound differences in annotation methods and data interpretation render direct primary literature comparisons difficult. Therefore, similarities and differences between the innate immune signalling pathways used by different sponges species remains to be systematically described.

Studies using *S. domuncula* indicated that this sponge recognizes Gram-negative and Gram-positive bacteria as well as fungal pathogens via cell surface receptors (107-109). Molecular determinants implicated in innate immunity identified in *S. domuncula* include TLRs and components of the TLR signalling pathway (108, 109). TLRs are type I membrane glycoproteins with a conserved cytoplasmic Toll/interleukin-1 (TIR) signalling domain and diverse external antigen-recognizing domains containing leucin-rich repeats (LRRs) capable of detecting a vast array of MAMPs, and therefore microbial types (72, 73, 121-123). The *S. domuncula* TLRs lack LRRs and are unusually short (108, 109). Additionally, the *A. queenslandica* genome encodes (i) two putative receptors with an intracellular TIR domain, (ii) Toll/Interleukin1 receptor-like immunoglobulins (Igs), (iii) 135 homologs of the intracellular Nod-like receptors (NLRs), as well as (iv) an NFkB homologue that is highly similar to the human protein, suggesting that NF-kB evolved prior to divergence of sponges and other metazoa (45, 112, 124). In addition to NLR, TLR and LPS-binding proteins, sponges also contain proteins with scavenger-receptor cysteine-rich (SRCR) domains that may be involved in innate immune

responses (115, 120).

Exposure of *S. domuncula* to components of Gram-negative bacteria cell walls, such as lipopolysaccharides (LPS) and lipoproteins, resulted in the increased expression of bacteriotoxic macrophage-expressed protein and the adaptor molecule MyD88 (108), or induced apoptosis of affected sponge cells by inducing a TLR-dependent cysteine-dependent aspatryl-specific protease (caspase) (109), respectively. The *S. domuncula* caspase is related to human caspase 7 (109), an effector caspase involved in mediating cellular changes associated with apoptosis (125, 126). Similarly, *A. queenslandica* encodes 3 putative caspases (45). Allograft and autograft fusion experiments in the sponge *Geodia cydonium* showed that apoptosis is induced during allograft rejection (107, 118). Thus, apoptotic cell death may be an important sponge response to pathogenic bacteria or viral infection. Besides pathogen recognition and resistance, mammalian TLRs, and NLRs are important for host tolerance of symbiotic microbiota (73, 95, 96, 127, 128). The conservation of molecules involved in microbial recognition across metazoa, together with the presence of key components of TLR and NLR signalling in sponges, suggest that signalling pathways activated by these receptors play a key role in the recognition of microbes by sponges.

1.5 A specific symbiosis: Dragmacidon mexicanum and Cenarchaeum symbiosum

The highly specific symbiosis between the demosponge *Dragmacidon mexicanum*, formerly known as *Axinella mexicana*, and the thaumarchaeaote Candidatus *Cenarchaeum symbiosum* was the first description of a sponge-archaeal association (129). *C. symbiosum* is a non-motile symbiont that occurs extracellularly in the sponge mesohyl, occupying spaces between host cells, where it represents up to 65% of the prokayotic community (129). In the original characterization of this symbiosis, the symbiont was found in over 40 *D. mexicanum*

specimens from 4 locations, and when maintained in flowing aquaria, the sponges did not expel the symbionts (130). Moreover, *C. symbiosum* cells were viable and divided within sponges kept in aquaria, underscoring the persistence and constancy of the symbiosis (129, 130).

Following the discovery of *C. symbiosum*, other closely related (about 96% sequence similarity across the SSU rRNA gene) thaumarchaeal species have been identified in sponge species related to *D. mexicanum* from the Mediterranean, Korean and Australian coasts (131-133). These thaumarchaeota and *C. symbiosum* form a monophyletic sponge-specific lineage based on their SSU rRNA gene sequences (21). Furthermore, a similar species-specific pattern of association to that of *C. symbiosum* with *D. mexicanum* is observed in these sponge-archaeal symbioses (131). This suggests that the specificity and nature of the symbiotic relationship is not unique to *D. mexicanum* but is shared within the sponge family despite its polyphyly (131, 134).

Although *C. symbiosum* is the only archaeal member of *D. mexicanum* microbial community, two different *C. symbiosum* populations, "A" and "B" type, have been described (51, 129, 135). Despite a difference in % GC composition, the symbiont "A" and "B" types are highly similar, >99.2% identity across the small and large subunit rRNA genes, and 80-90% similar at the nucleotide level with >90% protein sequence identity (51, 135). The two *C. symbiosum* types largely encode for the same proteins, in the same order and orientation and therefore it is not clear whether there is functional divergence between them. The "A" type is the dominant form of the symbiont, and its abundance in *D. mexicanum* allowed for the sequencing of its complete genome (51). The *C. symbiosum* "A" genome provides important clues to potential trophic exchanges and is an invaluable resource for investigating the putative molecular mechanisms mediating the *D. mexicanum-C. symbiosum* symbiosis. Based on the presence of genes encoding ammonia monooxygenase, urease and a urea transporter, *C. symbiosum* may

contribute to elimination of nitrogenous wastes in the sponge host (51, 52). Since *C. symbiosum* is considered non-motile and is found in the mesohyl intermingling with microbes that are consumed by the sponge as food, interactions with host cells and proteins are hypothesized to play important roles in establishing and maintaining this stable symbiosis.

The importance of archaea as symbionts is becoming increasingly evident, as sequencing, detection and visualization techniques improve. Methanogenic euryarchaea form associations with ruminants and termites, and are part of the human microbiota (101, 136-138). In addition to sponges, Thaumarchaea have also been reported in associations with other metazoans, including humans and ascidians (139, 140). However, the recognition mechanisms underlying archaeal-metazoan associations are unknown. Thus, investigating the *D. mexicanum-C. symbiosum* association is interesting from an evolutionary standpoint as novel mechanisms of archaeal recognition by multicellular hosts may be uncovered.

1.6 Research question and thesis objectives

The overall objective of this thesis was to detect and describe molecular mechanisms mediating symbiosis between D. mexicanum and C. symbiosum. Specifically, I hypothesized that the D. mexicanum – C. symbiosum symbiosis depends on modulation of host innate immune signalling by C. symbiosum-encoded "symbiosis factors" that allow the symbiont to colonize and thrive within the host milieu. To this end, I charted D. mexicanum microbial community structure, delineated C. symbiosum-encoded gene products with the potential to modulate host recognition and immune response and identified host genes and pathways involved in innate immune signalling pathways.

The specific aims for each data chapter are as follows:

- 1. Chapter 2: Chart the microbial community associated with *D. mexicanum*. Quantitative, taxonomic and metabolic characterization of the sponge microbial community with emphasis on diversity and structure and the prevalence of *C. symbiosum* in *D. mexicanum* and other sponge species.
- Chapter 3: Identify and characterize potential archaeal symbiosis factors. Comparative genomic analysis between the genomes of *C. symbiosum* and free-living thaumarchaea to identify *C. symbiosum* genes potentially involved in symbiosis. Biochemical approaches to assess biological activity of a putative symbiosis factor.
- 3. Chapter 4: Identify putative host pattern recognition receptors and innate immune pathways. Compare gene expression data from sponge hosts to metazoan innate immunity pathways involved in microbial recognition and response to reconstruct common and unique pathway component among and between metazoa.

Chapter 2: Microbial community structure in *Dragmacidon mexicanum* and other sponge species from northeast Pacific coastal waters

2.1 Introduction

Marine sponges are important members of benthic ecosystems and provide habitat to many invertebrate and fish species. As highly efficient suspension feeders that can filter 24L per g sponge per day, sponges are an important link between benthic and pelagic ecosystems and nutrient cycling (22, 25-28). Due to their filtering activities, sponges constantly engage in nutritional, pathogenic and beneficial interactions with billions of microbes. Consequently, the sponge immune system must be able to differentiate between symbionts, pathogens and food (43). Given their basal phylogenetic position with regards to animals, the biology of sponges can provide insight into the evolution of cell adhesion, signalling and innate immunity systems (44, 45), presenting an opportunity to chart conserved modes of microbial recognition and symbiont selection by animal hosts.

Sponge microbiota are distinct from those of ambient seawater, and comprise a complex mixture of extracellular and intracellular symbionts exhibiting transient and stable associations (21, 59, 60). Indeed, a number of bacterial and archaeal lineages form distinct sponge-specific phylogenetic clades (18, 21, 63). Some sponge-microbial associations are sponge-species-specific (141) as conspecific sponges from different locations have more similar communities than other sponges species from the same location (16). Most members of the sponge microbial community appear to be active as indicated by similarities between small subunit ribosomal RNA (SSU rRNA) gene and SSU rRNA clone libraries (142). Cultivation independent

sequencing approaches suggest the roles of sponge symbionts in production of secondary metabolites, biosynthesis of vitamins, cofactors and amino acids, as well as roles in carbon and nitrogen metabolism (50-55). However, factors underlying selective retention of symbiotic microbes by sponges are currently unknown.

The association of *Cenarchaeum symbiosum* with the demosponge *Dragmacidon mexicanum* is the first described example of a sponge-archaeal symbiosis (129). C. symbiosum is highly abundant and is the sole archaeal representative in its host (129). Related thaumarchaea have since been identified in stable associations with sponges from the Mediterranean, Korean and Australian coasts (131-133), and form a monophyletic sponge-specific lineage with C. symbiosum based on their SSU rRNA gene sequences (21). Only a few other animal-archaeal symbioses have been described, including associations of methanogenic archaea in humans and termites (101, 136, 137). Thus the D. mexicanum - C. symbiosum symbiosis is interesting from an evolutionary perspective in terms of conserved and novel mechanisms of host recognition and signalling. However, the D. mexicanum microbial community structure, including non-archaeal symbionts, remains cryptic. Holistic understanding of microbial community structure is needed to provide a baseline for differentiating between symbionts, pathogens and food in the sponge milieu. Moreover, to establish the D. mexicanum - C. symbiosum association as a model for investigating symbiont recognition and signalling, it is important to determine how the D. mexicanum community compares to other sponge species.

Here I use a combination of clone libraries, quantitative PCR and pyrotag sequencing to describe, with three-domain resolution, *D. mexicanum* microbiota in relation to 11 other species representing 3 sponge classes from coastal Northeast Pacific Ocean waters. Phylogenetic analysis and multivariate statistics were used to identify core and indicator microbes, and

intraspecific variation of *D. mexicanum* microbiota was examined to better define transient and host-specific associations.

2.2 Materials and methods

2.2.1 Sample collection and processing

Nineteen marine sponge specimens (Figure 2.1) were collected by SCUBA from Naples Reef, California, USA at a depth of 8-10 m in September 2007, and from Howe Sound, British Columbia, Canada at depths of 15-35 m in June 2009 (Table 2.1). Three additional sponges were collected from Howe Sound at an unknown date were maintained at the Vancouver Aquarium until they were obtained for this study in June 2009 (Table 2.1). All sponges were transported to the laboratory, kept in an aquarium filled with artificial sea water for 2 to 6 weeks and subsequently frozen at -80°C. Sponges were maintained in aquaria for several weeks prior to cryostorage to reduce the size of nonspecific microbial populations. Frozen sponges were ground using a mortar and pestle, with liquid nitrogen added to the specimens. Ground samples were weighed, placed in sterile tubes and stored at -80°C until further processing.



Figure 2.1 Sponges in sample set.

Twelve sponge specimens were collected from coastal waters in British Columbia (BC), top, and ten sponges were collected in California (SB), bottom.

2.2.2 DNA isolation and purification

Frozen, ground sponge tissue samples were washed twice in sterile phosphate-buffered saline solution, suspended in sucrose buffer and homogenized using a Dounce homogenizer. Next, lysozyme and RNase A were added to the homogenate to final concentrations of 1 mg/ml and 20 µg/ml, respectively. The subsequent incubation at 37°C for 1 hour, addition of Proteinase K and sodium dodecyl sulphate to final concentrations of 0.5 mg/ml and 1% (w/v) respectively, incubation at 55°C for 2 hours, phenol:chloroform extraction, buffer exchange and concentration were performed as previously described (2). Total DNA concentrations were determined using PicoGreen® reagent (Invitrogen, Carlsbad, CA, USA). DNA quality was assessed by gel electrophoresis and absorbance profiles observed on a NanoDrop Spectrophotometer (NanoDrop Technologies, Inc, Thermo Scientific, Waltham, MA, USA). Finally, DNA from each sponge

was further purified using cesium chloride density gradient centrifugation (3). Gel electrophoresis and PicoGreen® were used to assess quality and quantity of purified DNA.

2.2.3 Quantification of archaeal and bacterial taxonomic marker genes

Archaeal and bacterial abundance within sponge tissues was determined using quantitative polymerase chain reaction (qPCR). Total archaeal and bacterial SSU rRNA gene copy numbers per gram of sponge tissue were determined using archaeal and bacterial universal primer sets previously described (2). C. symbiosum specific SSU rRNA gene copy numbers were determined using the taxon - specific primers 369F (5'-TACACGGCAGGCTACGG) and 509R (5'-GCTAAAGAAATCTTTTACCGGTC) designed for this study. Each 20 ml reaction contained 10 ml iQ[™] SsoFast[™] EvaGreen[®] Supermix (Biorad Hercules, CA, USA), 300 nM (total archaea and bacteria SSU rRNA genes) or 500 nM (C. symbiosum SSU rRNA and serpin genes) final concentration of each primer, 1.5µl of cesium chloride purified, sponge tissue derived DNA, with the remaining volume made up by sterile nuclease-free water. All reactions were performed in white 96-well qPCR plates (BioRad) and were run on the CFX96TM PCR detection system (BioRad) under the following thermocycling conditions: initial denaturation at 98°C for 2 minutes, followed by 40 cycles of 98°C for 2 seconds, primer annealing at 55°C (bacteria), 65°C (archaea) or 52°C (C. symbiosum SSU rRNA) for 5 seconds. The fluorescence on the plate was measured after each cycle, and a melting curve analysis was carried out after the completion of 40 cycles. A 10-fold dilution series for each of the targets was used to generate the standard curve. Standards were prepared from clone libraries as described in (2). The dilution series, in copies per μ l, ranged from $4x10^1$ to $4x10^7$, $8x10^1$ to $8x10^7$, and $9x10^1$ to $9x10^7$ for bacteria, archaea, and C. symbiosum SSU rRNA genes, respectively. Real-time data were

analysed using CFX ManagerTM software (BioRad). The limit of detection for the experiments was established by comparing the Cq values between no template controls, the most dilute standard and sample dissociation curves. The Cq threshold was 26 cycles for total bacteria, 38 cycles for total archaea, and 30 cycles for *C. symbiosum* - specific SSU rRNA assays. Reaction efficiencies ranged from 92-105% for archaea, and 94-103% for bacteria gene quantification experiments. For *C. symbiosum* SSU rRNA quantification optimized primer efficiency was 90-108%, however experimental efficiency varied (118-120%). At least two experiments were performed per assay, with each sample in triplicate in each experiment.

2.2.4 Preparation of PCR amplicons for pyrosequencing

The V6-V8 loop of the SSU rRNA genes was amplified using three-domain universal primers 926F (5'-AAACTYAAAKGAATTGACGG) and 1392R (5'-ACGGGGGGGTGTGTGTRC) and the appropriate chemistry and thermocycling conditions (13). To allow for multiplexed sequencing, each sample was amplified with a unique MID adaptor - ligated (454 Life Sciences, Branford, CT, USA) fusion reverse primers. PCR products were purified using the QiaQuick PCR purification kit (Qiagen), eluted in 20 mM Tris pH 8, quantified using the Picogreen® reagent. SSU rRNA amplicons were pooled at 30 ng DNA for each sample. Purified PCR amplicons were sequenced on the Roche 454 GS-FLX platform using Titanium series chemistry (McGill University and Génome Québec Innovation Center).

2.2.5 Three-domain variable region sequence data analysis

Pyrotag sequences were filtered, clustered and mapped to reference taxonomy using the Quantitative Insights Into Microbial Ecology (QIIME) software package (14). Sequences were

removed if they were shorter than 200 bp, contained homopolymers or contained ambiguous bases. To identify singleton and chimeric sequences, data were clustered at 99% sequence similarity, and representative sequences were run through chimera prediction software. Sequences that fell into clusters containing only one sequence (i.e. singletons) or chimeric clusters were removed. The remaining sequences were re-clustered at 99% similarity since the presence of chimeric and singleton sequences may have affected their initial clustering. The resulting clusters were used to generate an OTU matrix. The taxonomic identity of each OTU was assigned by BLAST against SILVA database release 111 (19) clustered at 99% identity, and confirmed by querying against the Greengenes database (20).

For statistical analyses requiring even sampling effort 1,800 sequences from each sample were randomly subsampled for 10 iterations. For each iteration, sequences were clustered, assigned taxonomy and used for producing an OTU table in addition to calculating alpha diversity in QIIME. Calculation of mean alpha diversity, standard error and visualization of the resulting rarefaction curves was executed in SigmaPlot version 7.101 (Systat Software, San Jose, CA). Hierarchical cluster analysis (HCA) was performed using Manhattan distance and average linkage method for distance calculations implemented in R. Nonmetric multidimensional scaling (NMS) was used to investigate relationships between microbial communities in seawater and those of different sponge individuals using the following options: Sørensen distance, random starting coordinates, and 250 runs each of real and randomized data. NMS was performed both on subsampled and total OTU matrices, with each matrix relativized by sample or converted into presence-absence data. Ordinations showed highly similar patterns in each case. To compare sponges based on a specific microbial class, the entire OTU matrix was divided into separate class-level OTU matrices, which were randomly subsampled, at 250 reads per sponge, for
multiple iterations. Multi-response permutations procedure (MRPP) was performed to compare within- and between-group similarity. Indicator species analysis (ISA) was performed on relativized data to identify indicator OTUs (15) for sponge groups defined by HCA. PC-ORD version 5.10 (MjM Software, Gleneden Beach, OR, USA) was used for NMS, ISA and MRPP analyses.

2.2.6 Archaeal SSU rRNA gene amplification, library production and screening

SSU rRNA gene libraries were generated, and subsequently screened, for the Santa Barbara sponges using a previously described method (2). Resulting fingerprint patterns were inspected visually to identify unique restriction patterns. At least one representative of each unique restriction fragment length polymorphism (RFLP) pattern observed was sequenced (Michael Smith Genome Sciences Center). Sequences were edited using Sequencher software (Gene Codes Corporation, Ann Arbor, MI, USA). Chimeric sequences were identified using the open source application Bellerophon (4).

2.2.7 Eukaryotic SSU rRNA gene PCR amplification

Eukaryotic SSU rRNA genes in the purified sponge DNA were amplified by PCR using the universal primers EukF (5'-AACCTGGTTGATCCTGCCAGT) and 1391R (5'-GACGGGCGGTGWGTRCA) under the following PCR conditions: 3 minutes at 94°C, 30 cycles of denaturation at 95°C for 20 seconds, annealing at 58°C for 20 seconds, and extension at 72°C for 1 minute, followed by a final 10 minute extension at 72°C. Each 50 μ l reaction contained 2 μ l of template DNA, which was first incubated with 30 μ g of bovine serum albumin for 10 minutes at 95°C, 300 nM each forward and reverse primer, 1 mM deoxynucleotides, 1 unit Herculase II Fusion DNA polymerase (Stratagene, La Jolla, CA, USA), and the Stratagene PCR buffer at 1× concentration. Amplification products from triplicate reactions were pooled and purified using either gel purification or PCR purification kits (Qiagen, Germantown, MD, USA) and were directly sequenced (Macrogen or Genewiz). For samples in which PCR products consisted of more than one band, SSU rRNA libraries were constructed, screened using RFLP analysis, and representative clones were sequenced (Macrogen or Michael Smith Genome Sciences Center). Sequence data were processed manually using Sequencher software.

2.2.8 Phylogenetic analysis of SSU rRNA sequences

SSU rRNA sequences were analyzed using the ARB software package (5). First, sequences were aligned to the closest relative in the full-length SILVA 111 database using the SINA aligner (www.arb-silva.de) (6, 18) and aligned sequences were imported into ARB. For phylogenetic analysis of archaeal clones, assembled sequences were first assigned to operational taxonomic units (OTUs) at >99% identity (0.00 distance), a representative sequence for each OTU was selected using DOTUR (Schloss, 2005) prior to alignment of the representative sequences to the SILVA database. The sequences reported in this study and selected reference sequences were exported from ARB and re-aligned using MUSCLE (7). The resulting alignment was imported into Mesquite (8) for manual refinement. Maximum likelihood phylogenetic trees were inferred using an HKY+4G+I and GTR+G+I models of nucleotide evolution using PHYML (9) for archaeal and sponge trees, respectively. The program also estimated the proportion of invariable sites, G distribution, and the transition to transversion ratio. A consensus tree of 100 bootstrap replicates was assembled to determine the confidence at each node. The

archaeal and eukaryotic trees were visualized using NJplot (10) and Interactive Tree of Life (11), respectively.

Sponge ID	Sampling site	Sampling date	Taxonomy* (Family, Order, Class)	Bacterial SSU rRNA gene copies/g (SE)	Archaeal SSU rRNA gene copies/g (SE)	C. symbiosum SSU rRNA gene copies/g (SE)
SB1	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	$\begin{array}{c} 4.44 \text{x} 10^8 \\ (5.01 \text{x} 10^7) \end{array}$	$\begin{array}{c} 6.32 \times 10^8 \\ (4.09 \times 10^7) \end{array}$	$2.38 \times 10^{9} \\ (2.30 \times 10^{8})$
SB2	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	$\begin{array}{c} 4.60 \mathrm{x10^8} \\ (6.72 \mathrm{x10^7}) \end{array}$	$\begin{array}{c} 6.03 \text{x} 10^8 \\ (3.72 \text{x} 10^7) \end{array}$	3.30x10 ⁹ (1.63x10 ⁸)
SB3	Naples Reef, CA, USA	September 2007	Microcinidae, Poecilosclerida, Demospongiae	$\begin{array}{c} 1.51 \text{x} 10^7 \\ (2.48 \text{x} 10^6) \end{array}$	Not Detected (NA)	Not Detected (NA)
SB4	Naples Reef, CA, USA	September 2007	Tethyidae, Hadromerida, Demospongiae	1.14x10 ⁸ (2.90x10 ⁷)	1.53x10 ⁷ (2.78x10 ⁶)	Not Detected (NA)
SB5	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	$\begin{array}{c} 4.48 \text{x} 10^7 \\ (3.76 \text{x} 10^6) \end{array}$	1.05x10 ⁸ (4.54x10 ⁶)	$8.20 ext{x} 10^7$ (2.76x10 ⁶)
SB6	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	3.88x10 ⁸ (6.02x10 ⁷)	5.04x10 ⁸ (2.99x10 ⁷)	2.43x10 ⁹ (1.09x10 ⁸)
SB7	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	$\frac{1.07 \text{x} 10^8}{(2.37 \text{x} 10^7)}$	9.73x10 ⁷ (5.90x10 ⁶)	1.89x10 ⁹ (4.60x10 ⁷)
SB8	Naples Reef, CA, USA	September 2007	Axinellidae, Halichondrida, Demospongiae	1.95x10 ⁸ (3.29x10 ⁷)	3.37×10^8 (1.88 \text{10}^7)	1.59x10 ⁹ (1.31x10 ⁸)
SB10	Naples Reef, CA, USA	September 2007	Tethyidae, Hadromerida, Demospongiae	2.24x10 ⁸ (5.40x10 ⁶)	3.62×10^7 (2.52 \ 10^6)	Not Detected (NA)
SB11	Naples Reef, CA, USA	September 2007	Tethyidae, Hadromerida, Demospongiae	$\frac{1.63 \text{x} 10^8}{(2.42 \text{x} 10^7)}$	9.94x10 ⁷ (2.00x10 ⁷)	Not Detected (NA)
BC1	Howe Sound, BC, Canada	June 2009 **	Chalinidae, Haplosclerida, Demospongiae	5.90×10^{8} (7.85x10 ⁷)	$6.91 \times 10^{6} \\ (9.09 \times 10^{5})$	Not Detected (NA)
BC2	Howe Sound, BC, Canada	June 2009	Latrunculiidae, Poecilosclerida, Demospongiae	5.09x10 ⁸ (7.76x10 ⁷)	$\frac{1.46 \text{x} 10^8}{(1.10 \text{x} 10^7)}$	Not Detected (NA)

Table 2.1 Sponge sample collection, taxonomy summary and quantification of bacteria and archaea in sponges.

Sponge ID	Sampling site	Sampling date	Taxonomy* (Family, Order, Class)	Bacterial SSU rRNA gene copies/g (SE)	Archaeal SSU rRNA gene copies/g (SE)	C. symbiosum SSU rRNA gene copies/g (SE)
BC3	Howe Sound, BC, Canada	June 2009 **	Chalinidae, Haplosclerida, Demospongiae	$8.10 \times 10^{8} \\ (6.80 \times 10^{7})$	$\begin{array}{c} 4.06 \text{x} 10^7 \\ (1.21 \text{x} 10^6) \end{array}$	Not Detected (NA)
BC4	Howe Sound, BC, Canada	June 2009	Rossellidae, Lyssacinosida, Hexactinellida	7.85x10 ⁶ (1.30x10 ⁶)	Not Detected (NA)	Not Detected (NA)
BC5	Howe Sound, BC, Canada	June 2009	Halichondriidae, Halichondrida, Demospongiae	3.39x10 ⁹ (2.14x10 ⁸)	2.30x10 ⁸ (3.01x10 ⁷)	Not Detected (NA)
BC6	Howe Sound, BC, Canada	June 2009	Aphrocallistidae, Hexactinosida, Hexactinellida	1.45x10 ⁸ (2.00x10 ⁷)	$\begin{array}{c} 2.09 \text{x} 10^6 \\ (1.60 \text{x} 10^5) \end{array}$	Not Detected (NA)
BC7	Howe Sound, BC, Canada	June 2009 **	Chalinidae, Haplosclerida, Demospongiae	4.90x10 ⁸ (1.00x10 ⁸)	3.93x10 ⁷ (6.80x10 ⁶)	Not Detected (NA)
BC8	Howe Sound, BC, Canada	June 2009	Geodiidae, Astrophorida, Demospongiae	$\begin{array}{c} 2.23 \mathrm{x10^{10}} \\ (1.63 \mathrm{x10^{9}}) \end{array}$	3.48x10 ⁹ (2.69x10 ⁸)	Not Detected (NA)
BC9	Howe Sound, BC, Canada	June 2009	Hymeniacidonida e, Halichondrida, Demospongiae	3.70x10 ⁹ (4.31x10 ⁸)	Not Detected (NA)	Not Detected (NA)
BC10	Howe Sound, BC, Canada	June 2009	Chalinidae, Haplosclerida, Demospongiae	1.01x10 ⁹ (8.12x10 ⁷)	1.49x10 ⁷ (1.38x10 ⁶)	Not Detected (NA)
BC11	Howe Sound, BC, Canada	June 2009	Chalinidae, Haplosclerida, Demospongiae	6.48x10 ⁸ (5.21x10 ⁷)	6.54x10 ⁸ (2.91x10 ⁷)	Not Detected (NA)
BC12	Howe Sound, BC, Canada	June 2009	Clathrinidae, Clathrinida, Calcarea	$\begin{array}{c} 4.47 \text{x} 10^8 \\ (7.14 \text{x} 10^7) \end{array}$	$1.49 \text{x} 10^8$ (1.46x10 ⁷)	Not Detected (NA)

* based on BLAST hits of full-length SSU rRNA gene sequences with highest % coverage and identity; minimum 97% ID; ** Collection date of sample from the wild unknown; recieved from the Vancouver Aquarium in June 2009..

2.3 Results

2.3.1 Composition of sponge sample set

Twenty-two sponges encompassing 3 classes (Demospongiae, Hexactinellida and Calcarea), 8 orders, 11 genera and 12 species were collected from southern California (SB) and

British Columbia (BC) coasts (Figure 2.2 and Table 2.1). Most of the sponges, including all SB and nine BC were demosponges, the most abundant sponge class. British Columbia sponges also included two glass sponges and one calcareous sponge.



Figure 2.2 Distance tree of sponge SSU rRNA gene sequences.

Bootstrap values are based on 100 replicates using the maximum likelihood method for inferring phylogeny, and are shown for branches with greater than 50% support. The tree is rooted on Choanoflagellida. The scale bar represents 0.1 substitutions per site. Sponges are encoded by a filled symbol according to group membership determined by HCA shown Figure 2.4.

2.3.2 Quantification of archaea and bacteria in sponges

Quantification of bacterial and archaeal abundance is necessary for determining microbial distribution patterns among and between sponge hosts. To this end, I employed domain-specific SSU rRNA gene quantitative PCR assays to enumerate bacterial and archaeal abundance within sponge tissues. Two sponges, SB3 and BC4, were low microbial abundance (LMA) sponges and did not contain archaea in quantities detectable by qPCR (Table 2.1). The other twenty sponges were high microbial abundance (HMA) sponges. High microbial abundance sponges host microbial communities at concentrations 100 to 10,000 times greater than those in surrounding seawater, whereas LMA sponges contain smaller communities at concentrations 1 to 100 times greater than those in surrounding seawater (35, 56). The range of bacterial and archaeal abundances for LMA and HMA sponges was between 10^6 to 10^7 and 10^8 to 10^{10} SSU rRNA gene copies per gram tissue, respectively, and are consistent with quantification from other HMA and LMA sponge species (143, 144). With the exception of the six D. mexicanum sponges, all HMA sponges had a greater proportion of bacteria than archaea, or in the case of BC9, contained only bacteria. Conversely, D. mexicanum sponges generally had slightly higher, or at least equivalent, archaeal to bacterial abundances. To investigate the specificity of the D. mexicanum and C. symbiosum interaction, I developed a C. symbiosum-specific qPCR assay targeting the SSU rRNA gene. C. symbiosum was detected only in D. mexicanum using this approach (Table 2.1). C. symbiosum numbers generally exceeded total archaeal gene copy quantities likely due to differences in primer amplification efficiency and suggested that C. symbiosum was the predominant, or sole, archaeal member of the D. mexicanum community.

2.3.3 Community structure relationships between sponges

Following estimation of bacterial and archaeal abundance in LMA and HMA sponges, pyrotag sequencing was used to survey sponge microbiota with three-domain resolution. A total of 237,672 high quality SSU rRNA gene pyrotag sequences clustered into 9,865 operational taxonomic units (OTUs) at 99% sequence identity (Table 2.2). Richness was assessed using a rarefaction curve of observed OTUs as a function of sub-sampled sequence sets (Figure 2.3). Generally, observed OTU richness was lower in SB than in BC sponges and was lowest in *D. mexicanum* sponges. Consistent with this observation, Simpson's diversity index calculated using complete sequence sets was lower in SB sponges, especially *D. mexicanum* (Table 2.2). Highest richness and diversity were found in BC1 and BC12.



Figure 2.3 Observed species richness rarefaction curve of mean OTU diversity. Analysis is based on 10 random sub-sampling iterations, the error bars indicate standard error of the mean. Colours represent group membership determined by HCA shown in Figure 2.4.

Community structure relationships between sponges were determined using hierarchical cluster analysis (HCA) and non-metric multidimensional scaling (NMS). Hierarchical cluster analysis revealed that the twenty-two communities formed four distinct groups based on sponge species and location (Figure 2.4A). The *D. mexicanum* sponges clustered away from the other sponges, forming Group 1 (Figures 2.2 and 2.4A). Group 2 consisted of the remaining four SB sponges, including three conspecific individuals. Group 3 was comprised of three sponges of the same species. Curiously, a Group 3 conspecific sponge, BC1, was associated with Group 4 instead. Group 4 contained the most sponges encompassing 9 species. The MRPP results, p = 0.0000, A = 0.2681, indicated that sponges within groups were significantly more similar to each other than to members of other groups, thus the observed groups were statistically and ecologically significant. These four groups were subsequently used to structure downstream comparisons.





A. Hierarchical clustering of sponges based on relative OTU distribution patterns. **B**. Non-metric multidimensional scaling ordination of sponges based on community compositional profiles. Sponges are encoded by a filled symbol according to groups determined by hierarchical cluster analysis. Pearson correlations between microbial SSU rRNA

gene copy numbers and the ordination axes were calculated and shown as a vector ($r^2 \ge 0.30$ cutoff). Direction and length of vector indicates strength of correlation with each axis.

Non-metric multidimensional scaling produced a two-dimensional solution with a cumulative r^2 value of 0.780, implying that 78% of the total variance in the data was captured along the two ordination axes (Figure 2.4B). The final stress for the solution was 10.16 after 68 iterations and the final instability was 0.00. The stability of the solution was confirmed by inspecting a stress versus iteration number plot. The ordination revealed similar community structure relationships to HCA. However, partitioning of sponge microbiota by geography was more evident using NMS. Non-metric multidimensional scaling was also used to define correlations between OTU profiles and bacterial, archaeal and *C. symbiosum* SSU rRNA gene abundances determined by qPCR. Correlation analysis with each axis showed that only *C. symbiosum* had a significant correlation ($r^2 \ge 0.3$ with either axis) with the ordination ($r^2 = 0.492$ with axis 2). The observed relationship between *C. symbiosum* SSU rRNA gene abundance and position of sponges along axis 2 was due to the presence of large *C. symbiosum* populations in *D. mexicanum* sponges.

Sponge	# SSU rRNA sequences passing QC	Average sequence length (bp)	# SSU rRNA OTUs at 99% ID	Simpson's diversity index
SB1	6,853	423	554	0.8079
SB2	14,369	419	976	0.8612
SB3	5,055	419	775	0.9194
SB4	11,622	420	970	0.7755
SB5	6,648	421	673	0.8071
SB6	14,152	418	962	0.8932
SB7	11,515	413	783	0.8720
SB8	8,939	429	604	0.8049
SB10	5,390	428	1,019	0.9437
SB11	11,106	424	1,000	0.9177

Table 2.2 Amplicon pyrosequencing data summary and community diversity.

Sponge	# SSU rRNA sequences passing QC	Average sequence length (bp)	# SSU rRNA OTUs at 99% ID	Simpson's diversity index
BC1	10,223	418	2,062	0.9900
BC2	1,821	429	394	0.9561
BC3	13,962	409	1,910	0.9321
BC4	14,446	417	1,966	0.9442
BC5	19,879	417	1,372	0.9515
BC6	10,205	416	1,478	0.9456
BC7	4,303	418	745	0.9052
BC8	19,329	416	1,262	0.9794
BC9	14,811	390	1,794	0.9593
BC10	12,054	398	1,541	0.9211
BC11	11,484	404	891	0.8181
BC12	9,506	420	1,666	0.9813

Core microbiota, here defined as non-sponge sequences present in at least 75% of the sponges in a given group, were identified for groups 1-4 and compared to identify microbes either common to sponges in all four groups or unique to one group. Sponge species-specific microbiota in Groups 1 and 3, each consisting of conspecific sponges, harbored the largest proportion of unique core OTUs, 77% and 84%, respectively compared to Groups 2 and 4 (Figure 2.5). However, Group 2, in which three of the four sponges were conspecific, also harbored more unique than shared core OTUs (Figure 2.5). Only one OTU, assigned to *Colwellia*, was shared among all four groups. Due to sequence heterogeneity within populations of related OTUs, core microbiota were compared at the taxon level. Terminal taxa, that is the consensus taxon identity of the reference Silva database clusters, were used and ranged from order to subspecies levels of taxonomy, with most resolving at the genus level. The number of shared taxa exceeded the number of unique taxa in all groups (Figure 2.5). Group 3 had the largest proportion of unique taxa, while Group 1 shared most of its core taxa with other sponges and had only 6 unique taxa including Candidatus *Endobugula*. *C. symbiosum* was part of the

common core microbiota of the two SB sponge groups. Groups 1 - 3 shared more OTUs and taxa among each other than with Group 4 both at the OTU and taxon levels.



Figure 2.5 Core microbial communities of sponge groups as designated by HCA. Core membership of taxa or OTUs depended on the presence in >75% sponges within a given group (at least 3 sponges for Group 2 and Group 3, and 9 or more sponges for Group 4), with the exception of Group 1 core, in which case presence of an OTU or taxon in all 6 sponges was required.

2.3.4 Taxonomic composition of sponge microbiota

Sponge-associated archaea consisted of two phyla, Euryarchaeota and Thaumarchaeota (Figure 2.6). Euryarchaea were detected in only one sponge, in low abundance (0.05%) based on pyrotag data. Although Thaumarchaoeta were present in all sponges, their relative abundance varied greatly between individuals, ranging from 0.015% to 63.7% of total pyrotags per sponge, and ~1% in LMA sponges, conflicting with observations made by qPCR, possibly due to primer sensitivity. *Cenarchaeum symbiosum* comprised >99.8% of thaumarchaeal sequences and up to ~64% of the *D. mexicanum* community, but was present as part of the rare biosphere, $\leq 0.2\%$ (145, 146) in other sponge species.

Twenty-six bacterial phyla were identified across all sponges (Figure 2.6). BC1 had the highest bacterial diversity at the phylum level, with 24 phyla detected. BC2 had the least number, with 8 bacterial phyla. Only six phyla, Actinobacteria, Bacteriodetes, Cyanobacteria, Firmicutes, Planctomycetes and Proteobacteria, were found in all sponges. The proteobacteria were abundant in all sponges and were also very diverse. Although cosmopolitan phyla were present in all sponges, abundance and OTU composition differed among sponge species and conspecific individuals, with Planctomycetes exhibiting the least variation in composition between sponges. biogeographical Other phyla exhibited distinct distribution patterns. Tenericutes. Thermodesulfobacteria, Fusobacteria, Fibrobacteres and Deferribacteres were only present in BC sponges, and did not partition according to groups identified by hierarchical cluster analysis. Moreover, of the 4 candidate divisions identified, only 3 were found in BC sponges. The fourth candidate division TM6 was present in 12 individuals from BC and SB. Conversely, no bacterial phyla were detected in SB sponges exclusively. Dragmacidon mexicanum sponges had lower bacterial abundance compared to other sponges with bacterial sequences collectively representing 9-22% of total pyrotag reads.

Fewer non-poriferan eukaryote sequences were detected when compared to bacteria in all individuals (Figure 2.6). Only two eukaryotic groups, the opisthokonta and the SAR supergroup, were present in all sponges. For most individuals, porifera reads accounted for >90% of opisthokonta reads. The remaining opisthokonta sequences were affiliated with fungi, holozoa, and other metazoa including nematodes, annelids, and arthropods. BC5 had few sponge reads and opisthokonta in this sponge encompassed primarily the bacterivorous nematode *Halomonhystera disjuncta*. Biogeography was evident for some eukaryotic sponge associates as the algal groups Haptophyta and Cryptophyceae were found only in SB sponges, specifically in

two Group 1 sponges, and the RT5iin25 clade was present only in four of the nine Group 4 sponges. Taxonomy could not be assigned to one OTU.



Figure 2.6 Relative abundance and distribution of archaeal, bacterial and eukaryotic taxa.

The size of each circle represents the relative proportion of the taxon as a percentage of the total number of sequences within each sponge. Open circles representing *C. symbiosum* and Porifera percentages are overlaid on closed thaumarchaea and Opisthokonta circles.

2.3.5 Archaeal phylogeny in SB sponges

Archaeal SSU rRNA clone libraries were constructed from SB sponges to better determine the phylogeny of archaea in these sponges. No archaeal sequences could be amplified for SB3 and SB4 using 20F and 958R archaeal primers (section 2.2 in this thesis), and therefore no libraries were available for these two sponges. All clones from *D. mexicanum* sponges, with the exception of a single clone from SB8, were highly similar to *C. symbiosum* sequences and fell into the C1a-Porifera-A cluster, whereas neither SB10 nor SB11 contained *C. symbiosum* clones (Figure 2.7). These results support the highly specific nature of the *D. mexicanum - C. symbiosum* relationship, consistent with foundational observations made by Preston and colleagues (129). Thaumarchaeal sequences derived from two Group 2 sponges (SB10 and SB11) were part of the C1a- α group, which includes environmental sequences as well as clones from tunicates and other sponge species (Figure 2.7). Euryarchaeota sequences were recovered from three sponges and all clustered with methanogenic archaea of the genus *Methanosaeta*.



Figure 2.7 Distance tree of archaeal SSU rRNA gene sequences recovered from SB sponges. Bootstrap values are based on 100 replicates using the maximum likelihood method for inferring phylogeny, and are shown for branches with greater than 50% support.

2.3.6 Indicator species analysis

Indicator species analysis was performed using the PC-ORD software package to identify significant indicator microbes for each sponge group defined by hierarchical clustering. Collectively, indicator OTUs encompassed almost 90% of all the reads in Group 1 sponges, 44% and 42% for Groups 2 and 3, respectively, and just over 1% for Group 4. Although most indicator OTUs belonged to the rare biosphere, Groups 1 and 2 also had abundant indicator OTUs (Appendix A). Group 1 consisted of *D. mexicanum* sponges and contained the most

significant and highest ranking indicator OTUs. Of the 312 indicator OTUs identified in this group 24 were affiliated with unique terminal taxa and 17 orders (Figure 2.8 and Appendix A). Approximately 39% of all group 1 indicator OTUs were affiliated with *C. symbiosum*. Group 2 harbored 229 indicator OTUs affiliated with 75 unique terminal taxa and 35 orders, while Group 3 harbored 206 indicator OTUs representing 49 unique terminal taxa and 23 orders. The taxonomic composition of indicators was quite different between groups. A large proportion of indicator OTUs for Groups 1-3 consisted of Demospongiae-affiliated OTUs, reflecting the species-specific nature of each group. Group 4 harbored only 3 indicator OTUs representing three proteobacterial taxa (Figure 2.8A).

2.3.7 Dragmacidon mexicanum core microbiota

To assess the consistency of microbial associations with *D. mexicanum*, core (present in all six) and unique (present in only one sponge) microbiota were examined. Most of the OTUs in any one of the six *D. mexicanum* individuals were neither core nor unique, but rather, were shared between two to five sponges (Figure 2.9). However, even though core OTUs represented only ~20% of total diversity, they were highly abundant, representing ~90% of all pyrotag sequences recovered from individual sponges. Unique OTUs were part of the rare biosphere. Archaeal OTUs consisted solely of *C. symbiosum*, representing 79% of the core *D. mexicanum* microbiota (Figure 2.8B). Bacterial OTUs represented seven phyla, comprising 20% of the core *D. mexicanum* microbiota. Cyanobacteria, Bacteroidetes, Actinobacteria, Spirochaetes, Planctomycetes and Verrucomicrobia were affiliated with a single order per phylum as indicated in Figure 5B. Proteobacterial OTUs were the most abundant core bacterial phyla. The γ -proteobacteria were comprised of three orders, the Alteromonadales, which consisted of

Colwellia species and Candidatus *Endobugula*, the Oceanospirillales, comprised of MBAE14, *Amphritea*, *Neptuniibacter*, *Pseudospirillum* and *Neptunomonas*, and the sponge-associated E01-9C-26 marine group. Due to the high proportion of γ -proteobacteria in the *D. mexicanum* bacterial community, as well as apparent similarity in composition of this bacterial class in *D. mexicanum*, statistically significant differences in γ -proteobacteria across all 22 sponges were tested using MRPP and pairwise comparisons. MRPP results, *p* = 0.0000, A = 0.2320, indicated that sponges within groups were significantly more similar based solely on the γ -proteobacteria. Furthermore, pairwise comparisons indicated that *D. mexicanum* sponges were significantly distinct from the other sponges (*p* < 0.05 in each case, Table 2.3). The α -proteobacteria consisted of the OCS116 clade and Rhodobacterales, comprised of the *Leisingera*, *Ruegeria*, *Phaeobacter* and *Roseobacter* lineages. Uncultured alveolates were the core eukaryotic component of *D. mexicanum* microbiota.



Figure 2.8 Indicator and core community members.

A. (Top) Distribution of the p value against the indicator value for each OTU. The black box denotes significant indicator OTUs, with an indicator value greater than 65, and p value less than 0.01. (Bottom) Number and taxonomic distribution of indicator OTUs at the order-equivalent taxonomic level. **B** Core microbial community composition in *D. mexicanum* sponges. Relative proportions are based on OTU abundance in all six *D. mexicanum* sponges.



Figure 2.9 Richness and abundance of OTUs detected in *D. mexicanum* sponges.

The proportions of OTUs and total pyrotag read abundances either core (common to all 6 sponges), shared (between 2 to 5 individuals), or unique (present in 1 sponge) are shown for each individual sponge.

Groups compared	Т	Α	р*
1 and 2	-5.6601	0.2446	0.0014
1 and 3	-5.0308	0.3040	0.0015
1 and 4	-9.0805	0.2440	0.0001
2 and 3	-3.3760	0.1464	0.0096
2 and 4	-6.6295	0.1003	0.0003
3 and 4	0.6746	-0.0066	0.7315

Table 2.3 Statistical comparison of gamma-proteobacteria composition in sponges.

2.4 Discussion

2.4.1 Core and indicator sponge microbiota

Sponge-microbial associations were species-specific and secondarily dependent on location as determined by HCA and NMS. Comparison of core and indicator OTUs supported this observation, with few OTUs and indicators shared between sponge groups defined by HCA. Groups 1-3 harbored high-ranking indicators with more unique than shared core OTUs. Moreover, the most abundant core OTUs within conspecific sponges exhibited limited variation, even at the 99% identity threshold. Consistent with previous studies, sponge microbiota varied independently of sponge phylogeny (147) since representatives from three sponge classes formed part of the same group (Group 4). This suggests that individuals of the same species select for similar microbiota and that this pattern in microbial community structure can arise independently across multiple sponge lineages.

Sponge core microbiota largely contained similar taxa that differed at the OTU level, suggesting that sponge hosts select for functionally redundant taxa with strain specificity. While functional roles of symbionts cannot be predicted from taxonomy, it is reasonable to suppose that microbial OTUs from the same reference cluster or belonging to the same genus play common physiological roles, in line with recent evidence for functional redundancy of microbial communities in diverse sponges (148). The extent to which observed OTU diversity within groups reflects recurrent environmental acquisition or genetic drift between vertically transmitted symbionts remains to be determined.

Most of the sponges in this study including the calcareous sponge and one of the hexactinellid sponges were HMA. While HMA sponges harbored a range of archaeal to bacterial ratios, bacteria represented >95% of all pyrotags in LMA sponges suggesting anatomical or

behavioral differences promoting domain-specific symbiotic associations (58). Previous studies have reported that LMA sponges harbor less diverse microbiota than HMA sponges (57, 142). However, the results do not support this since no obvious differences between the LMA and HMA sponge-associated bacterial diversity were detected. Indeed, BC HMA sponges had the highest and lowest Simpson's diversity indices, with lowest richness observed in *D. mexicanum* sponges. Neither of the LMA sponges surveyed in this study harbored low bacterial diversity although both lacked archaea in abundances detectable by qPCR. The two LMA sponges had dissimilar microbiota and clustered into different groups that included HMA sponges. This is consistent with previous observations that communities within LMA species were distinct from other LMA and HMA sponges (57). Interestingly, the LMA BC4 community was dominated by a single *Clostridium* OTU reminiscent of the uneven representation of *C. symbiosum* within *D. mexicanum* microbiota. The discrepancies in LMA sponge community diversity reported in this and previous studies may be partially attributed to differences in methods used (57, 142) and invite more research utilizing more integrated molecular and microscopy approaches.

2.4.2 Taxonomic composition of sponge microbiota

The taxonomic composition of sponge microbiota was similar to those previously reported for other sponges and was consistent across three sponge classes and between LMA and HMA sponges (18, 21, 147, 149). Sponge-archaeal associations tend to involve primarily or exclusively Group C1a thaumarchaea, and to a lesser extent euryarchaea (21, 131, 150-152). Likewise, the archaeal component of sponges in this study was predominantly thaumarchaeal although a euryarchaeal OTU (Candidatus *Parvarchaeum*) was detected in BC9 by pyrosequencing and *Methanoseata*-like clones were recovered from SB8, SB10 and SB11

sponges. Since no archaeal clones could be recovered from the third *T. californiana* sponge, these archaea may represent either transient associations, or part of the planktonic community accumulated by filtration. All sponges contained ammonium-oxidizing thaumarchaea. With the exception of some *D. mexicanum* individuals, all sponges contained Group C1a- α thaumarchaea, a group often seen in sponges (150). No representatives of the sponge-specific Group C1a-Porifera C cluster identified in other sponges, including Halichondrid species, were detected in this study (132, 133, 148). The discrepancies in detecting archaea between methods employed in this study may be due to differences in primer specificity, assay sensitivity and template heterogeneity.

A total of 26 bacterial phyla were identified in BC and SP sponges, encompassing most of the taxonomic diversity documented in sponges on a global scale (16, 18, 147, 149). Four of the six cosmopolitan phyla identified in BC and SB sponges (Actinobacteria, Bacteriodetes, Firmicutes and Proteobacteria) have been identified in most sponges, including other NE Pacific Ocean sponges (153). In contrast to previous work defining microbes common to most sponge species, Poribacteria were found as a rare biosphere member in BC8 and were not identified in the remaining sponges (64, 147).

The presence of aerobic, facultative aerobic and anaerobic microbes in all sponges is indicative of fluctuations in oxygen concentrations that accompany changes in sponge pumping activity and suggests metabolic versatility across a range of redox conditions (154, 155). Although bacterial composition was similar among sponges at the phylum level, inter- and intraspecific differences were apparent at finer resolution. Intraspecific variation in bacterial composition generally occurred in less abundant phyla or less abundant families within abundant phyla. These taxa could represent transient associations or filtered food particles. Another reason for differences in community composition between conspecific sponges could be variations in spatial distribution of certain phyla (156). However, since visually similar whole tissue samples were homogenized, such differences should have been minimized.

Eukaryotic interactions with sponges are less described than those of bacteria or archaea. Similar to previous studies, fungal and protist groups were detected, although they represented small proportions of sponge communities (18, 157, 158). Non-poriferan eukaryotic OTUs were not indicator species for any sponge group. Eukaryotic communities varied more than bacteria in Antarctic sponges (159). Since planktonic eukaryotes are a major food source for sponges, it is plausible that the relatively few sequences identified may be food particles rather than symbionts. Sequences from non-sponge metazoa may either be derived from symbiotic animals inhabiting sponges, or possibly from filtered larvae.

2.4.3 Intraspecific variation between *Dragmacidon mexicanum* microbiota

The *D. mexicanum* microbiota differed markedly from other sponges. *Dragmacidon mexicanum* sponges were the only sponges in this study that contained more archaea than bacteria, similar to observations made in *Tentorium semisuberites* (152). I observed minimal intraspecific variation in *D. mexicanum*, considering the microbial community at large and *C. symbiosum* sub-populations. Although most OTUs present in these sponges were not common to all six individuals, shared OTUs represented the vast majority of the community based on SSU rRNA gene sequence abundance.

The *D. mexicanum* archaea were primarily composed of *C. symbiosum*, consistent with foundational observations made by Preston and colleagues (129). More intraspecific variation was detected in bacterial and non-poriferan eukaryotic taxa between different *D. mexicanum*

individuals, suggesting less consistent associations with these domains. The D. mexicanum bacteria differed from other sponge bacteria, and encompassed only 16 bacterial phyla, ranging from 9 to 13 phyla per individual sponge. However, the composition of these phyla differed between D. mexicanum individuals, and only seven bacterial phyla were represented by core OTUs across all six individuals. Among these core OTUs were indicator bacteria, which were all present at low to intermediate abundance, ranging from 0.1% to $\sim 4\%$ of the total community. In addition to representing the most abundant bacterial indicators, γ -proteobacteria were the most abundant core bacteria in D. mexicanum sponges, and exhibited compositional conservation within D. mexicanum that was distinct from other sponges species, consistent with a hostspecific interaction. Most of the bacteria detected in D. mexicanum were also found in other sponges. Bacterial phyla observed in other sponges, including the Gemmatimonadetes, Lentisphaerae, Acidobacteria, TM7 and Chloroflexi (60, 147) were not detected in most D. mexicanum sponges, consistent with a specific selection process. However, the functional basis for this selection is not clear. It would be interesting to determine whether sponges containing Group C1a-Porifera thaumarchaea exhibit similar patterns with respect to reduced bacterial diversity.

Minimal variation was observed between abundant OTUs affiliated with *C. symbiosum* among *D. mexicanum* sponges. Moreover, the core *D. mexicanum* microbiota was dominated by *C. symbiosum* and most of the significant indicator OTUs for Group 1 sponges were *C. symbiosum*, suggesting vertical transmission of the symbiont. Yet, *C. symbiosum* formed part of the rare biosphere in other BC and SB sponges, suggesting previously unrecognized horizontal or environmental acquisition of this symbiont. Regardless of transmission mode, *C. symbiosum* unequivocally dominates the *D. mexicanum* community, positing the existence of specific

signalling and recognition processes between host and a relatively simple symbiotic community. Given these observations, the persistent *D. mexicanum - C. symbiosum* symbiosis provides a model for investigating molecular mechanisms underlying symbiont selection by sponge hosts due to low complexity and numerical abundance. To resolve the currently equivocal mode of selection and illuminate innate immune pathways mediating stable symbioses within the *D. mexicanum* host, Chapters 3 and 4 describe studies utilizing genomics and gene expression profiling in combination with biochemical and cell biological assays to identify and characterize signalling and recognition molecules.

Chapter 3: Genomic and functional characterization of *C. symbiosum*-encoded genes

3.1 Introduction

Sponge symbionts contribute to host fitness and impact nutrient cycling via metabolic exchange, vitamin biosynthesis and nutrient transport (50, 61, 148, 160). Despite the organismal and ecological importance of sponge symbioses, mechanisms mediating symbiont recognition are poorly understood, in part due to the fact that most lineages of sponge-associated microbes do not have cultured representatives (161). The evolution of symbiosis and pathogenesis evolution is associated with genomic changes and adaptations in both the host and microbial partners (78). Thus, genomes of symbiotic microbes are essential for understanding the genetic adaptation of symbionts to the host environment (50, 82). Until recently, microbial adaptations to eukaryotic host colonization have been best characterized in pathogenic interactions. The success and transmission of pathogenic microbes highly depends in their virulence factors (68). These factors can help the colonizing microbe attach to host cells or extracellular matrix, invade the epithelium or intracellular compartments, acquire host micronutrients, and evade the host immune system, among others (68, 69). Symbionts use similar mechanisms enabling them to gain entry into the host milieu, avoid antimicrobial responses and replicate for mutual advantage (69-71).

Bacterial members of the mammalian gut microbiota utilize extracellular features including pili, fimbrae, the cell envelope, as well as secreted proteins to adhere to host mucosa and cell surfaces and protect themselves from host immune responses (78, 79). The cell envelope

contains exopolysaccharides (EPS), which are important for modification of surfaces and specific microbial recognition by host immune cells, contributing to symbiosis maintenance (78, 80). For example, EPS produced by the bacterium *Bifidobacterium breve* allows it to persist in the host, tolerate stress, and has been implicated in evasion of B-cell responses (91). Additionally, symbionts can directly modulate signalling and immune response dependent on pattern recognition receptors (39, 95, 96). Among symbiont proteins that help mediate host-microbial interactions and possibly modulate host immunity are the serine protease inhibitors (serpins) (78, 79). Multiple *Bifidoacterium* species encode serpins, some of which establish symbiotic recognition and may attenuate inflammation in the host (83-85). Further, serpins also have important roles in host-microbial interactions in insects (86-90). Given these observations, it is possible that serpin-mediated immunomodulation encoded in host and symbiont genomes may be conserved across metazoans.

Almost all that is known about host-microbe interactions come from studies on bacterial symbionts and pathogens. There are no known archaeal pathogens, and although archaea form important symbioses with metazoans, yet how they colonize, interact with, and succeed in the host is unknown (99-101, 162). *Cenarchaeum symbiosum* is the sole archaeal symbiont of *Dragmacidon mexicanum*, where it dominates the microbial community (Chapter 2). Given its numerical abundance, a population genome for this symbiont has been assembled, providing preliminary insight into its metabolic capacity (51, 52). Like other Thaumarchaea, *C. symbiosum* carries genes for ammonia oxidation and carbon fixation (51, 52). Thaumarchaea have been shown to tolerate lower ammonium concentrations than bacterial ammonium oxidizers and are considered to be important players in the marine nitrogen and carbon cycles (163, 164). Whether members of this archaeal phylum are strictly autotrophic or are mixotrophic is not yet clear since

47

thaumarchaeal genomes, including that if *C. symbiosum*, encode for a TCA cycle and putative transporters for amino acids and other organic substrates (51, 52, 163, 165, 166). *C. symbiosum* cells are non-motile and occur extracellularly in the sponge mesohyl (129, 130). Since *C. symbiosum* exists in close proximity with sponge cells and other extracellular microbes, *C. symbiosum* likely encodes adaptive traits to evade the sponge immune response and thrive in the host.

In this chapter, I use a combination of comparative genomics with free-living archaea, proteomic analysis and homology modeling to predict molecular determinants of sponge symbiosis, including serpins, using the highly specific *D. mexicanum - C. symbiosum* symbiosis. A series of biochemical experiments were used to test the biological activity of *C. symbiosum*-encoded serpins. The cultivation-independent approach described here provides a framework for identifying and characterizing putative functional genes encoded by sponge-associated microbes to better understand how symbiotic microbiota colonize or maintain a population in their hosts.

3.2 Materials and methods

3.2.1 Comparative genomics of C. symbiosum and other Thaumarchaeota

To identify genomic features potentially important for symbiosis establishment, the *C. symbiosum* genome was compared to 5 thaumarchaeal genomic datasets, including the genomes of the marine *Nitrosopumilus martimus* SCM1, *Nitrosoarchaeum limina* SFB1 from low-salinity waters, the full-length fosmid sequences of uncultivated thaumarchaea from 4000m at Hawaii Ocean Time Series Station ALOHA (HF4000 dataset), the rhizospheric *Nitrosoarchaeum koreensis* MY1 and *Nitrososphaera gargensis* Ga9.2 enriched from a biofilm at a hot spring outflow (Table 3.1). Predicted genes from the thaumarchaeal genomes were compared in a pair-

wise fashion using BLASTp. Quality cutoffs for BLASTp were a bit score ratio >0.4 and expectation value <1E-6. Graphical representations of comparisons were generated using R. Since adaptation to a symbiotic lifestyle can be accompanied by gene gain through lateral gene transfer, IslandPath was used to identify outatuce island regions in the *C. symbiosum* genome (167-169).

Two different *C. symbiosum* populations, "A" and "B" types have been previously described (51, 135). Furthermore, a number of closely related but distinct *C. symbiosum* OTUs were detected in the *D. mexicanum* microbiome characterization presented in Chapter 2. Population heterogeneity can have significant functional implications. Even small genomic variations between strains of the same species can impact host-microbial associations, such as change the host range of *Vibrio fisheri* (170). Thus, genome assembly of other *C. symbiosum* populations is necessary for a greater understanding of the selective pressures acting on *C. symbiosum* (Appendix B.1). However, since the "A" types is consistently more prevalent in all *D. mexicanum* surveyed, comparative genomic and functional analyses in this chapter are based on the "A" type genome (135).

Thaumarchaeote	Habitat	Accession #	Reference
Cenarchaeum symbiosum A	sponge (<i>D. mexicanum</i>) mesohyl	NC_014820	(51)
Nitrosoarchaeum koreensis MY1	soil, rhizosphere	osphere NZ_AFPU01000001	
Nitrosoarchaeum limnia SFB1	low-salinity estuary sediment	NZ_CM00158	(165)
<i>Nitrosopumilus maritimus</i> SCM1	planktonic, marine	NC_010085	(172)
Nitrososphaera gargensis Ga9.2	moderate thermophilic microbial mat	NC_018719	(164)
HF4000 dataset thaumaracheae	planktonic, marine	EU016559 - EU016674	(173)

 Table 3.1 Thaumarchaeal genomes used in comparative analysis

3.2.2 Protein extraction and peptide mapping

Total holobiont (sponge host and associated microbiota) protein was extracted from frozen sponge tissue from one D. mexicanum specimen (SB2) for proteomic sequencing. First, 1 ml of CelLyticTM MT cell lysis reagent (Sigma-Aldrich, St Louis, MO, USA) was added for every 50 mg of sponge tissue, followed by thorough homogenization in a pre-chilled homogenizer. Next, the lysed sample was centrifuged at 12,000 - 20,000 xg for 10 minutes to pelletize cell debris, and supernatants were transferred to a chilled test tube. Total protein concentration in the lysate was determined using a BCA assay and sample volume was determined. Next, urea and thiourea were added to the sample, to a final concentration of 7M and 2M, respectively. After the addition of dithiothreitol to a final concentration of 5 mM, the sample was incubated at 60°C for 30 minutes. Next, the sample was diluted 10-fold with 100 mM NH₄HCO₃ to reduce the salt concentration. Then, CaCl₂ was added to a concentration of 1 mM CaCl₂. The protein sample was then digested with trypsin for 3 hours. The sample was cleaned using a C18 solid phase extraction column (Sigma-Aldrich). Briefly, the column was first conditioned with methanol and then the column was rinsed with water containing 0.1% trifluoracetic acid (TFA). The sample was put through the column, after which the column was washed with 5% acetonitrile and 0.1% TFA and allowed to dry. Finally, the sample was eluted and concentrated. Protein concentration in the samples was measured again using a BCA assay. The sample was flash-frozen in liquid nitrogen and sent to collaborators at the Pacific Northwest National Laboratories (PNNL) for peptide identification by liquid chromatography-tandem mass spectrometry. Proteins were identified from peptide sequences at PNNL using SEQUEST. The C. symbiosum and N. maritimus genomes and a sponge (Oscarella carmela) EST library were used as reference databases for protein identification. The accuracy of peptide identification was

estimated by converting SEQUEST scores to a probability using PeptideProphetTM (174). Only peptides with a PeptideProphet probability score >0.95 were considered expressed and used for downstream analyses. Since protein length can affect the spectral count, normalized spectral abundance factor (NSAF) was used to normalize protein expression (175).

3.2.3 Sequence homology and structure prediction

In addition to identifying genes unique to C. symbiosum, potential symbiosis factors were identified using a series of sequence homology searches. Specifically, protein structure prediction and a hidden Markov model (HMM) search were done to assign functional annotation to hypothetical proteins, verify annotation and classify proteins into broad superfamily membership. All predicted C. symbiosum genes were queried against the Pfam-A database, version 25, using an E-value <0.001 (176). Options to predict active sites and to resolve clan overlaps were selected. The Pfam results were mapped to the genome and were subsequently used to inform homology structure predictions. Protein structures were predicted for C. symbiosum genes that did not have homologs in other thaumarchaea and (i) were hypothetical proteins containing domains with homology to proteins involved in microbe-host interactions, (ii) had annotations similar to known virulence factors, or (iii) were expressed hypothetical proteins in close genomic proximity to other proteins potentially involved in symbiotic interactions using Phyre2 (177). Special attention was given to putative proteins that were validated by proteomic analysis. Multiple sequence alignments were performed using ClustalW to identify regions of sequence conservation for a subset of C. symbiosum putative proteins that had similar annotations or contained similar conserved functional domains (178). The genome viewer Artemis was used to visualize the genomic arrangement of genes and domains identified

as described above to help delineate potential genes and genomic regions of interest (179). Genomic features identified in this study were visualized using Circos (180). Possible function of 23 *C. symbiosum* proteins containing metallo- β -lactamase or lactonase domains was predicted using a sequence similarity network (181-183). Over 6,000 proteins belonging to the metallo- β -lactamase superfamily were included, and were compared by BLASTp. Nodes (proteins) were connected with edges where sequence identity exceeded 50% with an e-value cutoff of 1E-10.

3.2.4 Biological activity of C. symbiosum serpins

3.2.4.1 Cloning, expression and purification of *C. symbiosum* serpins

Two expression systems, *Escherichia coli* pPET28a and the yeast, *Pichia pastoris*, were used to clone and express *C. symbiosum* serpins. However, since archaeal post-translational modification mechanisms are more similar to those of yeast than bacteria, functional assays were performed using recombinant proteins expressed using *P. pastoris*. Three *C. symbiosum* serpin homologues (CENSYa_0537, CENSYa_1229, CENSYa_1605) were successfully cloned, expressed and purified in *P. pastoris*. Serpin genes were cloned in frame, sequences were verified, and the pPICZ α A constructs were transformed into *P. pastoris*, as previously described (184). Expression was induced in *P. pastoris* cell cultures containing the construct as described (184). A control culture transformed with a vector without an insert was used. Culture supernatants containing the secreted His-tagged serpins were pH and ionic strength adjusted, filtered, and loaded onto a HisTrap column packed with Ni²⁺ sepharose resin (GE Healthcare, Little Chalfont, Buckinghamshire, UK). After sample addition and flow-through collection, the column was washed with a binding buffer, and eluted with elution buffer. The wash and elution fractions were collected. The binding buffer was 20 mM sodium phosphate buffer, pH 7.4, with

0.5M NaCl and 5mM or 20 mM imidazole, whereas the elution buffer was 20 mM sodium phosphate buffer, pH 7.4, with 0.5M NaCl and 0.5M imidazole. Elution profiles were generated by plotting the absorbance at 280nm in UV-clear plates. The protein concentration in the fractions was assessed by a BCA assay, and protein purity was assessed by denaturing polyacrylamide gel electrophoresis (SDS-PAGE) and silver-staining of the gel. Western blots, using a polyclonal rabbit antibody raised against E. coli expressed CENSYa 1605 (PL Laboratories, Port Moody, BC, Canada) were performed to confirm serpin presence. Silver staining of the SDS-PAGE gel confirmed that a protein of approximately 45kDa was present in the elution fractions. However, other non-specific bands were also visible following Ni²⁺ column purification. Therefore, all samples that were determined to contain serpin by Western blot were combined, buffer exchanged and purified again using anion-exchange on a HiTrap Q HP column (GE Healthcare) and a salt concentration gradient, ranging from 0.1-1 M in 0.1M increments. The anion exchange purification step yielded protein that migrated as a clean single band on silver-stained SDS-PAGE gel, indicating sufficient protein purity for further analysis. A native gel of the purified protein shows only protein of the expected size and no aggregation or cleavage of the expressed protein. Finally, the identity of the purified proteins was confirmed by mass spectrometry, performed by the CHiBi Proteomics core facility at UBC.

3.2.4.2 Protease inhibition assay

The inhibitory function of the three anion-exchange purified proteins against trypsin, α chymotrypsin, thrombin, subtilisin, and papain was assayed using the Pierce Fluorescent Protease Assay Kit (Thermo Scientific, Waltham, MA, USA), which provides fluoresceinlabeled casein (FITC-casein) as the fluorescent substrate. First, the ratio of protease to FITC- casein that produced the most linear response was determined for each of the five proteases. Fluorescence was measured at 485nm excitation and 538nm emission maxima every 30 seconds for 1 hour at 37°C. Next, a slow-binding kinetics experiment, using 1:1, 10:1, 100:1 serpin:protease ratios over 1 hour at ambient temperature (26°C) was performed (185). Fluorescence was measured in 30-second intervals. Since the *in situ* temperature that *C. symbiosum* serpins would function at in the sponge is lower than 26°C, the residual activity of the protease after incubation with serpin was measured, based on a method used in (186). To do this, first the serpin and protease were incubated for 1 hour at 20°C or 10°C, the expected functional temperature for *C. symbiosum* serpins, then FITC-casein was added and fluorescence was measured at 37°C. To confirm cleavage of the target by the protease, and determine whether the protease cleaves the serpin protein, an aliquot of each treatment was examined by SDS-PAGE. All experiments were performed in triplicate and included positive and negative controls and a BSA treatment.

3.2.4.3 Serpin-sponge lysate pull-down

A bait (serpin protein) and prey (sponge lysate) approach was used in a pulldown experiment using the PierceTM ProFound Pull-Down PolyHis kit (Thermo Scientific) to identify serpin-interacting proteins. Purified *C. symbiosum* serpin, CENSYa_0537, was dialysed with Tris buffer (25 mM, pH 7.2) and immobilized on a cobalt chelating resin packed column. After incubation, the column was washed to remove any unbound recombinant protein. Sponge lysate was prepared using the CellLyticTM lysis reagent. Freshly prepared sponge lysate protein was loaded onto the column with immobilized serpin. Following the incubation, the column was washed to remove and non-binding proteins. Wash fractions from each 54

step were collected and saved for downstream analysis. Trapped serpin-target covalent complexes were eluted with Imidazole and analysed by SDS-PAGE and gel silver-staining, and Western blotting using the anti-CENSYa_1605 polyclonal antibody. Two controls were used in this experiment: a column without immobilized serpin (resin only) and a column with immobilized serpin, but no lysate added as prey protein. Another approach to identify serpin-interacting partners in the sponge was to incubate recombinant serpin, CENSYa_1229, with sponge cell lysate. A constant amount of sponge lysate was incubated with no serpin, or 5 different concentrations of serpin overnight at 10°C, 15°C or room temperature. The fractions, along with serpin only, positive and negative controls were analysed by SDS-PAGE and Western blot.

3.2.4.4 Serpin protein and C-terminal peptide activity in NFKB reporter cell line

The effect of purified recombinant serpin, CENSYa_0537, and C-terminal peptides (Table 3.2) on NF κ B signalling was assayed in HEK-Blue reporter cell lines (a collection of engineered HEK293 cells), HEK-BlueTM-hTLR4 cells (InvivoGen, San Diego, CA, USA), and the control parental line HEK-BlueTM-Null2 cells (InvivoGen). The peptide sequences used are listed in Table 3.1. The experiments were performed in a laminar flow hood and included triplicates (Null2 cells) and five replicates (hTLR4 cells) of each of the following treatments: no treatment control, LPS only (1ng/ml), LPS + LipidIVA (10ng/ml), 4 concentrations of full-length serpin (5, 10, 50, 100 ng/ml) or 2 concentrations (10 or 50 µg/ml) of serpin C-terminal peptides, and serpin + LPS, for all serpin concentrations. Cells were grown in complete growth media, consisting of Dulbecco's modified eagle medium with 10% fetal bovine serum, 2 mM glutamax, 1 mM pyruvate, 100 U/mL penicillin, 100 µg/mL Streptomycin and 100 µg/mL normacin. Cell

type specific selection antibiotics were added to the media as follows: 100 µg/ml zeocin for the parental Null2 cells and 100 µg/ml zeocin, 200 µg/ml hygrogold and 30 µg/ml blasicidin. Cells were first detached from plate using Hank's based enzyme free cell dissociation buffer, then cells were collected into a 15 ml tube, with an aliquot reserved for counting, and centrifuges at 1,000 rpm for 3 minutes at room temperature. Cells were then resuspended in 1 ml complete growth media and selection antibiotics. Approximately 50,000 viable cells were seeded per well, in a 100-µl volume, in a 96 well tissue culture plate. Cells were grown for 48 hours in an incubator at 37°C and 5% CO₂, with an addition of 100 µl growth media to each well after the first day. Cells were visually inspected and washed with 100 µl pre-warmed growth media. Following the wash, cells were treated with stimulants diluted in growth media and placed back in the incubator for 24 hours. Following incubation, cells were visually observed, 20 µl supernatant from each well was transferred to a 96-well flat bottomed plate, and 180 µl QUANTI-BlueTM (InvivoGen) reagent was added to each well containing supernatant and gently mixed. Multiple absorbance readings were taken at 650 nm in regular intervals for 60 minutes.

Peptide	Peptide sequence	Charge	MW	Experiment(s)	[] range or amount used
CENSYa_ 1229	PFLFLIQDDESGTILFMGRVSEP	-3	2,612.0	PBMC and NFkB reporter cell stimulation	1.9-40 μM (5 - 100 μg/ml)
CENSYa_ 1605	PFLFLIQDDESGAVLFMGRVSEP	-3	2,567.9	PBMC and NFkB reporter cell stimulation	1.9 - 40 μM (5 - 100 μg/ml)
Human_A T1	PFVFLMIEQNTKSPLFMGKVVNPTQK	2	2,994.6	PBMC and NFkB reporter cell stimulation	1.6 - 33 μM (5 - 100 μg/ml)

Table 3.2 C. symbiosum serpin C-terminal peptide sequences.
Peptide	Peptide sequence	Charge	MW	Experiment(s)	[] range or amount used
CENSYa_ 1605WT	PQFKADRPFLFLIQDDESGAVLFMGRVSEP	-2	3,410.9	peptide pull- down	100 µg
CENSYa_ 1605mut	PQFKADRPALFLIQDDESGAVLAMGRVSEP	-2	3,258.7	peptide pull- down	100 µg

3.2.4.5 C-terminal peptide activity in human peripheral blood mononuclear cells

Inflammatory activity of serpin C-terminal peptides (Table 3.2) was assayed using human peripheral blood mononuclear cells (PBMCs) by measuring production of tumor necrosis factor α (TNF α), interleukin 6 (IL6) and monocyte chemotactic protein 1 (MCP1) by enzyme-linked immuno-sorbent assays (ELISA). PBMCs were isolated from blood collected by members of the Hancock lab from healthy volunteers, under UBC ethics approval and guidelines, into Vacutainer tubes (BD Bioscience, San Jose, CA, USA). Blood samples were diluted with an equal volume of phosphate buffered saline (PBS) (Invitrogen, Carlsbad, CA, USA) following collection, and cell types were separated by Ficoll gradient centrifugation. The layer containing mononuclear cells was carefully removed and washed twice with PBS. Following this, cells were resuspended in RPMI 1640 with 10% FBS, 2 mM L-glutamine, and 1 mM sodium pyruvate, and placed in a humidified incubator at 37°C with 5% CO₂. Cells were then seeded into 96-well or 48-well tissue culture plates, at a concentration of 1E6 cells/ml, and incubated for 1 hour. Cells were treated with 200, 100, 50, 20 or 5 µg/ml of CENSYa 1229 peptide, CENSYa 1605 peptide or human antitrypsin 1 C-terminus peptide, all the peptide conditions with an addition of 10 or 20 ng/ml LPS, LPS only treatments, or were untreated. Samples were then centrifuged at 1000 x g for 10 minutes to obtain cell-free supernatants, which were then stored at -20 °C. ELISAs were performed on supernatants collected 24 hours after treatment and were developed using TMB

Liquid Substrate System (Sigma-Aldrich) and measured with a Power Wave X340 plate-reader (Bio-Tek Instruments, Winooski, VT). Cytotoxicity of serpin C-terminal peptides was assessed using the lactose dehydrogenase (LDH) assay on cell-free supernatants following the protocol specified in the Cytotoxicity Detection kit (Roche).

3.2.4.6 C-terminal peptide pull-down in sponge and mammalian cells

To identify proteins with the potential to interact, directly or via a protein complex, with C-terminal serpin peptides, I performed a pull-down experiment using biotin-labeled peptides and whole cell protein lysate from D. mexicanum cells and two human cell lines. Two positions seem to be very important for the biological activity of C-terminal peptide (Dr. F. Jean, personal communication) so the peptides including those positions, and a corresponding double mutant peptide were used for this experiment (Table 3.2). Purity of synthesized peptides was assessed by an amino acid analysis at The Hospital for Sick Children (Toronto, ON, Canada). A few small, ~ 1 cm³, pieces of sponge tissue were rinsed with calcium and magnesium-free artificial seawater (CMFASW) three times. The mass of the starting material was noted. Following the washes, sponge tissue pieces were allowed to dissociate in CMFASW and 25 mM EDTA for 30 minutes at room temperature with gentle agitation. Then, the cell suspensions were filtered through a 70µm nylon mesh sieve to remove spicules and large undissociated pieces. Sponge cells were pelletized by centrifugation at 600 xg for 10 minutes at 4°C. The supernatant was collected and centrifuged at 10,000 xg to collect microbial cells, which were then placed at -80°C for storage. The sponge cell pellet was resuspended in CMFASW and cells were observed using light microscopy and quantified using the hemocytometer. The two human cell lines used were HuH7, a cell line of differentiated hepatocyte cellular carcinoma cells, and HEK293, a cell line derived

from human embryonic kidney cells. Whole-cell lysates were obtained by first harvesting $\sim 5 \times 10^6$ cells per sample. Then, a hypotonic buffer (10 mM Tris-HCl, pH 7.8) containing a complete protease inhibitor cocktail (Roche, Basel, Switzerland), phosphatase inhibitor, 0.1% Triton-X, 0.2 mM EDTA, and 2 mM dithiothreitol was added. The cells were incubated on ice for 30 mintues, with intermittent vortexing. After incubation, 300 mM KCl was added in 1:1 v/v ratio, and samples were centrifuged at 13,000 rpm for 30 minutes. The supernatants were collected and pre-cleared by rotating incubation at 4°C with avidin beads washed with 150 mM KCl. The beads were then removed by centrifugation at 500 x g for 30 seconds. Positive control peptides used in the Jean lab were used in this experiment. The biotinylated peptides were resuspended in PBS and conjugated to washed strepavidin beads to generate resin for the pull-down. Beads were washed to remove any unbound peptide. Then, freshly prepared and pre-cleared whole cell lysate from human or sponge cells was applied to the resin, and after incubation and washes, any interacting proteins were eluted by adding 2 bead volumes of 100 mM glycine, pH 2.8. Eluted proteins were precipitated with acetone and submitted to the Core Proteomics facility at UBC for LC-MS analysis. Proteins were injected into an SDS-PAGE gel for in-solution digest and processed on the LTG Orbitrap VELOS. Peptide pull-down, MASCOT database searches, and protein identification, control subtraction and interactome mapping were performed by Jean lab members.

3.3 Results

3.3.1 Comparative genomics between C. symbiosum and free-living Thaumarchaeota

Almost half of the predicted genes in the *C. symbiosum* "A" genome were not found in free-living thaumarchaea from diverse environments (Figure 3.1A). The majority of shared genes

likely represent core thaumarchaeal genes, as they are found in all the genomes examined (Figure 3.1A). These potential core thaumarchaeal genes represent ~31% of the C. symbiosum genome. Other than the core genes, C. symbiosum had few homologs in the genome of the moderately thermophilic N. gargensis Ga9.2 but had >380 homologs in the genomes of marine thaumarchaea and the rhizospheric N. koreensis MY1. Comparison of C. symbiosum with marine thaumarchaea revealed a similar pattern, where >900 C. symbiosum ORFs were unique to the symbiont, and most of the shared homologs were present in all datasets (Figure 3.1B). Approximately 45% of predicted C. symbiosum ORFs had homologs in N. maritimus SCM1 and N. limnia SFB1, which are both from temperate waters. C. symbiosum had the least number of non-core homologs with the thaumarchaea from Hawaii, and had the most homologs with its closest known free-living relative (~97% similarity across the SSU rRNA gene), N. maritimus (163). Over 740 of the genes unique to C. symbiosum are predicted to encode hypothetical proteins that have no known function or homologs. The unique genes appear to have a nonrandom distribution in the C. symbiosum genome, as multiple regions with >10 contiguous unique C. symbiosum genes were identified (Figure 3.2). The longest of these "unique" regions consisted of 94 predicted genes, including a DNA modification methylase, micrococcal nuclease-like protein, Flp pilus assembly protein, an ATPase and 89 hypothetical proteins. Additionally, 3 putative genomic islands, mostly containing genes encoding hypothetical proteins (Figure 3.2) were identified within three unique genomic regions.

Approximately 20% of the predicted *C. symbiosum*-encoded proteins were detected by proteomics. I identified 392 *C. symbiosum* proteins and another 227 *N. maritimus* proteins, all of which had homologs in *C. symbiosum* and likely represent products of core genes. Over 130 of the 392 expressed *C. symbiosum* proteins were unique to *C. symbiosum*, and 87 of these are

annotated as hypothetical proteins. Furthermore, ~40 of the unique expressed proteins were very long, exceeding 1,000 residues, and most of these were annotated as "hypothetical" in the genome sequence. Since many of the genes unique to *C. symbiosum* did not have a predicted function, functional homology approaches were used to identify conserved domains and help infer function. Assigning function to proteins of unknown function is an increasingly important field of research (187). Conserved domains were identified for 222 unique *C. symbiosum* genes, including 68 hypothetical proteins and 20 proteins that contained the well-conserved "domains of unknown function" (DUF). Moreover, functional structure prediction, PfamA searches and genome arrangement implicated a number of proteins unique to *C. symbiosum* in functions relevant to host colonization and signalling.



Figure 3.1 Comparative analysis between thaumarchaeal genomes.

A. Comparison of *C. symbiosum* to genomes of thaumarchaea from diverse environments. **B.** Comparison of *C. symbiosum* with genomes of each of the marine planktonic datasets used in part A.



Figure 3.2 Position of proteins of interest identified in this study on the C. symbiosum genome.

Nested circles represent from outermost to innermost, (i) gene positions, (ii) genes unique to *C. symbiosum* when compared to *N. maritimus*, *N. limnia* and HF4000 datasets, (iii) proteins detected by proteomics in a *D. mexicanum* sample (SB2), where the values are scaled based on normalized spectral abundance factor, peptides considered have a prophet peptide probability score > 0.95, (iv) a subset of the genes and genomic regions of interest. Inset: Legend.

3.3.2 C. symbiosum proteins implicated in host-microbe interactions

Although a few proteins that had homology to proteins that confer antimicrobial activity, antibiotic resistance and toxin production were identified, they had low (<90%) model homology. Other unique *C. symbiosum* genes include 2 proteins associated with the type IV (pilus assembly) and type II (hydrolase) secretion systems (Figure 3.2). Among the unique *C. symbiosum* genes are 5 serpin homologs, CENSYa_0537, CENSYa_1229, CENSYa_1605, CENSYa_1682 and CENSYa_1965 (51). However, only one of these (CENSYa_1605) was detected in the proteome. Interestingly, each of the five homologs was adjacent to at least one hypothetical unique protein (Figure 3.3). In one case, the predicted hypothetical protein CENSYa_1230 had homology to a protease domain. Additionally, there were 3 subtilisin-like protease homologs in *C. symbiosum* but not in the free-living thaumarchaea examined and a previously unrecognized trypsin-like serine protease was identified (Table 3.3).



Figure 3.3 Genomic arrangement of *C. symbiosum* serpins. Identity and genomic position of genes adjacent to predicted serpin-encoding genes.

Trypsin- like serine protease	Subtilisin -like serine protease	Surface layer- associated STABLE protease	periplasmic serine protease (ClpP class)	metal- dependent protease of the PAD1/JA B1 superfamily	membrane- associated Zn- dependent protease	secreted periplasmic Zn-dependent protease
CENSYa _1238	CENSYa _0382	CENSYa_ 2066	CENSYa_0375	CENSYa_1304	CENSYa_0056	CENSYa_1060
	CENSYa 0623		CENSYa_1165	CENSYa_0531	CENSYa_1688	CENSYa_1145
	CENSYa _0644					CENSYa_1580
						CENSYa_1791
						CENSYa_1874

Table 3.3 Proteases encoded by *C. symbiosum* as per original and updated annotation.

A genomic region containing glycosyltransferases not found in other thaumarchaea was identified, as well as sialidases involved in cell wall and membrane biogenesis (Figure 3.4), which could play a role in cell surface modifications. *C. symbiosum* encodes genes not found in other thaumarchaeota whose products have high (>99%) PHYRE2 confidence with eukaryotic proteins involved in cytoskeletal and innate immunity regulation, including 3 expressed proteins with homology to thioester - containing protein I (TEP) and 4 proteins with strong homology to actin interacting protein 1 (AIP-1) and Sro7 involved in endo- and exocytosis.



Figure 3.4 Putative cell surface modification operon encoded by *C. symbiosum.* Genomic arrangement of predicted genes encoding glycosyltransferases unique to *C. symbiosum.*



Figure 3.5 Protein sequence similarity network between *C. symbiosum* and reference metallo- β -lactamases. *C. symbiosum* proteins with lactonase or beta-lactamase Pfam hits are indicated as enlarged blue nodes. Edges between nodes are drawn if the sequence similarity by BLASTp was >50% with an e-value cutoff of 1E-10.

The *C. symbiosum* genome encodes 32 unique large proteins (556 - 11,910 a.a.), more than half of which are expressed and are predicted to form a β -propeller structure. This topology is found in a variety of enzyme families and can serve as a scaffold for transient multi-protein

complexes. Thirteen *C. symbiosum* β -propeller proteins contained domains associated with lactonases, which belong to the metallo- β -lactamase (m- β L) protein family. Because a number of these proteins were too large to model (> 1500 aa), a BLAST sequence similarity network was used to predict possible functions of these proteins (188). The sequences in network regions surrounding *C. symbiosum* m- β L proteins were microbe-encoded hydrolases with diverse substrates (Figure 3.5). Therefore, the annotation of these proteins is limited to broad superfamily membership. Although limited in resolution, these annotations provide insight into the possible function of a number of hypothetical proteins with potential antimicrobial or signalling roles.

3.3.3 Biological activity of C. symbiosum serpins

Serpins encoded by *C. symbiosum* share sequence similarity with secreted inhibitory serpins (Figure 3.6). To infer possible function of *C. symbiosum*-encoded serpins, the biological activity of these proteins was investigated using 3 recombinant proteins including CENSYa_0537, CENSYa_1229 and CENSYa_1605 purified from *P. pastoris* cultures. Serpin addition, regardless of serpin to protease ratio and serpin homolog identity, did not inhibit the cleavage, or decrease protease kinetics, of FITC-casein by trypsin, α -chymotrypsin, thrombin, subtilisin, or papain (Figure 3.7). Similarly, incubation of proteases with serpin prior to addition of substrate did not decrease activity protease activity. Although serpin genes could be quantified (~10⁹ copies/g sponge) by quantitative PCR in total DNA preparations, PCR performed on cDNA did not yield amplicons, suggesting that if serpins were expressed in the sponge, they were expressed at a very low level (Appendix B.2). Likewise, only 1 serpin homolog was detected in the proteome dataset and *C. symbiosum* serpins were not detected in the

								hA	
al-Antitrypsin Antithrombin III CENSYa_1965 CENSYa_1605 CENSYa_0537 CENSYa_1229 CENSYa_1622 M. mazei serpin I4 Serpin	MYSPGAGSGAAGER MRVI MRRA MPGGRPLRKA MNIK	SSVSWGILLAGLÓ KLCLLSLLIGALÓ YAAPILLLVPAA YLAAAVVIIVAAA YLAAAAVVIIVAAA YLAAAAVVIIVAAA YLAAAAVVTIAAAA MSRIKIILLACM SIVPFAVVMLACLI	CCLVPVSLA GCAICHGNPVDDICI GALQCGEEHVLTERPNGKI VAYSALSASVIQGGP VGYAAFPASDIQDAP VGYVUPASDTRDTP VAYSALSASVVQDAL LALLFSGCVENSA FC-VFAGC	ACVTDQTADRLGWMRVQDPC APVE	EDPQGDAAQK AKPRDIPVNP DPADRVSPDGLSVMHERQP VGPPDLASLP TGPPDLASLP TGPPDLASLS VFEKNTATLS TGTPGSDGTETATP	TDTSHHI LCIYRSPGKKATEED VBAPAL VVSASE VVSASE VVSASE VVSASE T	DODHPTFNKITPN SEQKVPEATNRR GGIGADARGMTDA IESAVIPREIADS IESAVIPQEIADS IESAVIPQEIADS VADSVRGYDIASA SHTGSAVESVVDA	** * - LAEFAFSLYR VWELSKANSFAGDFYA DARGAAONRSFAGDFYA NNGFAADFYR NNGFAADFYR NNGFAADFYR NNGFAADFYR NNAFAFDMYS NNMFAFEIYK	63 96 111 75 75 88 81 62 67
al-Antitrypsin Antithrombin III CENSYa_1965 CENSYa_1605 CENSYa_0537 CENSYa_1229 CENSYa_1682 M. mazei serpin I4 Serpin	S6B QLAHQSNS-TNIFI HLADSKNDN-DNIFI QVSDIQGNIFI QVSGEEGNIFI QVSGEEGNIFI QVSGEEGNIFI QVSGECGNIFI QLTQQVTGDYENILI QLSGENSKD-DNLFI	hB SPUSIATAFAMIS SPUSIATAFAMIS SPUSIATAFANIS SPUSMYTAFSLLY SPUSMYTAFSLLY SPUSMYTAFSLLY SPYSISAAMAICY LSPFSISSALALTY	hC :::::: KLGTKADTHDEILEGLNFN KLGACNDTLKQLMEVFKFT (EGAEGDTAAQLRDAFGLF (EGASGEAKELEDAFGFF (EGAGGEGAKELEDAFGFF (EGAGGEGAKELEDAFGFF (EGAEGTAQISSVFYFF (EGAKGETADQISSVFYFF	hD : LTEIPEAQI-HEGFQELLRT TISEKTSDQIHFFFAKLNCR PDDSARHASVKLLLES PDARARHNATAHAIAS PDAQARHNTAHTIAS PDAQARHNSTAHAMAS TNKTVLKVRLERINDR DNIETLRSGYQEVNAG	s2A 	hE SEGLKLVDKFLEDVKF SKSLTFNESYQDVSEY ADWIEAKEEYVGAVRI FVG-DVVPEYLNVAR FVG-NVRPEYLDVGRS FVG-SS-AINPEYLDVGRS SS-AINPEYLDVGRS SEGYPVKEEYLFNVKF SETYPFLKEYINTAKT	slA VYGAKLQPLDFK VYGAKLQPLDFK VYRSDVERIDF	hF .: *:: DTEEAKKQINDY ENEQSRVTINNW ESAENRGDEPARINOW GDDPMRINOW GDPIDRINGW GDNPVDRINGW RKPDDSRNTINEW NQPEESRVSINDW	184 221 223 196 185 198 191 183 187
al-Antitrypsin Antithrombin III CENSYa_1965 CENSYa_1605 CENSYa_0537 CENSYa_1682 M. mazei serpin 14 Serpin	VEKGTQGKIVDLVK VANKTEGRIKDVIP ASDNTRGRITDVTE TSDMTNGTITGIFG VSDATNEKIPEVID VSDATNEKISEVVG VETRTSDKIKDLVP VAEKTNDKIEDLIP	ELDRDTVFALV QGAINELTALVLV QDQFDDLTAAVIT PNSFP-DPILVIV PDSFA-DPILAIV ELNPYTVLVIT ELDSDTVLVIT EGMIDSMTRLVIT	S3A * ::*** * ::* NYIFFKGKWERFFEVKDT NTIFFKGKWERFFEVKDT NAVYFKGTWETQFPPEDT NAVYFKGTWETQFPPEDT NAIYFKGTWVKQFLPEHT NAIYFKGTWVKQFLPEHT NAIYFKGTWVLQFDKNMT	S4C S3C EEEDFHVDQVTTVKVPMMKRI REPFYRVDGQSCPVFMMYDI SKNDWHGEGDAEADYMNV KMDEWHGEGDAEADYMNV ENEWHGPSGSAEYMYAI ELDEWHGPSGSAEYMYAI EKRAFYPANGEEASVDMYM FEADFTTPSGETVTVEMMKR	s1B CGMFNIQH-CKKLSSWV CGKFKYRRVAEGTQV VGAFDYSE-RGAQV VGIFDYAE-YDGLQV VGIFDYEE-YDGLQV KDRFSYAE-YDGLQV YNNFNYGE-NSKAK-I FDDDAVYGYTE-TDGLQV	S2B : :: : : : : : : LLMKYLGNATA LELPFKGDDITM LRLPYEGGRISM LRMPYEGDRLSM LRMPYEGRNLSM IELPYKGNLSM IELPYKGNLSM	S3B IFFLP-DEGKLQH VIILPGFKSIAK VVILP-BPGGMKE TVVLP-BPGGMKE VVLP-EPGGMRE VVLP-EPGGMSN VVLP-KENDLKA	hG hH LLENELTHDIITKFLENE VEQELTPELQENLDEL LEQSVVHDGIPAWTSGM LEESISAEQMAVWTGGT LLEESISAEQMAVWTGGT LLEESISAEQMLEWTGGL LEESISAEQMLEWTGGL EFEFFINQYTELKNEM LAEDALSADNFEEIENSI	304 343 344 313 302 314 307 303 313

RCL

Figure 3.6 Multiple sequence alignment of predicted *C. symbiosum* serpins and reference homologs.



Figure 3.7 Serpin inhibition of protease activity assays. The representative result of an inhibition assay, in this case CENSYa_0537 serpin and subtilisin, is shown, as all protease-serpin interactions tested had a similar pattern.

lysate by Western blot or silver-staining the SDS-PAGE gel, indicating that these proteins were not expressed at a high level at time of sponge tissue preservation. To identify potential serpin targets expressed by the *D. mexicanum* holobiont, poly-histidine pull-down and co-incubation experiments using sponge lysate and CENSYa_0537 and CENSYa_1229 respectively were done. No SDS-stable serpin-protease complexes were detected by SDS-PAGE or Western blot with either approach (data not shown). Further, no new bands were observed between controls and elution of serpin-lysate incubation in the pull-down experiment.

Since no readily available sponge system exists, heterologous mammalian systems were used to study the activity of recombinant *C. symbiosum* serpins. To test whether *C. symbiosum*

serpins affect innate immune signalling, a reporter cell line expressing secreted embryonic alkaline phosphatase under the control of an IL-12 p40 minimal promoter and fused to binding sites of NF κ B and AP-1 transcription factors, and the recombinant serpin CENSYa_0537 were used. Treatment of the reported cells with the recombinantly-expressed serpin had no effect on LPS-stimulated signalling through TLR4 and did not stimulate AP-1 or NF κ B driven responses (Figure 3.8A).

Hydrolysis of the P1-P1' serpin scissile bond by the target protease results in the release of a C terminus peptide that has been shown to play roles in viral infection (189, 190). Thus, I tested synthetic peptides with sequences corresponding to serpin C termini in the reporter cell line. As observed with recombinant serpin, treating cells with these peptides did not affect TLR4 signalling or signalling through NF κ B and AP-1 transcription (Figure 3.8B). To test whether these peptides have immunomodulatory activity not necessarily dependent on TLR signalling I measured cytokine and chemokine production in human primary cells in response to peptide stimulation. Under one set of experimental conditions, these C-terminal peptides affect the amount of TNF α and IL6 produced in response to LPS stimulation (Figure 3.9A and B). Furthermore, serpin C-terminal peptides, in combination with LPS, have a weak additive effect on induction of the chemokine MCP-1 (Figure 3.9C). However, a similar effect was not observed in the 96-well experimental setup, where no effect on TNF α , MCP-a or IL6 levels was observed.



Figure 3.8 Effect on TLR signalling in NFKB reporter cells stimulated by *C. symbiosum* serpin protein and C-terminal peptides.

Results are shown as the mean (+ standard error of the mean) of two independent experiments. Statistical comparisons between LPS-stimulated cells either treated or not treated with a peptide or serpin were evaluated by a two-tailed Student's *t* test, * p < 0.05.



Figure 3.9 Cytokine and chemokine production in PBMCs stimulated with C-terminal peptides. Results are shown as the mean (+ standard error of the mean) of two independent experiments, and are shown for stimulation with CENSYa_1605 peptide. These results are representative of CENSYa_1229 stimulation, as well as experiments using 10 ng/ml LPS. Statistical comparisons between LPS-stimulated cells either treated or not treated with a peptide were evaluated by a two-tailed Student's *t* test, * p < 0.05, ** p < 0.01.

To further explore potential immunomodulatory roles of C-terminal serpin peptides a series of incubation studies were performed on mammalian cell lines. Specifically, the serpin CENSYa_1605 C-terminal peptide appears to interact with mammalian proteins in a non cell-

type specific manner, since 40% of the proteins pulled down by the WT peptide were the same in both mammalian cell lines. Approximately 70 proteins from each of the two human cell types used were pulled down by the WT peptide (Figure 3.10, Appendix B.3). The double FA substitution in the peptide sequence resulted in a significant loss of proteins pulled down by the peptide, with 31 proteins pulled down for Huh7.5.1 and only 16 proteins pulled down for HEK293. Furthermore, the mutant peptide pulled down a different set of proteins than the WT peptide, with just 1 protein, a heat shock protein, pulled down with both peptides in both cell types (Figure 3.10). The majority of proteins pulled down are localized to the cytoplasm and mitochondrion. Additionally, WT peptide, but not mutant, interacting proteins also localize to the endoplasmic reticulum and nucleus. Based on protein identities, the proteins pulled down likely function in translation, protein targeting, viral cycle and metabolism. In contrast, extremely few proteins from D. mexicanum cells were pulled down by WT or mutant peptides, even with additional database searches using a custom sponge database that included D. mexicanum sequences. Three sponge proteins, including a hypothetical protein, an F0F1 –type ATP synthase and a heat-shock protein were pulled down with the WT peptide. The mutant peptide pulled down 2 proteins including tubulin and actin gamma2.



Figure 3.10 Proteins pulled down by a synthetic CENSYa_1605 peptide in human cells lines. The number of proteins pulled down in each treatment is in brackets; mut and WT refer to the peptide sequence.

3.4 Discussion

3.4.1 Putative C. symbiosum symbiosis factors

Comparative analysis of the *C. symbiosum* genome suggests that the symbiont carries genetic adaptations for a symbiotic lifestyle. Even though, as with most microbial genomes, most of the predicted ORFs in the *C. symbiosum* are predicted to encode hypothetical proteins of known function, structure and domain homology modeling led to the identification of conserved domains suggesting function (164, 191, 192). A number of possible symbiosis factors with functions relevant to host colonization and signalling were expressed by this archaeal symbiont. The *C. symbiosum* gene with homology to Flp pilus likely encodes archaeal flagellin, which is similar to bacterial Type IV pili and machinery of the Type II secretion systems rather than bacterial flagellin (193, 194). Colonizing microbes need to be able to adhere to host cells. Many adhesins are found in host-associated bacteria, including fimbrae, or pili, and flagella (195, 196). Pili are important for efficient symbiosis of *Sinorhizobium meliloti* with its plant host and are important virulence factors of plant and animal bacterial pathogens (197-200). The presence of these proteins in *C. symbiosum* indicates either previously unrecognized motility in *C. symbiosum* or, alternatively, an ability to adhere to host cells using Flp pili.

Evasion of the sponge innate immune response and phagocytosis are likely important strategies, given *C. symbiosum*'s localization in the mesohyl (130). It is not yet clear whether sponge symbionts are specifically recognized and not ingested by sponge cells or if symbionts actively conceal themselves from detection by host cells (201). Bacterial symbionts in sponges encode proteins containing ankyrin repeats that may help bacteria inhibit phagocytosis by host cells, indicating that the microbes may modulate host behaviour (50, 148, 201, 202). However, no ankyrin repeat proteins were encoded or expressed by *C. symbiosum*, thus this archaeal

symbiont may employ a different strategy to protect itself. A possible mechanism possibly used by *C. symbiosum* to mask its surface is through cell surface modifications or extracellular factors, such as a polysaccharide capsule encoded by the operon of putative glycosyltransferases not shared with free-living thaumarchaea. A strategy adopted by pathogenic, and also presumably symbiotic, microbes is to modulate host cytoskeletal system, particularly through interactions with actin filaments for motility and phagocytosis evasion (195, 203, 204). *C. symbiosum* β propeller proteins with homology to AIP-1 suggest that *C. symbiosum* may directly interact with host cytosekeletal components for symbiosis, possibly evading uptake and digestion as AIP-1 disruption disrupts phagocytosis (205).

A role for tetratricopeptide repeat (TPR) proteins previously implicated in bacterial virulence has been proposed in sponge-bacterial interactions on the basis of their presence in the genomes of symbiotic Poribacteria and δ -proteobacteria (54, 61, 206). However, the taxonomic distribution of these TPR encoding genes suggests that they are not unique to sponge-associated microbes, as the genome of the sponge cynanobacterial symbiont *Synechococcus spongiarum* contained less TPR proteins than free-living cyanobacteria (207). Similarly, while *C. symbiosum* encodes for 8 TPR proteins only 1 is unique to *C. symbiosum* (51), suggesting that TPR proteins may not be integral to sponge-archaeal symbiosis. Conversely, a different group of protein-protein complex forming proteins, the beta propeller proteins, may have important role in *C. symbiosum* – *D. mexicanum* symbiosis. There were 23 *C. symbiosum* β -propeller proteins may act in quorum quenching, particularly for systems with acylhomoserine lactone autoinducer molecules, limiting populations of symbiotic or pathogenic bacteria in the sponge (208, 209). Additionally, a lactonase homologue from a thermophilic archaeon has

detoxification activity (210). The m- β L protein superfamily can hydrolyse β -lactam antibiotics, and includes various enzymes including zinc hydrolases, involved in many functions including detoxification (211, 212). Thus, *C. symbiosum* homologs may confer toxin for *C. symbiosum* and by extension the sponge host.

C. symbiosum expresses proteins with homology to thiosester bond-containing TEP1 or A2M proteins which also share structural homology with the complement component C3 (213, 214). In metazoans, proteins belonging to the TEP family, including the broad range protease inhibitor A2M and C3, are involved in immune responses (215-217). In arthropods, TEP acts in a complement-like system, where it is essential for sequestration and phagocytosis of microbes (213, 218). Bacterial A2M proteins appear to offer protection from proteolytic degradation, either through direct inhibition of a host protease, or by complex formation with a host A2M (219-221). The Streptococcus pyogenes A2M-like surface protein can bind the human A2M, which can inhibit host and S. pyogenes proteases, thus helping protect the cell surface of the pathogen (221). Although A2M-like proteins are widely distributed in bacteria, where they are the most abundant protease inhibitors, their evolutionary origins remain uncertain (219, 220, 222). Furthermore, archaeal forms of A2M-like proteins are rare, with only 1 homolog from a cold-adapted euryarchaeote described thus far (104, 222). The rare distribution of archaeal A2M proteins suggests horizontal acquisition of these proteins. Thus, it is possible that C. symbiosum TEP/A2M-like proteins help the symbiont colonize sponge tissue and either form complex with a host A2M protein or inhibit host proteases.

Inhibition of host proteases is an important immonomodulatory mechanism, and the *C. symbiosum* genome also contains homologs of the serpin superfamily of protease inhibitors. Serine proteases are ubiquitous in eukaryotic and prokaryotic organisms and mediate a multitude

of physiological and developmental processes including innate immune signalling (223). Most characterized serpins are from eukaryotes, rare in prokaryotes and are absent in free-living thaumarchaea (51, 222, 224, 225). Since the function of archaeal serpins is not well understood and as previously mentioned, serpins are important in host-microbe interactions, *C. symbiosum* serpins may be important symbiosis factors. Serpins comprise a superfamily of proteins from all domains of life with a conserved tertiary structure but diverse functions (226, 227). Most serpins inhibit serine proteases, although some serpins inhibit cysteine proteases (228). Still other serpins do not have an inhibitory role despite sharing a conserved structure with inhibitory serpins (228, 229).

3.4.2 Biological activity of C. symbiosum serpins

Serpin expression in *C. symbiosum* is likely regulated, since only 1 homolog was detected at a low level in the proteome, and no serpin expression was detected using other approaches. In members of the human microbiota symbionts, *Bifidobacteria*, serpin expression is induced in response to certain proteases (83). Serpin expression in *B. breve* UCC2003 is regulated by twocomponent regulatory system, which is adjacent to the operon encoding the serpin and a hypothetical membrane-associated protein gene (230). Other strains of this species had a similar genomic arrangement (230). It will be of interest to determine whether ORFs encoding hypothetical proteins adjacent to *C. symbiosum* serpins affect serpin expression or function.

Conventional biochemical approaches to characterize the inhibitory activity of *C*. *symbiosum* serpins failed to identify target proteases and no interaction was found between recombinant *C. symbiosum* serpins and a protein in the sponge lysate. It is possible that *C. symbiosum* serpins (i) are non-inhibitory, (ii) inhibit an untested class of proteases, (iii) require a

co-factor for activity, or that non-optimal ratios of serpin to protease were used (228, 231). Further, the lack of interaction between serpins and sponge lystae could be due to either low concentration of the target protease in the lysate, or absence of an interaction between the serpin and a protease. It is also possible that the serpin-protease interaction involves a rapid cleavage event, thus preventing the detection of a higher-order complex. Another possibility may be that any serpin post-translational processing required for activity did not occur because the cognate accessory proteins were not present in *P. pastoris*. Thus, the results of serpin protease inhibition are inconclusive and warrant further investigation.

Neither the recombinant serpin nor the synthesized C-terminal peptides had an effect on LPS-stimulated TLR4 signalling. The mammalian cells used expressed TLR3, TLR5 and NOD1 in addition to TLR4. Since NF κ B and AP-1 can be activated by a variety of stimuli, HEK293-hTLR4 cells can also be stimulated in a TLR4-independent manner. The results indicate that *C. symbiosum* serpins do not affect NF κ B or AP-1 signalling under the conditions tested. However, synthetic peptides for *C. symbiosum* serpins appear to promote an immunoprotective response in human PBMCs under certain experimental conditions, as 1605 and 1229 peptides suppress expression of the pro-inflammatory cytokines TNF α and IL6 and induce MCP-1, which regulates monocyte and macrophage migration to infection site (232-234). Consistent with these observations, serpins from Bifidobacteria limit damage and cell death in host tissues by inhibiting proteases involved in inflammation (78).

The identities of human proteins pulled down suggest internalization of the peptide and interaction of CENSYa_1605 derived peptide with host proteins heavily involved in viral cycle. These observations suggest that archaeal serpin C-terminal peptides may act similarly to human serpin peptides, such as serpin A1, which is internalized by surface proteins and scavenger

77

receptors, and can inhibit HIV-1 entry, replication and promoter activity (189, 190). Since different sets of proteins were pulled down by WT and mutant peptides, it seems plausible that the WT CENSYa 1605 C terminus peptide has bona fide biological activity. Indeed, work with synthetic peptides based on regions of serpin A1 C terminus identified sequences important to its activities, including a putative internalization signal pentapeptide, that allows the peptide to be transported to the nucleus after interaction with the cell membrane (235). Since the synthetic C. symbiosum serpin-based WT peptide included this putative internalization signal and sequence important for affecting virus promoter activity, and protein interactions that may be involved in these processes were observed, translocation to the nucleus and antiviral properties are likely evolutionarily conserved serpin properties (235). It was surprising that very few sponge proteins were pulled down by the WT peptides. However, since the database used to identify sponge proteins was based on transcriptomic data, it is possible that the interacting proteins were missing from the database. Otherwise, the cell fraction used for lysate preparation did not include the target proteins. Beyond their antiviral activities, serpin C-terminal peptides have also been shown to affect proliferation, alter adhesion and migration of endothelial cells, and even induce apoptosis in epithelial cell lines (236, 237). It would be of interest to explore the potential involvement of C. symbiosum serpins in these contexts as well.

Although the ability of *C. symbiosum*-encoded serpins to modulate metazoan immune responses, and thus their potential role in symbiosis, remains ambiguous, I show that these proteins interact with eukaryotic cells. This work is the first attempt to functionally characterize archaeal serpins and their C-terminal peptides, providing important insight into functional conservation and diversity of this protein superfamily. Since *C. symbiosum* encodes for a number of proteases, including serine proteases, it is possible that serpins could interact with endogenous

proteins to regulate an internal process. Thus, the molecular tools and methods presented here could be used to clone, express and test interactions between *C. symbiosum* serine proteases and serpins. In addition to identifying serpin homologs, comparative genomic analysis of *C. symbiosum* implicated a number of unique genes in possible interactions with its sponge host *D. mexicanum*, allowing the identification, modeling, and biochemical characterization of putative symbiosis factors. These observations provide a robust framework for inferring host-microbe interactions mediated by *C. symbiosum*, and open the door for more in-depth studies focused on specific signalling and recognition processes supporting stable symbioses.

Chapter 4: Microbial recognition and host defense systems in marine sponges

4.1 Introduction

The ability to distinguish self from non-self is foundational to the evolution of multicellular organisms and a harbinger of innate immune signalling pathways. Innate immunity is an evolutionarily conserved system that detects non-self organisms and provides a rapid first line of defense against invading pathogens (73, 128, 238). With respect to microbial detection, the innate immune response depends on the recognition of conserved structural features, termed microbial associated molecular patterns (MAMPs), by germline-encoded pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs), Nod-like receptors (NLRs), mannose receptors and C-type lectins (72-74). Bacterial and fungal MAMPs are well characterized and include essential, often highly expressed components such as flagellin, peptidoglycan, lipopolysaccharides, β -glucan, and lipoproteins (73, 75, 76). However, MAMPs are also present in nonpathogenic microbes. Thus, the innate immune system must differentiate between epitopes derived from pathogenic and symbiotic microbes.

Metazoan evolution is intimately associated with microbial interactions and symbioses, resulting in diversification and niche expansion (1, 2). The long history of co-evolution is evident in physiological, developmental and genomic dependencies between eukaryotic hosts and their specific core microbiota (1-5). Recognition between interacting partners is essential for establishing and successfully maintaining interspecies associations, as shifts in microbial composition can have detrimental effects on the host (13). Animal-microbial symbioses are defined by both host and symbiont factors. Indeed, the host immune response helps structure the microbiota by exerting a strong selective pressure on microbial species composition (39). Since

host cells, pathogens and symbiotic microbes can be in close proximity, as in the mammalian intestine and within sponge tissues, it is imperative that the host immune system be able to maintain a balance between pathogen response and symbiont maintenance (21, 39). Indeed, PRR signalling has been implicated in protection of host tissues and maintenance of gut homeostasis in mammals and is necessary for long-term microbial colonization by microbiota, indicating that innate immune signalling pathways evolved under selection from both pathogen invasion and symbiotic communication (39, 94, 239).

As the deepest-branching animals, sponges offer a deep time perspective on animal evolution, phylogeny and animal-microbe interactions (43-45, 240, 241). Genomic and transcriptomic sequence information exists for several sponge species, enabling partial reconstruction of the earliest innate immune signalling (summarized in Table 4.1) (44, 45, 107-115, 240, 242, 243). Toll-like receptors and components of the TLR signalling pathway have been identified in the sponges Suberites domuncula and Amphimedon queenslandica. Sponge TLRs are considered non-canonical and perhaps represent an ancestral form of the protein family (45, 108, 109). Other PRRs including LPS-binding protein and multiple scavenger-receptor cysteine-rich (SRCR) proteins have also been identified in sponges. Notably, the A. queenslandica genome contains 135 NLRs (124). This is in line with observations in other metazoans, and presents the possibility of wider microbial recognition by these PRRs through domain shuffling and rearrangement (124, 244-246). However, whether this NLR diversity is present in other sponge species and represents the ancestral state is not known. Furthermore, there is little information on the conservation of other responses and their possible roles in immunity across Porifera and Metazoa. The symbiosis between the sponge Dragmacidon mexicanum and Cenarchaeum symbiosum presents an opportunity to examine conserved pattern

recognition pathways as it is highly specific (129), selection of the symbiont by the host is implicated, and the microbial community is known (Chapter 2).

Here I describe gene expression profiles for two sponges species, including five *D. mexicanum* and one *Tethya californiana* individuals, to shed light on microbial recognition by sponges and help elucidate the molecular basis of *D. mexicanum - C. symbiosum* symbiosis. This is the first study to examine transcriptomes of sponge individuals whose microbiota are characterized (Chapter 2), and the only effort to compare multiple conspecific sponge individuals at the same developmental stage. Pathways and conserved domains associated with innate immune signalling pathways, from microbial recognition, inflammation, to clearance of infected cells, are identified and compared across the animal kingdom to delineate evolutionarilyconserved and divergent mechanisms of host defense and symbiont recognition.

Sponge species	Data type	Number of reads	Assembly size (Mbp)	References
Dragmacidon mexicanum	adult RNA, non-pooled replicates	752,999,502 Illumina reads	647, ~130/replicate	this study
Tethya californiana	adult RNA	159,593,058 Illumina reads	108	this study
Amphimedon queenslandica	larval and adult RNA	237,000,000 SOLiD reads		(110)
Amphimedon queenslandica	larval and embryo DNA	2,920,000 reads	167	(45, 111)
Aphrocallistes vastus	adult RNA*	78,150,000 Illumina reads	65	(115)
Chondrilla nucula	adult RNA*	159,450,000 Illumina reads	29	(115)
Cliona varians	adult and explant RNA*	122,504,240 Illumina reads	89	(247)
Corticium candelabrum	adult RNA*	96,670,000 Illumina reads	65	(115)
Crella elegans	adult (different reproductive stages) RNA	124,881,683 Illumina reads	91	(113)

Table 4.1 Summary of sponge sequence data

Sponge species	Sponge species Data type		Assembly size (Mbp)	References
Crella elegans	adult RNA*	25,951,906 Illumina reads	27	(114)
Ircinia fasciculate	adult RNA*	60,900,000 Illumina reads	17	(115)
Oscarella carmela	larval, adult, embryonic RNA	11,520 ESTs	9	(44)
Petrosia ficiformis	adult RNA*	64,100,000 Illumina reads	40 (2 assemblies)	(114, 115)
Pseudospongosorites suberitoides	adult RNA*	89,050,000 Illumina reads	28	(115)
Spongilla lacustris	adult RNA*	115,070,000 Illumina reads	48	(115)
Suberites domuncula	adult RNA	13,694 ESTs + individual clones	10	(106, 108, 248, 249)
Sycon coactum	adult RNA*	64,810,000 Illumina reads	23	(115)

* no polyA selection, no ORF prediction

4.2 Materials and methods

4.2.1 RNA isolation and purification

Transcriptome profiles were generated for a subset of the sponges used in the holobiont characterization study (Chapter 2), and included five adult *D. mexicanum* and one adult *T. californiana* sponges. Sponge mRNA was extracted from subsamples of homogenized frozen sponge tissue described in Chapter 2 using RNABee reagent (AMSBIO, Milton Park, Abingdon, UK). The homogenization and chloroform steps were performed in a clean fume hood. First, 1ml RNABee was added per 50 mg frozen sponge tissue and was immediately homogenized thoroughly using a glass homogenizer until no tissue pieces were observed. Following homogenization, 200 μ l chloroform was added to the homogenized tissue for every 1 ml RNABee used. The tubes were then shaken for 30 seconds, incubated on ice for 10 minutes and then centrifuged for 15 minutes at 12,000xg at 4°C to separate the phases. Following

centrifugation, the aqueous layer was transferred to a new tube, and the chloroform extraction step was repeated until the interface was clear. An equal volume of isopropanol was added to the aqueous phase and incubated at room temperature for 10 minutes to precipitate the RNA. After incubation, the tubes were centrifuged for 5 minutes at 12,000xg at 4°C, and the pellets were then washed with 75% ethanol and centrifuged at 7,500xg for 5 minutes at 4°C. After centrifugation, the supernatant was removed and the RNA pellets were air-dried, but not allowed to over-dry. The RNA pellet was dissolved in RNase-free water. Subsamples of prepared RNA were used to assess RNA quality by visualizing the extract on a MOPS-formaldehyde gel and taking A_{260/280} readings using a NanoDrop spectrophotometer (NanoDrop Technologies, Inc, Thermo Scientific, Waltham, MA, USA), and to measure the total RNA concentration using Ribogreen reagent (InvitrogenTM, Life Technologies, Carlsbad, CA, USA). To store the remaining RNA, 3 volumes 100% ethanol and 0.1 volume 5 M NH₄OAc or 3 M NaOAc were added and then were placed at -80°C until further use. Prior to use, the ethanol-precipitated RNA was washed with 70% ethanol three times, any residual ethanol was removed, and the RNA pellet was allowed to air dry. The pellet was then dissolved in RNase-free water and 1/40 vol/vol RNase Inhibitor (Ambion®, Life Technologies, Austin, TX, USA) was added to prevent RNA degradation. Any contaminating genomic DNA was digested using TURBOTM DNase enzyme (Ambion®, Life Technologies, Austin, TX, USA). The DNase treatment consisted of an addition of 1 µl DNase and 0.1 volume DNase buffer (Ambion[®], Life Technologies, Austin, TX, USA), incubation at 37°C for 30 minutes, followed by another addition of 1 µl DNase and a further 30minute incubation. To stop the digestion, 5µl DNase Inactivation reagent (Ambion®, Life Technologies, Austin, TX, USA) was added to the tube, incubated at room temperature for 5 minutes, and centrifuged at 10,000xg for 1.5 minutes. Purified RNA was transferred to a fresh

tube and more RNase Inhibitor was added. Aliquots were taken for quantification, reverse transcription and formaldehyde gel visualization. Reverse transcription PCR targeting sponge housekeeping genes was used to confirm absence of DNA contamination.

4.2.2 Sponge housekeeping genes PCR protocol

To test RNA extraction, purification and reverse transcription protocols, PCR assays targeting sponge housekeeping genes were developed. Sponge actin and tubulin sequences were identified on fosmid ends that were part of the original fosmid library generated for C. symbiosum genome assembly as follows (51, 135). A small proportion of the fosmids in the library were assigned to Eukayota by MEGAN taxonomic analysis (250). Open reading frames (ORFs) were predicted on these fosmid end sequences using FGENESH (SoftBerry), optimized for eukaryotes. A BLASTx search of the predicted ORFs was performed against the nr database to annotate the putative genes. Predicted actin and β -tubulin sequences were aligned to homologs from a number of animal species using MUSCLE multiple sequence aligner (251). Primers targeting conserved regions were designed and tested. Actin sequences were amplified by PCR (5'using the primers Actin F (5'-ATCCAGACGAAGGATGG) and Actin R2 ATCACACTTTCTACAACGAG) under the following PCR conditions: 3 minutes at 95°C, 36 cycles of denaturation at 95°C for 40 seconds, annealing at 57°C for 45 seconds, and extension at 72°C for 1 minute, followed by a final 10 minute extension at 72°C. Tubulin sequences were amplified using the TUBB F2 (5'-CCAGCAGATGTTTGATGCC) and TUBB R2 (5'-TGCCTTCACCAGTGTACC) under the following PCR conditions: 3 minutes at 95°C, 36 cycles of denaturation at 95°C for 40 seconds, annealing at 62°C for 45 seconds, and extension at 72°C for 1 minute, followed by a final 10 minute extension at 72°C. Each 25 µl reaction

contained 2 μ l of template cDNA, 500 nM each forward and reverse primer, 1 mM deoxynucleotides, (Bioshop, Burlington, ON, Canada), 1.5 mM MgCl₂ and the BioShop PCR buffer at 1× concentration. Additionally, SSU rRNA was amplified using the same primer set and protocol used to amplify the SSU rDNA described in Chapter 2.

4.2.3 cDNA library production, sequencing and assembly

Total RNA was submitted to the Genome Sciences Center (GSC) (Vancouver, BC, Canada) for polyA-selection, cDNA library construction, sequencing and *de novo* raw read assembly. RNA integrity for each sample was tested again at the GSC using the Agilent Bioanalyzer system prior to mRNA purification and plate-based RNA-seq library production. Libraries were sequenced using 1 lane, 50 base PET per sponge on the Illumina HiSeq platform. Shotgun transcriptome sequences were assembled into contigs using the *de novo* assembler ABySS 1.2.7 (252, 253), using default parameters for all k-mer sizes between 26 and 48. The multi-k assemblies were filtered and merged to generate non-redundant contigs.

4.2.4 Transcriptome annotation

Following assembly, transcriptome contigs were filtered, assigned taxonomy and annotated using a combination of BLAST-based and BLAST-independent sequence homology approaches. Raw assemblies were analysed using MetaPathways, a gene prediction and annotation pipeline (254). First, contig sequences were filtered to remove sequences below 180 bp length threshold and any sequences with incompletely specified bases. Only sequences passing the above criteria were used for ORF prediction. The MetaPathways pipeline uses Prodigal for calling and translating ORFs (255). Since this algorithm was originally designed to

86

predict prokaryotic rather than eukaryotic sequences, a subset of randomly selected contigs were used for gene prediction using FGENESH. The two ORF finding methods had equivalent performance on this dataset. Nucleotide sequences corresponding to ORFs were conceptually translated, and resulting amino acid sequences were queried against KEGG release 53 (256), COG (accessed in 2007) (257), RefSeq release 56 (258), and InnateDB downloaded in July 2010 (human and mouse gene lists) (259) databases using the BLASTp algorithm. The minimum bitscore and maximum E-value were set at 50 and 1E-6, respectively. A consensus annotation for each ORF was determined by the top hit for each database with a minimum bit-score ratio of 0.4 (260). The numbers of predicted ORFs with a BLAST hit and consensus annotation are summarized in Table 4.3. To confirm annotations obtained based on predicted ORFs, a BLASTx search, which is not dependent on a pre-defined ORF, was performed for a subsample of contig sequences. The similarity of BLASTx results to those of BLASTp increases confidence in the accuracy of the predicted ORFs. The taxonomic assignments of predicted protein-coding genes were determined using MEGAN based on the RefSeq BLASTp output (250). MEGAN utilizes NCBI taxonomy and a lowest common ancestor algorithm to assign taxonomy to each sequence (250).

Fragment recruitment was performed to identify any *C. symbiosum* transcripts. To this end, contigs were mapped to *C. symbiosum* genome using bwa fragment recruitment. A small proportion of each set of *D. mexicanum* contigs (< 1%; 98-7,000) mapped to the *C. symbiosum* genome, primarily to the 23S and 16S rRNA sequences. In the *T. californiana* sponge, 7 contigs mapped to the *C. symbiosum* genome, all of these sequences were very short ~60bp, and 6 of those mapped to the LSU and SSU rRNA genes.

A protein family search was done to confirm BLAST-based annotations of sponge transcriptomes as well as to predict function for sequences that did not have a good BLAST hit. To this end, conserved domains and functional motifs present in sponge transcriptomes were identified using a protein-protein search against the PfamA.hmm database, version 27.0 (261). This approach was used to find homology between sequences using Hidden Markov models (HMMs) associated with each protein family. The "resolve clan overlaps" option was selected and a significance threshold of E-value ≤ 0.01 was set. Significant HMM matches were tabulated for each sponge and were combined into a Pfam matrix. The matrix was converted to a binary presence-absence matrix for comparative analyses in R. Hierarchical cluster analysis (HCA) was performed using Bray-Curtis dissimilarity and average or Ward's linkage methods for clustering calculations, implemented in R. Both linkage methods yielded similar results.

4.2.5 Identification of putative pathways and interactions between expressed genes

A combination of analyses in InnateDB and MEGAN were used to identify the pathways and cellular processes represented in the transcriptome datasets. The BLASTp results against the human gene set in InnateDB were used to identify pathways expressed by sponges using InnateDB's pathway overrepresentation, network and interactor analyses. For most pathways, the source database was KEGG. A KEGG pathway analysis was performed using MEGAN and the RefSeq BLASTp output. The MEGAN software package matches each sequence with a KEGG Orthology accession number, based on the best BLAST hit for which the KEGG accession number is known (262). Sponge transcriptomes were compared based on the pathways that contain the identified KEGG orthologs as well the distribution and composition of KEGG orthologs. Hierarchical cluster analysis was done in R as described in section 4.2.4.

4.2.6 Identification of protein-coding genes from existing Oscarella carmela EST data

To expand the comparative dataset of identified sponge host defense and immunity genes, public transcriptomic information from an *O. carmela* EST library (44) was mined for sequences with homology to genes involved in innate immunity. Putative *O. carmela* protein sequences were predicted using the Fgenes bamg.pl script. Predicted protein sequences were then queried against NCBI's nr database and InnateDB human and mouse gene list using BLASTp, using an E-value < 1E-6 cutoff. All annotations were examined manually. Since the *O. carmela* data was available before the *A. queenslandica* genome or any sponge transcriptome sequencing efforts, it was used as the query database to identify peptides in the proteome of a single *D. mexicanum* sponge (SB2). See Materials and methods in Chapter 3 for a description of methods used for proteome profiling.

4.3 Results

4.3.1 Composition of sponge transcriptome datasets

To reconstruct sponge pathways for host defense and innate immunity genes and pathways with possible roles in symbiont recognition, transcriptomes for five *D. mexicanum* and one *T. californiana* adult sponges were generated. Each of the six transcriptomes contained between 147 and 160 million reads (Table 4.2). The average read length was similar in all sponge datasets, with a mean of 284 bp (Table 4.2). High quality reads were assembled *de novo* using ABySS (253). The number of contigs assembled for each sponge varied, ranging from ~0.6 to ~2.8 million contigs. The average contig lengths were shorter in *T. californiana* (780 bp) than *D. mexicanum*, which had an average contig length of ~970 bp across all five individuals (Table

4.12. Overall, the N50, average read length, number of contigs and total amount of sequence information in this study exceeded the values of these metrics in other sponge transcriptome sequencing efforts (Table 4.2) (110, 113-115). Contigs larger than 180 bp were used for ORF prediction, taxonomic characterization and annotation. Consistent with contig length differences, predicted ORFs from *T. californiana* tended to be shorter than ORFs predicted in the *D. mexicanum* transcriptome.

4.3.1.1 Taxonomic composition of transcriptomes

The taxonomic distribution of ORFs with a RefSeq BLAST hit was similar across all six transcriptomes, with ~92% ORFs assigned to a metazoan taxon (Table 4.3). Taxonomy could not be assigned for <0.2% ORFs used passing QC filtering. Only ~35% ORFs in each sponge were assigned to Porifera, and more specifically, *A. queenslandica*, indicating that about 2/3 of the predicted ORFs were more similar to other animals than *A. queenslandica*. Since *A. queenslandica* is the only sponge species in the RefSeq database used, it is possible that with additional sponge sequences, more ORFs in this study would be assigned to Porifera. However, BLAST analyses indicate that at least some sequences predicted for sponges in this study are more similar to mammalian than sponge homologs.

Sponge	# Reads passing QC	Average read length	Total # of contigs	N50	Average contig size (bp)	# Contigs for ORF prediction	# ORFs >60 a.a.	Average ORF length (a.a.)	Length of longest predicted ORF (a.a.)
SB1	149,368,622	298	938,092	1,575	930	142,430	118,454	306	36,837
SB2	155,334,066	291	1,187,979	1,403	881	201,333	159,912	276	13,533
SB5	147,650,040	280	787,597	1,530	963	104,590	96,311	303	5,528
SB6	153,924,808	281	592,143	1,676	1,050	114,096	106,855	324	5,403
SB8	146,721,966	283	608,279	1,653	1,017	113,930	102,285	322	23,571
SB10	159,593,058	273	2,874,679	1,211	780	139,201	116,416	248	3,887

Table 4.2 Summary of sponge transcriptome sequencing

Table 4.3 Taxonomic distribution of sponge transcripts.

Sponge	Opisthokonta	Archaea	Bacteria	Viruses	Not assigned
SB1	67,389	965	1,337	6	134
SB2	85,529	584	1,836	29	171
SB5	56,703	103	1,301	16	115
SB6	63,732	80	1,369	13	130
SB8	60,971	421	1,369	0	126
SB10	63,369	66	1,067	107	85

Sponge	# ORFs	RefSeq *	COG	KEGG	InnateDB (Mouse)	InnateDB (Human)	# HMM matches	# ORFs with HMM match	# different HMMs
SB1	118,454	73,321	7,889	67,630	23,463	39,146	101,834	59,295	4,240
SB2	159,912	93,023	32,238	86,697	29,953	49,934	124,709	75,159	4,293
SB5	96,311	61,571	22,736	57,497	19,511	33,365	84,422	51,933	4,146
SB6	106,855	69,137	25,420	64,370	22,712	37,382	97,359	57,647	4,194
SB8	102,285	66,127	24,640	61,445	21,791	36,175	94,003	54,769	4,182
SB10	116,416	68,230	24,267	63,895	19,589	35,135	82,664	55,888	4,021

Table 4.4 Distribution of sponge transcriptome BLASTp hits (Evalue < 1E-6, Bitscore > 50) and Pfam matches (Evalue <0.1)

* Used for taxonomic assignment
4.3.1.2 Functional annotation of transcriptomes

Homologs were identified for ~62% ORFs in each sponge transcriptome by querying reference databases by BLAST (Table 4.4). Most of the annotations were derived from matches to sequences in the RefSeq and KEGG databases, and <10% of the sequences with BLAST hits were similar to uncharacterized proteins from other metazoans. Generally, there was agreement between databases, and a common annotation for predicted ORFs that had homologs in more than one database could be assigned. To further validate and expand on BLAST-based annotations, a BLAST-independent method based on Pfam searches was used to identify functional domains found on predicted ORFs. An average of 97,500 Pfam matches that had an E-value <0.1 were found in each transcriptome, on about half the ORFs (Table 4.4). Over 4,200 different Pfams were identified in each transcriptome, with a total 4,634 Pfams shared between all six sponges, indicating little variation of expressed genes between samples. The transcriptomes of *D. mexicanum* and *T. californiana* contain all the protein families highlighted for roles in innate immunity for other sponge species, including SRCR domains, ankyrin, NHL and WD40 repeats, fibronectin, and A2M (45, 114, 115, 120).

4.3.2 Diversity and distribution of pathways and predicted ORFs in sponge transcriptomes

There was little difference in biological pathway representation between the *D. mexicanum* and *T. californiana* datasets based on the distribution of KEGG pathways (Figure 4.1). All six transcriptomes contained pathways classified into six broad categories, represented by similar proportions between samples. The "Metabolism" pathway category contained the largest number of ORFs in all six sponges, with the majority mapping to carbohydrate, lipid and amino acid metabolism pathways (Figure 4.1). An equivalent number of ORFs mapped to the "Organismal systems" and "Human diseases" categories, with the largest component being "immune system" and "cancer" pathways, respectively. Other organismal systems pathways included hormone production, excretion, contraction (circulatory system), long-term depression and potentiation (nervous system), and sensory systems. The "Cellular processes" category encompassed pathways involved in cell signalling, motility, cell growth and death, as well as transport and catabolism. The "Environmental-" and "Genetic information processing" categories were represented by ORFs in pathways involved in signal transduction, membrane transport, as well as replication, repair and gene expression. It is important to note that some ORFs mapped to multiple pathways, as well as multiple KEGG pathway categories.



Figure 4.1 KEGG pathway composition and distribution in sponge transcriptomes.

Results are shown for one *D. mexicanum* sponge (SB1) and the *T.californiana* sponge. All *D.mexicanum* individuals had a similar pathway composition and distribution.

Although there were no pathways that were expressed solely in one the two sponge species, some differences between D. mexicanum and T. californiana, and between D. mexicanum individuals, could be observed at the gene, or the KEGG ortholog, level. Similarities between transcriptomes were investigated by HCA based on the presence or absence of all KEGG orthologs in each transcriptome (Figure 4.2A). The T. californiana (SB10) sponge clusters away from the five *D. mexicanum* sponges, suggesting that the two species contain a different complement of homologs involved in similar pathways. The same pattern was observed when only orthologs associated with immune responses were considered (Figure 4.2B). However, the differences are due to less than half of the predicted proteins, since most of the immune system orthologs (61%) were shared between all six transcriptomes (Figure 4.3A). Approximately 9% "immune system" KEGG orthologs were found only in T. californiana, whereas another 6% were found in all five D. mexicanum sponges only (Figure 4.3A). Similarly, HCA was also performed to test the relationships of the diversity and distribution of Pfams across the six sponge transcriptomes. The Pfam HCA dendrogram showed a similar pattern as KEGG orthologbased clustering, although the difference between the two species was less pronounced (Figure 4.2C), as 79% Pfams are shared between transcriptomes of all six individuals, and only 3% were unique to T. californiana, and 6% unique to D. mexicanum (Figure 4.3B). The identities of Pfams and KEGG orthologs unique to D. mexicanum and T. californiana are listed in Appendix C.1 and C.2. Since proteases play important roles in many essential cell processes, including innate immunity pathways, proteases expressed by the sponge were categorized into five recognized classes (263) (Appendix C.3).



Figure 4.2 Comparison of sponge transcriptomes based on pathway, gene and Pfam distribution.



Figure 4.3 Shared and species-specific expressed genes and protein families.

4.3.3 Innate immunity genes and pathways in *D. mexicanum* and *T. californiana*

Based on a combination of KEGG pathway mapping, Pfam protein family and InnateDB pathway over-representation analyses, a diversity of innate immunity pathways were identified in *D. mexicanum* and *T. californiana* including pathways not previously reported in sponges. The search targeting innate immunity genes in O. carmela EST data (44), identified molecules that may be involved in innate immunity in this sponge, including the TLR signalling pathway. Specific immunity pathways identified in these three sponge species were summarized, and compared to pathways in eumetazoans and other sponge species where possible (Figures 4.4-4.9). For A. queenslandica, I used a combination of reported pathways and genes (45, 110-112) and mapping to KEGG pathway mapping available http://www.genome.jp/dbget-bin/get linkdb?at t+pathway+genome:T02284 for comparing innate immunity across the sponge phylum. The conservation of key innate immunity receptors across Metazoa was surveyed and tabulated (Figure 4.10). The presence of all, with the exception of peptidoglycan recognition receptors (PGRPs), was indicated in sponges, whereas the model invertebrates gained PGRPs but not NLRs or LPS-binding proteins.

4.3.3.1 Toll-like receptor signalling

The TLR signalling cascade is the best-characterized poriferan innate immunity pathway, with the adaptor, signalling molecules and downstream transcription factor described in other sponge species. Both *D. mexicanum* and *T. californiana* express all necessary genes for MyD88-dependent TLR signalling, which results in activation of the nuclear factor κ -light-chain-enhancer of activated B cells (NF κ B) and activator protein 1

(AP-1) transcription factors (Figure 4.4). Both NF κ B and AP-1 were detected in D. mexicanum, but only NFkB was found in T. californiana. The absence of the TRIF adaptor protein and interferon regulatory factor (IRF) transcription factors indicates the lack of expression of MyD88-independent TLR signalling in these sponge species. Additionally, transcripts with sequence similarity to LPS-binding protein (LBP) as well as the inhibitory adaptor protein have Toll interacting protein (TOLLIP) were identified. Canonical TLRs were not identified in *D. mexicanum* or *T. californiana*, as contigs with homology to TLR sequences contained only Toll/interleukin-1 receptor TIR domains, and lacked leucinerich-repeats (LRR). However, both species contained multiple transcripts containing LRRs. A total of three different types of TLR-like genes were found in D. mexicanum and T. californiana transcriptomes. Two of the five D. mexicanum specimens expressed two types of TLR-like genes (TLR6-like and TLR2-like), whereas the other three specimens expressed only one type (TLR6-like). On average, 5 contigs in each D. mexicanum sponge were similar to TLR6, and contained a TIR 2 domain only, and 2 contigs had homology to TLR2, containing either only a TIR domain or a TIR and Ig 3 Pfams. The top protein hit (~50% ID) for the TLR6-like sequences was the TLR protein from S. domuncula, followed by mammalian matches TLR6 proteins. Conversely, the top protein hit (~30% ID) for the TLR2-like genes was the A. queenslandica TLR-like gene, followed by avian TLR sequences. Tethya californiana expressed two TLR variants, 5 contigs containing TLR1like sequences and 1 contig with a TLR2-like sequence. The closest sequence to the T. californiana TLR2-like sequence was the A. queenslandica TLR-like gene (~30% ID), whereas the TLR1-like gene was most similar to S. domuncula (~40% ID) and The TLR2like sequences were identical between the two D. mexicanum sponges, but shared $\sim 70\%$

identity at the nucleotide level, with 50% query coverage, with the *T. californiana* homolog. No TLR-like genes were found in *O. carmela*, however, components of MyD88-dependent TLR signalling were expressed in this sponge (Figure 4.4).



Figure 4.4 Toll-like receptor signalling in sponges.

Presence of predicted *D. mexicanum* and *T. californiana* TLR signalling components is based on their identification in the transcriptomes by BLAST.

4.3.3.2 Nod-like receptor signalling

The NLR signalling pathway components were well represented in the *D. mexicanum* and *T. californiana* transcriptomes (Figure 4.5). Sponge NLR-like proteins could be involved in inflammasome signalling and caspase-1 activation via NLRP3-like genes, found in all six transcriptomes, and NLRP1-like genes found in *D. mexicanum* but not *T. californiana* datasets. Additionally, sponge NLR included sequences with homology to NOD1 and NOD2 genes, and the downstream molecules that act to activate MAPK signalling and NF κ B. The NOD1-like genes in *D. mexicanum* (about 7 contigs per sponge) and *T. californiana* (1 contig) contained only LRR_6 domains, whereas the NOD2-like genes were represented by contigs that either contained NACHT or LRR_6 domains. There was a greater diversity (more individual contigs) of NOD2-like sequences within each transcriptome than of NOD1. Although no individual contig contained LRR, NACHT and interaction domain, all domains that form NLRs were found in the transcriptomes. The potential interaction domains in *T. californiana* and *D. mexicanum* could be CARD, DED and death domains. Unlike the differences observed in TLR sequences, the NLR (NOD1 and NOD2) sequences did not differ greatly between *T. californiana* and *D. mexicanum*. When compared to reference datasets, both NOD1 and NOD2 expressed by these two sponge species shared ~35% ID at the protein level with, variably, mammalian and reptilian NLRs, *A. queenslandica* and choanoflagellate predicted genes. An NLR, NLRC3, was found in *O. carmela*, but no evidence for inflammasome signalling was found for this sponge.



Figure 4.5 NOD-like receptor signalling in sponges.

4.3.3.3 Phagocytosis and autophagy

Although sponges use phagocytosis to obtain nutrition, the pathway involved in this process has not been characterized in sponges. Further, phagocytosis plays an essential role in eumetazoan host defense (264). In order to better understand the evolution of phagocytosis mechanisms and function, transcriptomes were interrogated for components of this pathway in sponges. Both T. californiana and D. mexicanum express genes involved in Fcy receptor-mediated phagocytosis. While no Fcy receptors were detected, all genes necessary for downstream signalling were expressed, as well as genes with homology to the protein tyrosine phosphatase CD45. Phagocytosis involves regulation of actin cytoskeleton, membrane remodeling, phagosome formation and maturation, particle internalization, and finally digestion. As a phagosome matures, it fuses with a lysosome to form phagolysosome, where reactive oxygen species (ROS) are released and together with lysosomal hydrolases digest the engulfed materials. All the sponges in this study expressed genes necessary for ROS production, as well as lysosome-associated genes. Moreover, almost all lysosomal membrane and acid hydrolase transcripts were expressed in D. mexicanum and T. californiana. Since lysosomes are also involved in digestion of material acquired by endocytosis and autophagy, sponge transcriptomes were examined for the presence of these pathways. A full complement of transcripts involved in clathrindependent and clathrin-independent endocytosis was identified in both sponge species, including the receptors. There were minimal differences in the expression of phagocytosis, endocytosis and lysosome genes between D. mexicanum individuals and between D. mexicanum and T. californiana.

Autophagy is of great functional importance in immunity and inflammation, however this pathway has not been described in sponges. Strong evidence for the presence of this pathway in both *D. mexicanum* and *T. californiana* was found, as most of the autophagy related (ATG) and vacuolar protein genes comprising the pathway, including beclin-1, were detected in all six specimens (Figure 4.6). Components of the autophagy pathway are also present in the *A. queenslandica* genome, although the homologs mapped were slightly different from the two sponges species profiled in this study. I find only 1 gene, ATG8, in the *O. carmela* EST sequences.



Figure 4.6 Autophagy-associated genes in *D.mexicanum* and *T. californiana*.

Presence of predicted *D. mexicanum* and *T. californiana* autophagy genes is based on their identification in the transcriptomes by BLAST.

4.3.3.4 Lectins, complement and coagulation

Lectin and malectin-like genes were found in the transcriptomes of all six sponges and galectin homologs were expressed in *D. mexicanum*. However, while some components of the complement system were expressed in a subset of the sponges, no genes of the complement cascade were expressed in all six sponges. I found that most of the genes involved in the complement system and membrane attack complex (MAC) formation were not detected in *T. californiana* or *D. mexicanum*, and therefore this pathway is incomplete (Figure 4.7). It is not likely that sequences were erroneously mapped to this pathway since all the genes identified by BLAST were confirmed by Pfam HMM searches, and a subset of the proteins detected by transcriptome sequencing were also detected in the proteome. Curiously, I found a few of genes involved in the coagulation cascade, which can activate the complement system. Similar to the observations made for the complement system, none of the genes involved in coagulation were detected in all six specimens, and the majority of the genes involved were absent (Table 4.5). Although serpins inhibit proteases involved in both complement and coagulation cascades, I do not find any serpins expressed in either *D. mexicanum* or *T. californiana*.

F7	coagulation factor VII			
F11	coagulation factor XI			
	coagulation factor II			
F2	(thrombin)			
F13	coagulation factor XIII			
	tissue factor pathway			
TFP1	inhibitor			
PROC	protein C			
PROS1	protein S			
PLG	Plasminogen			
A2M	alpha-2-macroglobulin			

Table 4.5 D. mexicanum genes with roles in coagulation.





Presence of predicted *D. mexicanum* and *T. californiana* complement genes is based on their identification in the transcriptomes by BLAST.

4.3.3.5 Viral recognition mechanisms

The ability to detect and respond to viral infection was indicated in *D. mexicanum* and *T. californiana* but not *O. carmela*. The retinoic acid-inducible gene 1 (RIG-I)-like receptor (RLR) signalling pathway detects and initiates an antiviral response to infection by RNA viruses. *D. mexicanum* and *T. californiana* express the RLR genes RIG-I, melanoma differentiation-associated protein 5 (MDA5) and RIG-I-like receptor 2 (LGP2), as well as most of the downstream adaptor and signalling molecules (Figure 4.8). These two sponge species also express the components of molecular machinery to detect cytosolic DNA, and thus possibly DNA viruses. However, many of the genes involved in this pathway were not detected in either sponge species. Furthermore, no Type I interferons or the inflammatory cytokines, for example TNF α , important for eliminating viral pathogens were found.



Figure 4.8 RIG-I-like receptor signalling in sponges.

4.3.3.6 Apoptosis, transendothelial migration and adaptive immunity pathways

In addition to the specific host defense pathways described above, an almost complete set of genes required for apoptosis was identified, which has been previously described in other sponge species (45, 118, 265). Similarly, I find genes involved in the extrinsic and intrinsic apoptosis pathways in *D. mexicanum* and *T. californiana* (Figure 4.9). Additionally, the presence of pathways associated with adaptive immune responses including leukocyte transendothelial migration pathways, antigen processing and presentation, T cell receptor signalling, B cell receptor signalling and FceRI-mediated signalling pathways was indicated. The latter three pathways were mapped largely due to expression of genes in the MAPK, phosphatidyl inositol, and calcium signalling cascades, kinases, and transcription factors that are also involved in other pathways. Genes mapped to antigen processing and presentation included proteases, heat shock protein (HSP) 70 and HSP90, and transcription factors. Although sponges do not have leukocytes and sponge cells are not likely to move across the epithelium, *T. californiana* and *D. mexicanum* express a full collection of genes associated with transendothelial migration.

	D. mexicanum	T. californiana
Fas-L	\bigcirc	\bigcirc
TRAIL	\bigcirc	\bigcirc
TNFα	\bigcirc	\bigcirc
IL1	\bigcirc	\bigcirc
Fas	\bigcirc	\bigcirc
TRAIL-I	\sim \bigcirc	\bigcirc
TNF-R1	\bigcirc	\bigcirc
IL1R	\bigcirc	\bigcirc
FADD		
TRADD	\bigcirc	\bigcirc
RIP1	\bullet	\bigcirc
TRAF2	\bigcirc	\bigcirc
MyD88		
IRAK	\bullet	\bullet
NGF	\bigcirc	0
IL3	\bigcirc	\bigcirc
TrkA		
IL3R	\bigcirc	\bigcirc
PI3K		
Bcl2		\bullet
Bcl-xL		\bigcirc
Bad	\bigcirc	\bigcirc
IAPs		
Apaf-1		\bullet
cytC		\bullet
Bax, Bak		\bigcirc
AIF		\bullet



Figure 4.9 Apoptosis-associated genes expressed in *D. mexicanum* and *T. californiana*.

	- 	×	<u></u>	ン	
TLRs	+	+	+	+*	+
NLRs	+	+	-	-	+
Cytokines	+	+	+	-	+
Lectins	+	+	+	+	+
Compelement system	+	+	+	+	+
PGRPs	+	+	+	+	-
SRCRs	+	+	+	+	+
LBP/BPI/CETP	+	-	-	+	+

*Not involved in immunity

Figure 4.10 Evolutionary conservation of key innate immune signalling molecules.

4.4 Discussion

4.4.1 Distribution of *D. mexicanum* and *T. californiana* transcripts

Almost all transcripts from *T. californiana* and five *D. mexicanum* specimens likely represent host mRNA, as >90% were associated with Metazoa, and 30% were mapped to *A. queenslandica*. Similarly, there were more BLAST hits mapping to Metzoa than to Porifera in the transcriptome of the sponge *Cliona varians* (247). This could be because the only sponge sequences in RefSeq are from *A. queenslandica*. When comparing transcriptome sequences against the NCBI nr database, I observed that *D. mexicanum* and *T. californiana* shared greater sequence identity with *S. domuncula* than *A. queenslandica* where homologs for both reference sponges were available. Since both *D. mexicanum* and *T. californiana* are more closely related to *S. domucnula* than *A. queenslandica*, a phylogenetic pattern of poriferan functional gene sequence conservation is implicated. There are over 8,500 sponge species, more than 7,000 of which are demosponges, thus

there are likely to be considerable differences between demosponge sequences (19). Even though there has been increased activity in transcriptome sequencing of diverse sponge species, it is difficult to perform direct comparisons between the datasets from this study and the new-published transcriptomes since they relied on BLASTx for function prediction and no ORF predictions or selection for eukaryotic RNA were performed in those studies (114, 115). In agreement with the observation that most *D. mexicanum* and *T. californiana* sequences were most similar to deuterostomes, the EST libraries of the demosponges *S. domuncula* and *Lubomirska baikalensis* also shared more homologs with deuterostomes than with the more closely related *Caenorhabditis elegans* or *Drosophila melanogaster* (248), suggesting possible sequence divergence within Porifera.

The functional composition of transcripts expressed by the five *D. mexicanum* and one *T. californiana* sponges were highly similar, although subtle differences between the two sponge species were indicated by HCA. In addition to comparing the distribution of predicted genes and protein families, I mapped expressed genes to known pathways to predict how those genes might interact with each other. The same pathways, with small variations in the identity of homologs involved, were detected in *D. mexicanum* and *T. californiana* transcriptomes. Furthermore, there were only minor differences in gene expression in the five *D. mexicanum* specimens, suggesting similar individual responses to the aquarium environment and nutritional conditions, as well as confirming reproducibility of the methods employed. Among the "unique" KEGG orthologs were functionally similar proteins, such as TLR6 homologs in *D. mexicanum* only, and TLR1 homologs in *T. californiana*. Expression of proteins (detected in the proteome) with galectin Pfam homology, which may play roles in sponge aggregation or microbial recognition (266,

267), was detected in *D. mexicanum* but not *T. californiana*, indicating differences in interactions with cell surfaces.

4.4.2 Microbial recognition and response mechanisms in *D. mexicanum* and *T. californiana*

The diversity of molecules associated with innate immunity expressed by the demosponges *D. mexicanum* and *T. californiana* suggest that these metazoans are able to detect and respond to extracellular and intracellular microbial signals. Both of these sponge species have the capability to produce ROS and activate transcription factors that regulate cytokine and antimicrobial peptide production. *D. mexicanum* and *T. californiana* express pathways that could be used to detect invading RNA and DNA viruses, a novel finding in sponges, indicating that sponges may indeed have anti-viral immune responses, as was proposed when a 2',5 -oligo A synthetase homolog was reported in *G. cydonium* (117).

TLR signalling is the best-understood immunity pathway in sponges, and components of the pathway downstream of the receptor are well conserved (45, 106, 109, 111, 112, 115, 249). Specifically, sponge TLRs are likely to signal via the MyD88-dependent signalling pathway, which is used by most known TLRs (268). However, similarity of sponge TLR receptors remains poorly defined. Sponge TLR-like sequences have been reported for four demosponge species, *S. domuncula*, *A. queenslandica*, *Ircinia fasciculata* and *Petrosia ficiformis*, and one homoscleropmorph sponge, *Corticium candelabrum*, and all lack conventional LRR domains (109, 111, 115). In agreement, *D. mexicanum* and *T. californiana* TLRs also lack LRR domains on the same contig, although multiple LRRs were identified on other ORFs. Since all sponge TLR-like proteins

described thus far do not contain the conventional domain structure, it is possible that the poriferan homologs represent the ancestral state of the protein family (45). The *D. mexicanum* and *T. californiana* TLRs had homology to TLR6 and TLR1, respectively, which were most similar to the *S. domuncula* TLR. Both sponges also had TLR2-like sequences, which were most similar to the *A. queenslandica* TLR. Thus, I detect the entire complement of sponge TLRs described thus far in *D. mexicanum and T. californiana*. It is likely that these TLRs are used to recognize bacterial cells since the ligands of mammalian homologs include bacterial lipoprotein and peptidoglycan (75). Indeed, the *S. domuncula* TLR, which is constitutively expressed in the epithelium, initiates a signalling cascade resulting in caspase up-regulation in response to lipoprotein (269). Heterodimerization of different TLRs, allows for the recognition of additional MAMPs not targeted by either individual TLR type (269). Therefore, it is possible that a similar strategy to increase the diversity and specificity of microbial signals recognized could be at play in sponges.

Conversely, the NLR family of intracellular receptors has previously been reported only in a single sponge, *A. queenslandica*, which has >130 NLRs (124, 244). NLRs play an essential role in detecting microbes in mammals that have entered the host cell by recognizing MAMPs within the cytosol (128, 239). Although I do not find the same extent of NLR diversity identified in the *A. queenslandica* genome in either *D. mexicanum* or *T. californiana*, I find that sponge NLRs may be involved in inflammasome signalling, as well as MAPK signalling and NF κ B activation (270) previously not known in sponges. NLRs occur in all metazoa, which can be categorized into two monophyletic groups (124). All *A. queenslandica* NLRs form part of the group that also contains all human NLRP and most of the human NLRC genes, including NOD1 and NOD2 (124). Due to sequence homology, it is likely that the D. mexicanum and T. californiana predicted NLR genes also fall into this NLR lineage, implicating that sponge NLRs may represent the ancestral form of the gene. None of the NLR-containing sequences in T. californiana and D. mexicanum had all three (LRR, nucleotide-binding and interaction) domains. A possible explanation for the observed TLR and NLR domain distributions could be genetic recombination that could allow for greater flexibility in the PRRs in these earliest-branching animals. In the cnidarian *Hydra*, which also lacks canonical TLRs, LRR and TIR containing proteins form a complex as part of host response to bacterial pathogens (271). Recombination, domainshuffling, gene duplication and gene loss all likely played important roles in NLR evolution (124). PRR expansion is evident in the genome of the purple sea urchin, which encodes a diverse 222 TLRs and 203 NLRs (245). The hypervariability of sea urchin PRRs may expand the diversity and precision of the innate immune response (245, 246). Perhaps the large A. queenslandica inventory allows the sponge to specifically recognize a more diverse set of microbes than mammalian NLRs.

The sponges in this study have the necessary mechanisms to take up and digest particles from outside the cell and within the cell, via endocytosis, phagocytosis and autophagy. Phagocytosis is an important host defense mechanism that is involved in recognition of microbes and apoptotic cells (272). In vertebrates, phagocytosis is involved in adaptive immunity and innate immunity (264, 273). Indeed, one of the functions involving phagocytosis that evolved in jawed vertebrates is for antigen processing and presentation (274). Therefore, it is possible that this pathway is found in *D. mexicanum* and *T. californiana* due to conserved immunity mechanisms. Conversely, this pathway may

function solely for food particle uptake as in protists (275, 276) although this seems less likely. Immune responses in eumetazoan invertebrates include PRR signalling leading to antimicrobial peptide production. hemoplymph coagulation, melanization. prophenoloxidase activation, lectin complement activation, and phagocytic systems (89, 277-280). In sponges, phagocytosis is primarily carried out by the large, motile and totipotent archaeocytes (278). Motility of the sponge cells is likely the reason that both D. *mexicanum* and *T. californiana* expressed genes mapping to the transendothelial migration pathway. It is accepted that phagocytosis in sponges is non-selective (278). Yet, uptake mechanisms can vary depending on size and type of particle. (281). Further, sponges take up specific bacterial strains, implying some degree of specificity in poriferan phagocytosis (106).

Particles to be internalized can be recognized either directly by receptor proteins or opsonins, such as LBP found in *D. mexicanum* and *T. californiana* transcriptomes, which cover the particle and interact with specific surface receptors on phagocytes (272, 282). The lectin activation of the complement pathway, and ultimately phagocytosis and pathogen killing, involves recognition of microbial surface carbohydrates by PRRs (283). Mannose binding lectin (MBL) is a circulating PRR and opsonin that can activate the complement system through complexes with MBL-associated serine protease (MASP) (272, 283, 284). MBL, lectin and malectin-like genes were found in the transcriptomes of all six sponges and galectin homologs were expressed in *D. mexicanum*. Although, I do not find MASP in *D. mexicanum* or *T. californiana*, MASP homologs were expressed in *C. candelabrum* (45). Together with expression of C3 and other complement genes, these observations suggest earlier evolution of lectin-serine protease complement activation than

previously thought.

Autophagy is a conserved process important for maintain cellular homeostasis and stress response that eliminates intracellular targets (285). Autophagy is an ancient process, found in protists and fungi, where it plays a role in stress survival (286, 287). This process is also involved in immune response to intracellular bacterial and viral pathogens, either by direct removal or in concert with PRR signalling in vertebrates and invertebrates (288-293). These results indicate that this pathway is present in sponges, where it may a role in innate immunity mechanisms in addition to homeostasis (290).

Innate immunity pathways regulate and interact with each other and other pathways (294, 295). In addition to antimicrobial responses, PRR signalling can promote or inhibit apoptosis, which is expressed in *D. mexicanum* and *T. californiana* (289, 295, 296). With the exception of PGRPs, I show that sponges are equipped with key innate immunity molecules and pathways to recognize invading pathogenic microbes (73, 75). The same mechanisms are also involved in symbiont recognition (39, 93, 239). It has been proposed that invertebrates do not require a complex immune system capable of highly specific recognition in part because they host relatively simple resident microbial communities (14, 297). Marine sponges live in a microbe-rich environment and host microbiota whose diversity and composition are more similar to that of mammals than the squid, *C. elegans* or *Drosophila* species (39, 298-300). Thus, perhaps sponges appear to have more complex innate immunity systems than *C. elegans* and *D. melanogaster*, with possible diversification of key PRR molecules, because of the more complex nature of their symbiotic microbial communities.

The innate immunity in D. mexicanum and T. californiana is a complex and interdependent system that is capable of detecting microbes, eliminating invading pathogens and removing infected sponge cells. Identification of adaptive-immunity pathways in sponges highlights the conservation of molecules across animals, and indicates that the molecular mechanisms necessary for mammalian immunity pathways evolved prior to the split of eumetazoans and sponges. This observation supports the hypothesis proposed by Nichols and colleagues based on work with O. carmela that eumetazoans evolved new characteristics by combining existing cell signalling, development and adhesion molecules (44). That T. californiana and D. mexicanum express the same innate immune signalling pathways suggests that the two sponge species use similar mechanisms in their interactions with microbes. However, these two sponge species host distinct and specific microbial communities (Chapter 2). Therefore, it could be the differences in receptor sequences that drive the specificity and selection of the microbiota. The D. mexicanum transcriptomes represent the largest poriferan dataset and provide insight into the complexity of sponge innate immunity pathways, evolution of innate immunity genes, and serves as an invaluable resource for future cell biology and biochemical studies aimed at understanding hostsymbiont recognition.

Chapter 5: Conclusion

Animals have co-evolved with microbes, and therefore the animal immune system evolved to recognize symbiotic as well as pathogenic microbes. Indeed, the evolution of eukaryotic cells is a result of a symbiotic association. Interactions with symbiotic microbes are necessary for the health and development of their animal hosts. Sponges represent the most ancient extant animal phylum, and therefore hold much insight into evolution of animal-microbial symbioses. Marine sponges are important members of marine ecosystems, providing habitat, creating and stabilizing reefs, as well as coupling benthic and pelagic food webs throughout the world's oceans (20-27). As filter feeders, sponges contain large microbial populations with higher cell densities than surrounding waters that contribute to multiple ecosystem functions including primary production and nutrient cycling (32, 50-55). Despite their simple anatomy and physiology, sponges have persisted for ~600 million years, and their symbiotic microbes are likely a key contributing factor to their evolutionary success (21). Moreover, as sessile filter feeders, sponges are important indicators of water quality and ocean health. Therefore, understanding how spongeassociated communities come together is of high importance from both an evolutionary and ecological standpoint. Despite the ecological importance of sponge symbioses, mechanisms mediating symbiont recognition and host immune surveillance are unknown. Furthermore, sponge-associated microbial communities are diverse, specific and largely uncultivated, necessitating the use of cultivation-independent methods (15, 21, 59, 60). In this thesis, I used the symbiosis between the sponge *Dragmacidon mexicanum* and the thaumarchaeaote Cenarchaeum symbiosum to better understand the molecular and metabolic processes underlying sponge-microbial associations. I used a combination of community structure 118

characterization, comparative genomics, homology modeling, biochemical techniques, proteomics and transcriptomics to characterize the specificity of the interaction and identify the likely molecular determinants of the specific symbiosis.

5.1 Dragmacidon mexicanum hosts a specific microbial community

Selection of symbionts by the host is implicated as different sponge species host microbial communities distinct in composition and structure. This is the first study to compare three-domain holobiont composition in three sponge phylogenetic classes. Further, the core communities and intraspecific variation for three demosponge species, *D. mexicanum*, *T. californiana* and *Haliclona* sp. were identified, further suggesting that sponges select species-specific assemblages.

Since *C. symbiosum* represents up to 65% of the *D. mexicanum* microbial community (129, 130), the identities of the constituents of 35% of the microbiota in this sponge were unknown prior to the work presented in this thesis. Understanding the other members of the community is critical to predicting interactions between microbial populations with each other and the host with reciprocal and integrated effects on function and evolution (301-303). The *D. mexicanum* bacterial community included hundreds of low-abundance taxa, from up to 13 bacterial phyla, including Bacteroidetes, Proteobacteria and Firmicutes. *D. mexicanum* has a specific bacterial community distinct from other sponges, consistent with species-specific interactions. The most abundant bacterial indicators and bacterial core were γ -proteobacterial species, Candidatus *Endobugula* and the sponge-associated E01-9C-26 marine group. Candidatus *Endobugula* is a symbiont of marine bryozoans that produces polyketide lactones (bryostatins) for chemical defense that

protect the host from predation (304-306). It is possible that *D. mexicanum*-associated bacteria have a similar functional role. It is not known how these symbionts are transmitted in sponges and whether they are present in seawater, requiring further investigation.

Although the symbiosis between *D. mexicanum* and *C. symbiosum* was described almost 20 years ago, no methods to accurately measure absolute *C. symbiosum* abundance in the sponge were available (130). Therefore, I developed tools to quantify specific populations, both at the domain and species-level using SSU rRNA and *C. symbiosum* serpins as taxonomic marker genes. This is a simple approach that has established the constancy and predominance of this archaeal symbiont in its host. As part of this foundational thesis, it provides motivation for the other investigations into the nature of this highly specific association.

5.2 *Cenarchaeum symbiosum* acquired genes not found in other thaumarchaea to interact with the host

The high abundance, constancy and persistence of viable *C. symbiosum* cells in *D. mexicanum* (130) suggest that this archaeal symbiont performs an important function that benefits the host. Analysis of the *C. symbiosum* genome suggests a role in metabolic exchange and detoxification services through ammonia and urea uptake and oxidation (51). Yet, the mechanisms used by *C. symbiosum* to establish and maintain a large population within the host extracellular matrix are not known. This thesis provides insight into potential distinguishing features between the genomes of free-living thaumarchaea and that of *C. symbiosum*. This is a valuable contribution to the field of comparative genomics and

forms a basis for inferring the evolution and functional requirements for archaeal taxa to occupy a symbiont niche within multicellular hosts.

Comparative analysis of the *C. symbiosum* genome suggests that the symbiont carries genetic adaptations for a symbiotic lifestyle. Specifically, *C. symbiosum*'s repertoire of protein encoding genes is significantly different from other free-living archaea and may indicate formative events of genome expansion leading to the incorporation of ORFs. Unique genes were identified in *C. symbiosum* included those whose products are implicated in cell surface modifications, processing or cleavage of signalling molecules, hydrolysis of diverse substrates, as well as genes with homology to eukaryotic proteins involved in cytoskeletal rearrangement and innate immunity regulation. It is possible that prior interactions with host organisms was the vehicle for introduction of recognition-associated genes and signal processing functionalities into the *C. symbiosum* genome by horizontal gene transfer, either from other microbial symbionts or from the host itself. This is an interesting contrast to the better-characterized instances of genome reduction in obligate symbiont microbes (168, 307-309).

As most microbes, especially ones engaged in symbioses, are uncultivated, heterologous systems must be adopted to understand the biological activity and infer function of specific microbial proteins. This thesis presents the initial characterization of proteins encoded by an archaeal symbiont, which serves as an important proof of concept. Although the activity and function of *C. symbiosum*-encoded serpins remains elusive, some conservation of function is implicated since serpin C-terminal peptides interact with eukaryotic cells in a manner reminiscent of eukaryotic serpins.

121

5.3 Sponges have the potential to recognize a variety of microbes

The species-specific nature of the taxonomically diverse sponge microbial communities implies a certain degree of sophistication and complexity of innate immunity and recognition in these early-branching animals (65). Sponges, like other invertebrates, do not possess highly specialized adaptive immunity, yet they successfully maintain a specific community structure while inhabiting an environment filled with microbes. This suggest that sponges may have alternative strategies to distinguish microbes and respond in a specific way that we do not yet understand (65). The results presented in this thesis indicate that sponges are equipped with key innate immunity molecules and pathways to recognize intracellular and extracellular microbes. *D. mexicanum* and *T. californiana* express similar pathways, however they vary in the sequence of receptors involved. Moreover, receptor molecules in many pathways either were not identified, or did not have a canonical structure described in other animal species. It is possible that these "missing" receptors are the key underlying specific recognition of microbes.

Overall, the transcriptomes of *D. mexicanum* and *T. californiana* contained most of the genes described in other sponge species, including genes encoding conserved hypothetical proteins (45, 114). These genes could represent sponge-specific genes that evolved after the divergence from eumetazoa or were lost soon after the split, and could be important for understanding sponge biological systems. Since the datasets examined are transcriptomes, there are surely pathways and genes that are encoded by the sponge but not expressed at time of sampling. It is possible that the presence of *C. symbiosum* or other symbionts down-regulates the expression of specific host genes, as observed in the symbiosis between the intracellular photosynthetic eukaryote *Symbiodinium* and the sponge

Cliona varians (247). Future studies into sponge gene expression in response to various microbes will help elucidate the recognition mechanisms used by sponges. Genes not previously reported in sponges were expressed in the *D. mexicanum* and *T. californiana*. Moreover, interactions between sponge genes identified in this work were predicted and mapped into immunity pathways, previously not done for sponge datasets. Placing genes in the context of pathways is a significant contribution allowing for identification of mechanisms that have not been previously reported in sponges and thus proposed to have evolved after sponge divergence from other animals.

5.4 **Emerging themes**

The surveys of *C. symbiosum* genes and the *D. mexicanum* transcriptome indicate the presence of archaeal enzyme activities for possible signal processing and to facilitate symbiont colonization, as well as the conservation of innate immunity pathway components in the host organism. The sponge innate immune system appears to have the capability to recognize microbes using multiple pathways, many of which are associated with recognition of structural components of surface features, or direct interactions with the microbial surface (Figure 5.1). Sponge antimicrobial responses, including the vast arsenal of antimicrobial, antifungal and antiviral compounds produced undoubtedly help shape their microbiota (21, 310). Host immune responses and phagocytic sponge cells pose a potential threat to *C. symbiosum*. One of the main questions in the field of host-microbial interactions is how do symbiotic and pathogenic microbes that have similar structural and molecular characteristics elicit such different host responses (93, 94, 311). The many unique genes present in *C. symbiosum* suggest that at least in some cases, the symbiont is

more actively involved in fine-tuning the specificity of its recognition by the host organism (94), using a variety of symbiosis factors, possibly in very specific combinations, to colonize, or maintain a population in, the host. The molecular components, such as innate immune system gene products in *D. mexicanum* and prevalent β -propellor scaffolds in *C. symbiosum* suggest that there is a biochemical configuration in place for recognition, host immune response modulation and signalling pathways to participate in maintaining a stable symbiosis. A plethora of novel hypothetical and unique putative hydrolytic activity-encoding genes were identified in *C. symbiosum*, yet the nature and source organism of their targets are unknown. Hence, an important subsequent step for framing the *D. mexicanum* – *C. symbiosum* interaction is identifying small molecule targets and substrates for the host and symbiont proteins that might be relevant to maintaining a stable association. Cell-surface characteristics, tissue remodeling, protease activity, cell adhesion and symbiont interactions with sponge mucus are likely key mechanisms involved in the *C. symbiosum* – *D. mexicanum* symbiosis.



Figure 5.1 Emerging patterns of sponge-microbial interactions in *D. mexicanum*.

5.5 Future directions

The identification of genes involved in sponge immunity and potential symbiosis determinants encoded by the symbiont informs future projects and provides sequence information for cloning and expression of gene products to investigate their biological activity and function in heterologous systems. Due to the inconclusive results of *C. symbiosum* serpin protein and C-terminal peptide biochemical characterization experiments, it will be important to revisit PBMC stimulation and pull-down experiments using both the recombinant proteins peptides with the internalization signal, with different amount of LPS and other stimulants.

To understand how sponge and symbiont cells interact, it is important to determine the spatial arrangements and microbial localization in sponge tissue sections, including the mesohyl, epithelium and around the choanosome. To this end, transmission electron and dual beam slice and view microscopy and florescence *in situ* hybridization can be used (312). Specific microbial cells can be isolated from sponge tissue sections for taxonomic identification using laser-capture microdissection microscopy. Further, MALDI imaging MS will be used to measure the distribution of *C. symbiosum* peptides and target proteins in sponge tissue to help predict activity and possible function in host-microbe interactions (313). Since cell surfaces play a large role in cell signalling and recognition, but thaumarchaeal cell surfaces are not well-characterized, *C. symbiosum* cells will be imaged using scanning electron microscopy to describe cell surface features. Isolated cells could be used for single-cell genomics and transcriptomics for both host and symbiont to better understand responses to specific conditions and treatments.

To functionally characterize and evaluate the symbiotic roles of host and symbiont genes outlined in this thesis, it will be necessary to develop *D. mexicanum* cell cultures. Although no sponge cell lines currently exist, there has been some success with short-term primary cultures, and these methods should be adapted to *D. mexicanum* (131, 314-316). In addition to evaluating sponge responses to immune challenges and specific microbial taxa, sponge cell lines will allow for experiments to investigate the function of specific *D. mexicanum* genes by using RNA interference, recently used to knock-down sponge genes in *Tethya wilhelma* and *Ephydatia muelleri* RNAi (317). Comparative genomic analysis between *C. symbiosum* and thaumarchaeal symbionts from other sponges belonging to the Axinellid family will help determine whether the putative symbiosis determinants

identified in this thesis are shared features of sponge-associated archaea. This will be important for understanding the evolution and age of the *D. mexicanum* – *C. symbiosum* association.

5.6 Closing remarks

In this thesis, I presented a robust framework for inferring host-microbe interactions governing stable symbioses in uncultivated systems. This thesis is the only study to date to examine community composition, host gene expression and symbiont gene expression from the same tissue sample, and not from different individuals separated by time, space and changing environmental parameters. The specific and predominate symbiosis between D. mexicanum and C. symbiosum provides a useful system in which to explore the evolution and function of symbiotic associations. Comparative genomics and gene expression identified potential symbiont-encoded proteins important for host-microbe interactions, which are unknown for archaea. Innate immunity pathways that have not been previously reported in sponges were detected in the dataset presented, expanding the known complexity and repertoire of sponge immunity. Further, these observations indicate that many immunity-associated genes and interactions between them have a more ancient origin than previously appreciated. I propose that the complexity of sponge innate immune signalling may reflect the complex composition of resident microbiota, which have in turn evolved strategies to thrive in the host. In conclusion, the work presented in this dissertation provides invaluable insight into the microbial adaptations to a symbiotic lifestyle, molecular interactions between archaea and eukaryotic cells, and the evolution of host-microbial interactions and recognition.

Bibliography

- 1. Douglas AE (2014) Symbiosis as a General Principle in Eukaryotic Evolution. *Cold Spring Harbor Perspectives in Biology* 6(2).
- 2. McFall-Ngai M, et al. (2013) Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences of the United States of America 110(9):3229-3236.
- 3. Gilbert SF, Sapp J, & Tauber AI (2012) A symbiotic view of life: we have never been individuals. *Quarterly Review of Biology* 87(4):325-341.
- 4. Sharon G, *et al.* (2010) Commensal bacteria play a role in mating preference of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* 107(46):20051-20056.
- 5. Whalan S & Webster NS (2014) Sponge larval settlement cues: the role of microbial biofilms in a warming ocean. *Scientific Reports* 4.
- 6. An DD, *et al.* (2014) Sphingolipids from a Symbiotic Microbe Regulate Homeostasis of Host Intestinal Natural Killer T Cells. *Cell* 156(1-2):123-133.
- 7. Blumer N, *et al.* (2007) Perinatal maternal application of Lactobacillus rhamnosus GG suppresses allergic airway inflammation in mouse offspring. *Clinical and Experimental Allergy* 37(3):348-357.
- 8. Lemus JD & McFall-Ngai MJ (2000) Alterations in the proteome of the Euprymna scolopes light organ in response to symbiotic Vibrio fischeri. *Applied and Environmental Microbiology* 66(9):4091-4097.
- 9. Troll JV, *et al.* (2009) Peptidoglycan induces loss of a nuclear peptidoglycan recognition protein during host tissue development in a beneficial animal-bacterial symbiosis. *Cellular Microbiology* 11(7):1114-1127.
- 10. Rosengaus RB, Zecher CN, Schultheis KF, Brucker RM, & Bordenstein SR (2011) Disruption of the Termite Gut Microbiota and Its Prolonged Consequences for Fitness. *Applied and Environmental Microbiology* 77(13):4303-4312.
- 11. Ringo J, Sharon G, & Segal D (2011) Bacteria-induced sexual isolation in Drosophila. *Fly* 5(4):310-315.
- 12. Shin SC, *et al.* (2011) Drosophila Microbiome Modulates Host Developmental and Metabolic Homeostasis via Insulin Signaling. *Science* 334(6056):670-674.
- 13. Maynard CL, Elson CO, Hatton RD, & Weaver CT (2012) Reciprocal interactions of the intestinal microbiota and immune system. *Nature* 489(7415):231-241.
- 14. McFall-Ngai M (2007) Adaptive immunity Care for the community. *Nature* 445(7124):153-153.
- 15. Hentschel U, Usher KM, & Taylor MW (2006) Marine sponges as microbial fermenters. *Fems Microbiology Ecology* 55(2):167-177.
- 16. Webster NS, *et al.* (2010) Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environmental Microbiology* 12(8):2070-2082.
- 17. Carlos C, Torres TT, & Ottoboni LMM (2013) Bacterial communities and speciesspecific associations with the mucus of Brazilian coral species. *Scientific Reports* 3.
- 18. Simister RL, Deines P, Botté ES, Webster NS, & Taylor MW (2012) Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environmental Microbiology* 14(2):517-524.
- 19. Van Soest RWM, et al. (2012) Global Diversity of Sponges (Porifera). Plos One 7(4).
- 20. Bell JJ (2008) The functional roles of marine sponges. *Estuarine Coastal and Shelf Science* 79(3):341-353.
- 21. Taylor MW, Radax R, Steger D, & Wagner M (2007) Sponge-associated microorganisms: Evolution, ecology, and biotechnological potential. *Microbiology and Molecular Biology Reviews* 71(2):295-+.
- 22. Marliave JB, Conway KW, Gibbs DM, Lamb A, & Gibbs C (2009) Biodiversity and rockfish recruitment in sponge gardens and bioherms of southern British Columbia, Canada. *Marine Biology* 156(11):2247-2254.
- 23. Wulff J (2001) Assessing and monitoring coral reef sponges: Why and how? *Bulletin of Marine Science* 69(2):831-846.
- 24. Diaz MC & Rutzler K (2001) Sponges: An essential component of Caribbean coral reefs. *Bulletin of Marine Science* 69(2):535-546.
- 25. Vogel S (1977) Current-induced flow through living sponges in nature. *Proceedings of the National Academy of Sciences of the United States of America* 74(5):2069-2071.
- 26. Ribes M, Coma R, & Gili JM (1999) Natural diet and grazing rate of the temperate sponge Dysidea avara (Demospongiae, Dendroceratida) throughout an annual cycle. *Marine Ecology-Progress Series* 176:179-190.
- 27. Lesser MP (2006) Benthic-pelagic coupling on coral reefs: Feeding and growth of Caribbean sponges. *Journal of Experimental Marine Biology and Ecology* 328(2).
- 28. Yahel G, Whitney F, Reiswig HM, Eerkes-Medrano DI, & Leys SP (2007) In situ feeding and metabolism of glass sponges (Hexactinellida, Porifera) studied in a deep temperate fjord with a remotely operated submersible. *Limnology and Oceanography* 52(1).
- 29. Reiswig HM (1971) In situ pumping activities of tropical Demos-pongiae. *Mar Biol Berlin* 9.
- 30. de Goeij JM, *et al.* (2013) Surviving in a Marine Desert: The Sponge Loop Retains Resources Within Coral Reefs. *Science* 342(6154):108-110.
- 31. Yahel G, Sharp JH, Marie D, Hase C, & Genin A (2003) In situ feeding and element removal in the symbiont-bearing sponge Theonella swinhoei: Bulk DOC is the major source for carbon. *Limnology and Oceanography* 48(1).
- 32. Wulff JL (2006) Ecological interactions of marine sponges. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 84(2):146-166.
- 33. Cheshire AC & Wilkinson CR (1991) Modeling the photosynthetic production by sponges on davies-reef, great-barrier-reef. *Marine Biology* 109(1):13-18.
- 34. van Duyl FC, Hegeman J, Hoogstraten A, & Maier C (2008) Dissolved carbon fixation by sponge-microbe consortia of deep water coral mounds in the northeastern Atlantic Ocean. *Marine Ecology-Progress Series* 358.
- 35. Reiswig HM (1981) Partial carbon and energy budgets of the bacterio sponge verongiafistularis porifera demospongiae in barbados west-indieS. *Marine Ecology* 2(4).
- 36. Schroder HC, *et al.* (2004) Differentiation capacity of epithelial cells in the sponge Suberites domuncula. *Cell and Tissue Research* 316(2):271-280.

- 37. Leys S & Hill A (2012) The physiology and molecular biology of sponge tissues. *Adv Mar Biol* 62:1 56.
- 38. De Goeij JM, *et al.* (2009) Cell kinetics of the marine sponge Halisarca caerulea reveal rapid cell turnover and shedding. *Journal of Experimental Biology* 212(23):3892-3900.
- 39. Artis D (2008) Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. *Nature Reviews Immunology* 8(6):411-420.
- 40. Imsiecke G (1993) Ingestion, digestion, and egestion in spongilla-lacustris (porifera, spongillidae) after pulse feeding with chlamydomonas-reinhardtii (volvocales). *Zoomorphology* 113(4):233-244.
- 41. Willenz P, Vray B, Maillard MP, & Vandevyver G (1986) A quantitative study of the retention of radioactively labeled escherichia-coli by the fresh-water sponge ephydatia fluviatilis. *Physiological zoology* 59(5):495-504.
- 42. Willenz P & Vandevyver G (1986) Ultrastructural evidence of extruding exocytosis of residual bodies in the fresh-water sponge ephydatia-fluviatilis. *Journal of Morphology* 190(3):307-318.
- 43. Webster NS & Blackall LL (2009) What do we really know about sponge-microbial symbioses? *Isme Journal* 3(1):1-3.
- 44. Nichols SA, Dirks W, Pearse JS, & King N (2006) Early evolution of animal cell signaling and adhesion genes. *Proceedings of the National Academy of Sciences of the United States of America* 103(33):12451-12456.
- 45. Srivastava M, *et al.* (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* 466(7307):720-U723.
- 46. Sharp KH, Eam B, Faulkner DJ, & Haygood MG (2007) Vertical transmission of diverse microbes in the tropical sponge Corticium sp. *Applied and Environmental Microbiology* 73(2):622-629.
- 47. Schmitt S, *et al.* (2012) Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J* 6(3):564-576.
- 48. Taylor MW, Hill RT, Piel J, Thacker RW, & Hentschel U (2007) Soaking it up: the complex lives of marine sponges and their microbial associates. *Isme Journal* 1(3):187-190.
- 49. Lee YK, Lee JH, & Lee HK (2001) Microbial symbiosis in marine sponges. *Journal of Microbiology* 39(4):254-264.
- 50. Thomas T, *et al.* (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.*
- 51. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. Proceedings of the National Academy of Sciences of the United States of America 103(48):18296-18301.
- 52. Hallam SJ, *et al.* (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *Plos Biology* 4(4):520-536.
- 53. Hochmuth T, *et al.* (2010) Linking chemical and microbial diversity in marine sponges: Possible Role for Poribacteria as Producers of Methyl-Branched Fatty Acids. *Chembiochem* 11(18):2572-2578.
- 54. Liu MY, Kjelleberg S, & Thomas T (2011) Functional genomic analysis of an uncultured delta-proteobacterium in the sponge Cymbastela concentrica. *Isme Journal* 5(3).

- 55. Hentschel U, Piel J, Degnan SM, & Taylor MW (2012) Genomic insights into the marine sponge microbiome. *Nature Reviews Microbiology* 10(9).
- 56. Hentschel U, *et al.* (2003) Microbial diversity of marine sponges. *Prog Mol Subcell Biol* 37:59-88.
- 57. Giles EC, *et al.* (2012) Bacterial community profiles in low microbial abundance sponges. *FEMS Microbiology Ecology*:n/a-n/a.
- 58. Weisz JB, Lindquist N, & Martens CS (2008) Do associated microbial abundances impact marine demosponge pumping rates and tissue densities? *Oecologia* 155(2).
- 59. Grozdanov L & Hentschel U (2007) An environmental genomics perspective on the diversity and function of marine sponge-associated microbiota. *Current Opinion in Microbiology* 10(3):215-220.
- 60. Webster NS & Taylor MW (2012) Marine sponges and their microbial symbionts: love and other relationships. *Environmental Microbiology* 14(2):335-346.
- 61. Siegl A, *et al.* (2011) Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *Isme Journal* 5(1):61-70.
- 62. Weisz JB, Massaro AJ, Ramsby BD, & Hill MS (2010) Zooxanthellar Symbionts Shape Host Sponge Trophic Status Through Translocation of Carbon. *Biological Bulletin* 219(3):189-197.
- 63. Hentschel U, *et al.* (2002) Molecular evidence for a uniform microbial community in sponges from different oceans. *Applied and Environmental Microbiology* 68(9):4431-4440.
- 64. Fieseler L, Horn M, Wagner M, & Hentschel U (2004) Discovery of the novel candidate phylum "Poribacteria" in marine sponges. *Applied and Environmental Microbiology* 70(6):3724-3732.
- 65. Loker ES, Adema CM, Zhang SM, & Kepler TB (2004) Invertebrate immune systems not homogeneous, not simple, not well understood. *Immunological Reviews* 198:10-24.
- 66. Steger D, *et al.* (2008) Diversity and mode of transmission of ammonia-oxidizing archaea in marine sponges. *Environmental Microbiology* 10(4):1087-1094.
- 67. Bright M & Bulgheresi S (2010) A complex journey: transmission of microbial symbionts. *Nature Reviews Microbiology* 8(3):218-230.
- 68. Medzhitov R (2007) Recognition of microorganisms and activation of the immune response. *Nature* 449(7164):819-826.
- 69. Pallen MJ & Wren BW (2007) Bacterial pathogenomics. *Nature* 449(7164):835-842.
- 70. Zamioudis C & Pieterse CMJ (2012) Modulation of Host Immunity by Beneficial Microbes. *Molecular Plant-Microbe Interactions* 25(2):139-150.
- 71. Kereszt A, Mergaert P, Maroti G, & Kondorosi E (2011) Innate immunity effectors and virulence factors in symbiosis. *Current Opinion in Microbiology* 14(1):76-81.
- 72. Kopp EB & Medzhitov R (1999) The Toll-receptor family and control of innate immunity. *Current Opinion in Immunology* 11(1):13-18.
- 73. Kawai T & Akira S (2009) The roles of TLRs, RLRs and NLRs in pathogen recognition. *International Immunology* 21(4):317-337.
- 74. Athman R & Philpott D (2004) Innate immunity via Toll-like receptors and Nod proteins. *Current Opinion in Microbiology* 7(1):25-32.

- 75. Akira S, Uematsu S, & Takeuchi O (2006) Pathogen recognition and innate immunity. *Cell* 124(4):783-801.
- 76. Perovic-Ottstadt S, *et al.* (2004) A (1 -> 3)-beta-D-glucan recognition protein from the sponge Suberites domuncula Mediated activation of fibrinogen-like protein and epidermal growth factor gene expression. *European Journal of Biochemistry* 271(10):1924-1937.
- 77. Jarrell K, Ding Y, Nair D, & Siu S (2013) Surface Appendages of Archaea: Structure, Function, Genetics and Assembly. *Life* 3(1):86-117.
- 78. Ventura M, Turroni F, Motherway MO, MacSharry J, & van Sinderen D (2012) Hostmicrobe interactions that facilitate gut colonization by commensal bifidobacteria. *Trends in Microbiology* 20(10):467-476.
- 79. Schell MA, *et al.* (2002) The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. *Proceedings of the National Academy of Sciences of the United States of America* 99(22):14422-14427.
- 80. Liu CH, Lee SM, VanLare JM, Kasper DL, & Mazmanian SK (2008) Regulation of surface architecture by symbiotic bacteria mediates host colonization. *Proceedings of the National Academy of Sciences of the United States of America* 105(10):3951-3956.
- 81. Mazmanian SK & Kasper DL (2006) The love-hate relationship between bacterial polysaccharides and the host immune system. *Nature Reviews Immunology* 6(11):849-858.
- 82. Ventura M, et al. (2009) Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nature Reviews Microbiology* 7(1):61-U77.
- 83. Turroni F, *et al.* (2010) Characterization of the Serpin-Encoding Gene of Bifidobacterium breve 210B. *Applied and Environmental Microbiology* 76(10):3206-3219.
- 84. Ivanov D, *et al.* (2006) A serpin from the gut bacterium Bifidobacterium longum inhibits eukaryotic elastase-like serine proteases. *Journal of Biological Chemistry* 281(25):17246-17252.
- 85. Burg ND & Pillinger MH (2001) The neutrophil: Function and regulation in innate and humoral immunity. *Clinical Immunology* 99(1):7-17.
- 86. Holt RA, *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science* 298(5591):129-+.
- 87. Christophides GK, *et al.* (2002) Immunity-related genes and gene families in Anopheles gambiae. *Science* 298(5591):159-165.
- 88. Suwanchaichinda C & Kanost MR (2009) The serpin gene family in Anopheles gambiae. *Gene* 442(1-2):47-54.
- 89. Iwanaga S & Lee BL (2005) Recent advances in the innate immunity of invertebrate animals. *Journal of Biochemistry and Molecular Biology* 38(2):128-150.
- 90. Michel K, Budd A, Pinto S, Gibson TJ, & Kafatos FC (2005) Anopheles gambiae SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *Embo Reports* 6(9):891-897.
- 91. Fanning S, *et al.* (2012) Bifidobacterial surface-exopolysaccharide facilitates commensalhost interaction through immune modulation and pathogen protection. *Proceedings of the National Academy of Sciences of the United States of America* 109(6):2108-2113.
- 92. Molloy MJ, *et al.* (2013) Intraluminal Containment of Commensal Outgrowth in the Gut during Infection-Induced Dysbiosis. *Cell Host & Microbe* 14(3):318-328.

- 93. Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, & Medzhitov R (2004) Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* 118(2):229-241.
- 94. Chu HT & Mazmanian SK (2013) Innate immune recognition of the microbiota promotes host-microbial symbiosis. *Nature Immunology* 14(7):668-675.
- 95. Cario E & Podolsky DK (2005) Intestinal epithelial TOLLerance versus inTOLLerance of commensals. *Molecular Immunology* 42(8):887-893.
- 96. Kelly D, *et al.* (2004) Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-gamma and RelA. *Nature Immunology* 5(1):104-112.
- 97. Brestoff JR & Artis D (2013) Commensal bacteria at the interface of host metabolism and the immune system. *Nature Immunology* 14(7):676-684.
- 98. Kamke J, *et al.* (2013) Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges. *Isme Journal* 7(12):2287-2300.
- 99. Moissl-Eichinger C & Huber H (2011) Archaeal symbionts and parasites. *Current Opinion in Microbiology* 14(3):364-370.
- 100. Wrede C, Dreier A, Kokoschka S, & Hoppert M (2012) Archaea in Symbioses. *Archaeaan International Microbiological Journal*.
- 101. Ohkuma M, Noda S, & Kudo T (1999) Phylogenetic relationships of symbiotic methanogens in diverse termites. *Fems Microbiology Letters* 171(2):147-153.
- 102. Gill EE & Brinkman FSL (2011) The proportional lack of archaeal pathogens: Do viruses/phages hold the key? *Bioessays* 33(4):248-254.
- 103. Shiffman ME & Charalambous BM (2012) The search for archaeal pathogens. *Reviews in Medical Microbiology* 23(3):45-51.
- 104. Allen MA, *et al.* (2009) The genome sequence of the psychrophilic archaeon, Methanococcoides burtonii: the role of genome evolution in cold adaptation. *Isme Journal* 3(9):1012-1035.
- 105. Eichler J & Adams MWW (2005) Posttranslational protein modification in Archaea. *Microbiology and Molecular Biology Reviews* 69(3):393-+.
- 106. Bohm M, et al. (2001) Molecular response of the sponge Suberites domuncula to bacterial infection. *Marine Biology* 139(6):1037-1045.
- 107. Muller WEG & Muller IM (2003) Origin of the metazoan immune system: Identification of the molecules and their functions in sponges. pp 281-292.
- 108. Wiens M, *et al.* (2005) Innate immune defense of the sponge Suberites domuncula against bacteria involves a MyD88-dependent signaling pathway Induction of a perforin-like molecule. *Journal of Biological Chemistry* 280(30):27949-27959.
- 109. Wiens M, *et al.* (2007) Toll-like receptors are part of the innate immune defense system of sponges (Demospongiae : Porifera). *Molecular Biology and Evolution* 24(3):792-804.
- 110. Conaco C, *et al.* (2012) Transcriptome profiling of the demosponge Amphimedon queenslandica reveals genome-wide events that accompany major life cycle transitions. *Bmc Genomics* 13.
- 111. Gauthier MEA, Du Pasquier L, & Degnan BM (2010) The genome of the sponge Amphimedon queenslandica provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evolution & Development* 12(5):519-533.

- 112. Gauthier M & Degnan BM (2008) The transcription factor NF-kappa B in the demosponge Amphimedon queenslandica: insights on the evolutionary origin of the Rel homology domain. *Development Genes and Evolution* 218(1):23-32.
- 113. Perez-Porro AR, Navarro-Gomez D, Uriz MJ, & Giribet G (2013) A NGS approach to the encrusting Mediterranean sponge Crella elegans (Porifera, Demospongiae, Poecilosclerida): transcriptome sequencing, characterization and overview of the gene expression along three life cycle stages. *Molecular Ecology Resources* 13(3):494-509.
- 114. Riesgo A, *et al.* (2012) Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology* 9.
- 115. Riesgo A, Farrar N, Windsor PJ, Giribet G, & Leys SP (2014) Early evolution of molecular complexity in metazoans: an analysis of transcriptomes from all four sponge classes. *Integrative and Comparative Biology* 54:E339-E339.
- 116. Blumbach B, *et al.* (1999) Cloning and expression of new receptors belonging to the immunoglobulin superfamily from the marine sponge Geodia cydonium. *Immunogenetics* 49(9):751-763.
- 117. Wiens M, Kuusksalu A, Kelve M, & Muller WEG (1999) Origin of the interferoninducible (2 '-5 ')oligoadenylate synthetases: cloning of the (2 '-5 ')oligoadenylate synthetase from the marine sponge Geodia cydonium. *Febs Letters* 462(1-2):12-18.
- 118. Wiens M, Krasko A, Perovic S, & Muller WEG (2003) Caspase-mediated apoptosis in sponges: cloning and function of the phylogenetic oldest apoptotic proteases from metazoa. *Biochimica Et Biophysica Acta-Molecular Cell Research* 1593(2-3):179-189.
- 119. Pancer Z, Skorokhod A, Blumbach B, & Muller WEG (1998) Multiple Ig-like featuring genes divergent within and among individuals of the marine sponge Geodia cydonium. *Gene* 207(2):227-233.
- 120. Muller WEG, Blumbach B, & Muller IM (1999) Evolution of the innate and adaptive immune systems Relationships between potential immune molecules in the lowest metazoan phylum (Porifera) and those in vertebrates. *Transplantation* 68(9):1215-1227.
- 121. Kanzler H, Barrat FJ, Hessel EM, & Coffman RL (2007) Therapeutic targeting of innate immunity with Toll-like receptor agonists and antagonists. *Nature Medicine* 13(5):552-559.
- 122. Turvey SE & Hawn TR (2006) Towards subtlety: Understanding the role of Toll-like receptor signaling in susceptibility to human infections. *Clinical Immunology* 120(1):1-9.
- 123. Georgel P, Macquin C, & Bahram S (2009) The Heterogeneous Allelic Repertoire of Human Toll-Like Receptor (TLR) Genes. *Plos One* 4(11).
- 124. Yuen B, Bayes JM, & Degnan SM (2014) The Characterization of Sponge NLRs Provides Insight into the Origin and Evolution of This Innate Immune Gene Family in Animals. *Molecular Biology and Evolution* 31(1):106-120.
- 125. Nadiri A, Wolinski MK, & Saleh M (2006) The inflammatory caspases: Key players in the host response to pathogenic invasion and sepsis. *Journal of Immunology* 177(7):4239-4245.
- 126. Fink SL & Cookson BT (2005) Apoptosis, pyroptosis, and necrosis: Mechanistic description of dead and dying eukaryotic cells. *Infection and Immunity* 73(4):1907-1916.
- 127. Tsuji Y, *et al.* (2012) Sensing of Commensal Organisms by the Intracellular Sensor NOD1 Mediates Experimental Pancreatitis. *Immunity* 37(2):326-338.

- 128. Fritz JH, Ferrero RL, Philpott DJ, & Girardin SE (2006) Nod-like proteins in immunity, inflammation and disease. *Nature Immunology* 7(12):1250-1257.
- 129. Preston CM, Wu KY, Molinski TF, & DeLong EF (1996) A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen nov, sp, nov. *Proceedings of the National Academy of Sciences of the United States of America* 93(13):6241-6246.
- 130. Preston CM (1998) Prokaryotic Diversity in Marine Sponges: A Description of a Specific Association Between the Marine Archaeon, Cenarchaeum symbiosum, and the Marine Sponge, Axinella mexicana.
- 131. Holmes B & Blanch H (2007) Genus-specific associations of marine sponges with group I crenarchaeotes. *Marine Biology* 150(5):759-772.
- 132. Lee EY, Lee HK, Lee YK, Sim CJ, & Lee JH (2003) Diversity of symbiotic archaeal communities in marine sponges from Korea. 20:299-304.
- Margot H, Acebal C, Toril E, Amils R, & Puentes JLF (2002) Consistent association of crenarchaeal Archaea with sponges of the genus Axinella. *Marine Biology* 140(4):739-745.
- 134. Gazave E, *et al.* (2010) Polyphyly of the genus Axinella and of the family Axinellidae (Porifera: Demospongiaep). *Molecular Phylogenetics and Evolution* 57(1):35-47.
- 135. Schleper C, *et al.* (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon Cenarchaeum symbiosum. *Journal of Bacteriology* 180(19):5003-5009.
- 136. Scanlan PD, Shanahan F, & Marchesi JR (2008) Human methanogen diversity and incidence in healthy and diseased colonic groups using mcrA gene analysis. *Bmc Microbiology* 8.
- 137. Zhang HS, et al. (2009) Human gut microbiota in obesity and after gastric bypass. Proceedings of the National Academy of Sciences of the United States of America 106(7):2365-2370.
- 138. Jarvis GN, *et al.* (2000) Isolation and identification of ruminal methanogens from grazing cattle. *Current Microbiology* 40(5):327-332.
- 139. Hoffmann C, *et al.* (2013) Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *Plos One* 8(6).
- 140. Erwin PM, Pineda MC, Webster N, Turon X, & Lopez-Legentil S (2014) Down under the tunic: bacterial biodiversity hotspots and widespread ammonia-oxidizing archaea in coral reef ascidians. *Isme Journal* 8(3):575-588.
- 141. Erwin PM, Olson JB, & Thacker RW (2011) Phylogenetic Diversity, Host-Specificity and Community Profiling of Sponge-Associated Bacteria in the Northern Gulf of Mexico. *Plos One* 6(11).
- 142. Kamke J, Taylor MW, & Schmitt S (2010) Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. *Isme Journal* 4(4):498-508.
- 143. Bayer K, Kamke J, & Hentschel U (2014) Quantification of bacterial and archaeal symbionts in high and low microbial abundance sponges using real-time PCR. *FEMS Microbiology Ecology*:n/a-n/a.
- 144. Cassler M, *et al.* (2008) Use of real-time qPCR to quantify members of the unculturable heterotrophic bacterial community in a deep sea marine sponge, Vetulina sp. *Microbial Ecology* 55(3):384-394.

- 145. Hugoni M, *et al.* (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences*.
- 146. Sogin ML, *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* 103(32):12115-12120.
- 147. Schmitt S, *et al.* (2011) Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J* 6(3):564-576.
- 148. Fan L, *et al.* (2012) Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proceedings of the National Academy of Sciences* 109(27):E1878–E1887.
- 149. Lee OO, *et al.* (2011) Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. *ISME J* 5(4):650-664.
- 150. Turque AS, *et al.* (2010) Environmental Shaping of Sponge Associated Archaeal Communities. *PLoS ONE* 5(12):e15774.
- 151. Bayer K, Schmitt S, & Hentschel U (2008) Physiology, phylogeny and in situ evidence for bacterial and archaeal nitrifiers in the marine sponge Aplysina aerophoba. *Environmental Microbiology* 10(11):2942-2955.
- 152. Pape T, *et al.* (2006) Dense populations of Archaea associated with the demosponge Tentorium semisuberites Schmidt, 1870 from Arctic deep-waters. *Polar Biology* 29(8):662-667.
- 153. Lee OO, Wong YH, & Qian PY (2009) Inter- and Intraspecific Variations of Bacterial Communities Associated with Marine Sponges from San Juan Island, Washington. *Applied and Environmental Microbiology* 75(11):3513-3521.
- 154. Pfannkuchen M, Fritz GB, Schlesinger S, Bayer K, & Brummer F (2009) In situ pumping activity of the sponge Aplysina aerophoba, Nardo 1886. *Journal of Experimental Marine Biology and Ecology* 369(1):65-71.
- 155. Schlappy ML, Weber M, Mendola D, Hoffmann F, & Beer Dd (2010) Heterogeneous oxygenation resulting from active and passive flow in two Mediterranean sponges, Dysidea avara and Chondrosia reniformis. *Limnology and Oceanography* 55(3):1289-1300.
- 156. Yang Z & Li Z (2012) Spatial distribution of prokaryotic symbionts and ammoxidation, denitrifier bacteria in marine sponge Astrosclera willeyana. *Sci. Rep.* 2.
- 157. Ding B, Yin Y, Zhang FL, & Li ZY (2011) Recovery and Phylogenetic Diversity of Culturable Fungi Associated with Marine Sponges Clathrina luteoculcitella and Holoxea sp in the South China Sea. *Marine Biotechnology* 13(4):713-721.
- 158. Cerrano C, *et al.* (2000) Diatom invasion in the antarctic hexactinellid sponge Scolymastra joubini. *Polar Biology* 23(6):441-444.
- 159. Webster NS, Negri AP, Munro M, & Battershill CN (2004) Diverse microbial communities inhabit Antarctic sponges. *Environmental Microbiology* 6(3):288-300.
- 160. Maldonado M, Ribes M, & van Duyl FC (2012) Nutrient fluxes through sponges: biology, budgets, and ecological implications. *Advances in Sponge Science: Physiology, Chemical and Microbial Diversity, Biotechnology* 62:113-182.
- 161. Schmitt S, Wehrl M, Bayer K, Siegl A, & Hentschel U (2007) Marine sponges as models for commensal microbe-host interactions. *Symbiosis* 44(1-3):43-50.

- 162. Cavicchioli R, Curmi PMG, Saunders N, & Thomas T (2003) Pathogenic archaea: do they exist? *BioEssays* 25(11):1119-1128.
- 163. Pester M, Schleper C, & Wagner M (2011) The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Current Opinion in Microbiology* 14(3):300-306.
- 164. Spang A, *et al.* (2012) The genome of the ammonia-oxidizing Candidatus Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. *Environmental Microbiology* 14(12):3122-3145.
- 165. Mosier AC, Allen EE, Kim M, Ferriera S, & Francis CA (2012) Genome Sequence of "Candidatus Nitrosoarchaeum limnia" BG20, a Low-Salinity Ammonia-Oxidizing Archaeon from the San Francisco Bay Estuary. *Journal of Bacteriology* 194(8):2119-2120.
- 166. Spang A, *et al.* (2010) Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends in Microbiology* 18(8):331-340.
- 167. Hsiao W, Wan I, Jones SJ, & Brinkman FSL (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 19(3):418-420.
- 168. Ochman H & Moran NA (2001) Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* 292(5519):1096-1098.
- 169. Finan TM (2002) Evolving insights: Symbiosis islands and horizontal gene transfer. *Journal of Bacteriology* 184(11):2855-2856.
- 170. Mandel MJ, Wollenberg MS, Stabb EV, Visick KL, & Ruby EG (2009) A single regulatory gene is sufficient to alter bacterial host range. *Nature* 458(7235):215-U217.
- 171. Kim BK, et al. (2011) Genome Sequence of an Ammonia-Oxidizing Soil Archaeon, "Candidatus Nitrosoarchaeum koreensis" MY1. Journal of Bacteriology 193(19):5539-5540.
- 172. Walker CB, *et al.* (2010) Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proceedings of the National Academy of Sciences of the United States of America* 107(19):8818-8823.
- 173. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *Isme Journal* 2(10):1052-1065.
- 174. Keller A, Nesvizhskii AI, Kolker E, & Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry* 74(20):5383-5392.
- 175. Zybailov B, *et al.* (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *Journal of Proteome Research* 5(9):2339-2347.
- 176. Finn RD, et al. (2010) The Pfam protein families database. Nucleic Acids Research 38:D211-D222.
- 177. Kelley L & Sternberg M (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 4:363 371.
- 178. Larkin MA, et al. (2007) Clustal W and clustal X version 2.0. Bioinformatics 23(21):2947-2948.
- 179. Rutherford K, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10):944-945.
- 180. Krzywinski M, *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Research* 19(9):1639-1645.

- 181. Barber AE & Babbitt PC (2012) Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics* 28(21):2845-2846.
- 182. Brown SD & Babbitt PC (2012) Inference of Functional Properties from Large-scale Analysis of Enzyme Superfamilies. *Journal of Biological Chemistry* 287(1):35-42.
- 183. Baier F & Tokuriki N (2014) Connectivity between Catalytic Landscapes of the Metallobeta-Lactamase Superfamily. *Journal of Molecular Biology* 426(13):2442-2456.
- 184. Weidner M, Taupp M, & Hallam SJ (2010) Expression of Recombinant Proteins in the Methylotrophic Yeast Pichia pastoris. (36):e1862.
- 185. Tanaka S-i, Koga Y, Takano K, & Kanaya S (2011) Inhibition of chymotrypsin- and subtilisin-like serine proteases with Tk-serpin from hyperthermophilic archaeon Thermococcus kodakaraensis. *Biochimica et Biophysica Acta (BBA) Proteins & amp; Proteomics* 1814(2):299-307.
- 186. Schechter NM & Plotnick MI (2004) Measurement of the kinetic parameters mediating protease-serpin inhibition. *Methods* 32(2):159-168.
- 187. Buttigieg PL, *et al.* (2013) Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes. *PLoS ONE* 8(3):e50869.
- 188. Hicks MA, *et al.* (2011) The evolution of function in strictosidine synthase-like proteins. *Proteins-Structure Function and Bioinformatics* 79(11):3082-3098.
- 189. Congote LF (2007) Serpin A1 and CD91 as host instruments against HIV-1 infection: Are extracellular antiviral peptides acting as intracellular messengers? *Virus Research* 125(2):119-134.
- 190. Congote LF (2006) The C-terminal 26-residue peptide of serpin A1 is an inhibitor of HIV-1. *Biochemical and Biophysical Research Communications* 343(2):617-622.
- 191. Teh BA, *et al.* (2014) Structure to function prediction of hypothetical protein KPN_00953 (Ycbk) from Klebsiella pneumoniae MGH 78578 highlights possible role in cell wall metabolism. *Bmc Structural Biology* 14.
- 192. Galperin MY & Koonin EV (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Research* 32(18):5452-5463.
- 193. Thomas NA, Bardy SL, & Jarrell KF (2001) The archaeal flagellum: a different kind of prokaryotic motility structure. *Fems Microbiology Reviews* 25(2):147-174.
- 194. Ng SYM, Zolghadr B, Driessen AJM, Albers SV, & Jarrell KF (2008) Cell surface structures of archaea. *Journal of Bacteriology* 190(18):6039-6047.
- 195. Bhavsar AP, Guttman JA, & Finlay BB (2007) Manipulation of host-cell pathways by bacterial pathogens. *Nature* 449(7164):827-834.
- 196. Haiko J & Westerlund-Wikström B (2013) The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology* 2(4):1242-1267.
- 197. Zatakia HM, Nelson CE, Syed UJ, & Scharf BE (2014) ExpR Coordinates the Expression of Symbiotically Important, Bundle-Forming Flp Pili with Quorum Sensing in Sinorhizobium meliloti. *Applied and Environmental Microbiology* 80(8):2429-2439.
- 198. Wairuri CK, van der Waals JE, van Schalkwyk A, & Theron J (2012) Ralstonia solanacearum Needs Flp Pili for Virulence on Potato. *Molecular Plant-Microbe Interactions* 25(4):546-556.
- 199. Bahar O, Goffer T, & Burdman S (2009) Type IV Pili Are Required for Virulence, Twitching Motility, and Biofilm Formation of Acidovorax avenae subsp citrulli. *Molecular Plant-Microbe Interactions* 22(8):909-920.

- 200. Sauer FG, Mulvey MA, Schilling JD, Martinez JJ, & Hultgren SJ (2000) Bacterial pili: molecular mechanisms of pathogenesis. *Current Opinion in Microbiology* 3(1):65-72.
- 201. Nguyen M, Liu M, & Thomas T (2014) Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Molecular Ecology* 23(6):1635-1645.
- 202. Liu M, Fan L, Zhong L, Kjelleberg S, & Thomas T (2012) Metaproteogenomic analysis of a community of sponge symbionts. *Isme Journal* 6(8).
- 203. Finlay BB & McFadden G (2006) Anti-immunology: Evasion of the host immune system by bacterial and viral pathogens. *Cell* 124(4):767-782.
- 204. Le Clainche C & Drubin DG (2004) Actin lessons from pathogens. *Molecular Cell* 13(4):453-454.
- 205. Konzok A, *et al.* (1999) DAip1, a Dictyostelium homologue of the yeast actin-interacting protein 1, is involved in endocytosis, cytokinesis, and motility. *Journal of Cell Biology* 146(2):453-464.
- 206. Cerveny L, et al. (2013) Tetratricopeptide Repeat Motifs in the World of Bacterial Pathogens: Role in Virulence Mechanisms. *Infection and Immunity* 81(3):629-635.
- 207. Gao ZM, et al. (2014) Symbiotic Adaptation Drives Genome Streamlining of the Cyanobacterial Sponge Symbiont "Candidatus Synechococcus spongiarum". *Mbio* 5(2).
- 208. Liu DL, et al. (2005) Three-dimensional structure of the quorum-quenching N-acyl homoserine lactone hydrolase from Bacillus thuringiensis. Proceedings of the National Academy of Sciences of the United States of America 102(33):11882-11887.
- 209. Liu DL, *et al.* (2007) Structure and specificity of a quorum-quenching lactonase (AiiB) from Agrobacterium tumefaciens. *Biochemistry* 46(42):11789-11799.
- 210. Merone L, *et al.* (2010) Improving the promiscuous nerve agent hydrolase activity of a thermostable archaeal lactonase. *Bioresource Technology* 101(23):9204-9212.
- 211. Palzkill T (2013) Metallo-beta-lactamase structure and function. *Antimicrobial Therapeutics Reviews: the Bacterial Cell Wall as an Antimicrobial Target* 1277:91-104.
- 212. Daiyasu H, Osaka K, Ishino Y, & Toh H (2001) Expansion of the zinc metallo-hydrolase family of the beta-lactamase fold. *Febs Letters* 503(1):1-6.
- 213. Levashina EA, *et al.* (2001) Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. *Cell* 104(5):709-718.
- 214. Dodds AW & Law SKA (1998) The phylogeny and evolution of the thioester bondcontaining proteins C3, C4 and alpha(2)-macroglobulin. *Immunological Reviews* 166:15-26.
- 215. Armstrong PB (2006) Proteases and protease inhibitors: a balance of activities in hostpathogen interaction. *Immunobiology* 211(4):263-281.
- 216. Nonaka M (2011) The complement C3 protein family in invertebrates. *Isj-Invertebrate Survival Journal* 8(1):21-32.
- 217. Sekiguchi R, Fujito NT, & Nonaka M (2012) Evolution of the thioester-containing proteins (TEPs) of the arthropoda, revealed by molecular cloning of TEP genes from a spider, Hasarius adansoni. *Developmental and Comparative Immunology* 36(2):483-489.
- 218. Chaikeeratisak V, Somboonwiwat K, & Tassanakajon A (2012) Shrimp Alpha-2-Macroglobulin Prevents the Bacterial Escape by Inhibiting Fibrinolysis of Blood Clots. *Plos One* 7(10).

- 219. Budd A, Blandin S, Levashina EA, & Gibson TJ (2004) Bacterial alpha(2)macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome? *Genome Biology* 5(6).
- 220. Doan N & Gettins PGW (2008) α-Macroglobulins are present in some Gram-negative bacteria: characterization of the α2-macroglobulin from Escherichia coli. *Journal of Biological Chemistry* 283(42):28747-28756.
- 221. Rasmussen M, Muller HP, & Bjorck L (1999) Protein GRAB of Streptococcus pyogenes regulates proteolysis at the bacterial surface by binding alpha(2)-macroglobulin. *Journal of Biological Chemistry* 274(22):15336-15344.
- 222. Kantyka T, Rawlings ND, & Potempa J (2010) Prokaryote-derived protein inhibitors of peptidases: A sketchy occurrence and mostly unknown function. *Biochimie* 92(11):1644-1656.
- 223. Hedstrom L (2002) Serine protease mechanism and specificity. *Chemical Reviews* 102(12):4501-4523.
- 224. Irving JA, et al. (2002) Serpins in prokaryotes. *Molecular Biology and Evolution* 19(11):1881-1890.
- 225. Roberts TH, Hejgaard J, Saunders NFW, Cavicchioli R, & Curmi PMG (2004) Serpins in unicellular Eukarya, Archaea, and Bacteria: Sequence analysis and evolution. *Journal of Molecular Evolution* 59(4):437-447.
- 226. Gooptu B & Lomas DA (2009) Conformational Pathology of the Serpins: Themes, Variations, and Therapeutic Strategies. *Annual Review of Biochemistry* 78:147-176.
- 227. Kaiserman D, Whisstock J, C., & Bird P, I. (2006) Mechanisms of serpin dysfunction in disease. *Expert Reviews in Molecular Medicine* 8(31):1-19.
- 228. Gettins PGW (2002) Serpin structure, mechanism, and function. *Chemical Reviews* 102(12):4751-4803.
- 229. Silverman GA, *et al.* (2001) The serpins are an expanding superfamily of structurally similar but functionally diverse proteins Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *Journal of Biological Chemistry* 276(36):33293-33296.
- 230. Alvarez-Martin P, et al. (2012) A Two-Component Regulatory System Controls Autoregulated Serpin Expression in Bifidobacterium breve UCC2003. Applied and Environmental Microbiology 78(19):7032-7041.
- 231. Gettins PGW & Olson ST (2009) Exosite Determinants of Serpin Specificity. *Journal of Biological Chemistry* 284(31):20441-20445.
- 232. Deshmane SL, Kremlev S, Amini S, & Sawaya BE (2009) Monocyte Chemoattractant Protein-1 (MCP-1): An Overview. *Journal of Interferon and Cytokine Research* 29(6):313-326.
- 233. Easton DM, Nijnik A, Mayer ML, & Hancock REW (2009) Potential of immunomodulatory host defense peptides as novel anti-infectives. *Trends in Biotechnology* 27(10):582-590.
- 234. Nijnik A, *et al.* (2010) Synthetic Cationic Peptide IDR-1002 Provides Protection against Bacterial Infections through Chemokine Induction and Enhanced Leukocyte Recruitment. *Journal of Immunology* 184(5):2539-2550.

- 235. Congote LF (2008) Multi-Functional Anti-HIV Agents Based on Amino Acid Sequences Present in Serpin C-Terminal Peptides. Anti-Infective Agents in Medicinal Chemistry (Formerly 'Current Medicinal Chemistry - Anti-Infective Agents) 7(2):126-133.
- 236. Congote LF & Temmel N (2004) The C-terminal 26-residue peptide of serpin A1 stimulates proliferation of breast and liver cancer cells: role of protein kinase C and CD47. *Febs Letters* 576(3):343-347.
- 237. Koskimaki JE, *et al.* (2012) Serpin-Derived Peptides Are Antiangiogenic and Suppress Breast Tumor Xenograft Growth. *Translational Oncology* 5(2):92-97.
- 238. Medzhitov R & Janeway CJ (2000) Advances in immunology: Innate immunity. *New England Journal of Medicine* 343(5):338-344.
- 239. Robertson SJ & Girardin SE (2013) Nod-like receptors in intestinal host defense: controlling pathogens, the microbiota, or both? *Current Opinion in Gastroenterology* 29(1):15-22.
- 240. Muller WEG, *et al.* (2001) Contribution of sponge genes to unravel the genome of the hypothetical ancestor of Metazoa (Urmetazoa). *Gene* 276(1-2):161-173.
- 241. Muller WEG (1995) Molecular phylogeny of Metazoa (animals) monophyletic origin. *Naturwissenschaften* 82(7):321-329.
- 242. Zilberberg C, *et al.* (2008) Innate immune response and potential biomarkers of Suberites domuncula (Porifera : Demospongiae) after exposure to gram-negative bacteria. *Marine Environmental Research* 66(1):174-175.
- 243. Wang XJ & Lavrov DV (2007) Mitochondrial genome of the homoscleromorph oscarella carmela (Porifera, Demospongiae) reveals unexpected complexity in the common ancestor of sponges and other animals. *Molecular Biology and Evolution* 24(2):363-373.
- 244. Lange C, *et al.* (2011) Defining the Origins of the NOD-Like Receptor System at the Base of Animal Evolution. *Molecular Biology and Evolution* 28(5):1687-1702.
- 245. Hibino T, *et al.* (2006) The immune gene repertoire encoded in the purple sea urchin genome. *Developmental Biology* 300(1):349-365.
- 246. Rast JP, Smith LC, Loza-Coll M, Hibino T, & Litman GW (2006) Review Genomic insights into the immune system of the sea urchin. *Science* 314(5801):952-956.
- 247. Riesgo A, *et al.* (2014) Transcriptomic analysis of differential host gene expression upon uptake of symbionts: a case study with Symbiodinium and the major bioeroding sponge Cliona varians. *BMC Genomics* 15(1):376.
- 248. Harcet M, *et al.* (2010) Demosponge EST Sequencing Reveals a Complex Genetic Toolkit of the Simplest Metazoans. *Molecular Biology and Evolution* 27(12):2747-2756.
- 249. Bohm M, Schroder HC, Muller IM, Muller WEG, & Gamulin V (2000) The mitogenactivated protein kinase p38 pathway is conserved in metazoans: Cloning and activation of p38 of the SAPK2 subfamily from the sponge Suberites domuncula. *Biology of the Cell* 92(2):95-104.
- 250. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, & Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research* 21(9):1552-1560.
- 251. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.
- 252. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research* 19(6):1117-1123.

- 253. Birol I, et al. (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25(21):2872-2877.
- 254. Konwar KM, Hanson NW, Page AP, & Hallam SJ (2013) MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *Bmc Bioinformatics* 14.
- 255. Hyatt D, *et al.* (Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
- 256. Kanehisa M & Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1):27-30.
- 257. Tatusov RL, *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29(1):22-28.
- 258. Pruitt KD & Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* 29(1):137-140.
- 259. Lynn DJ, et al. (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular Systems Biology* 4.
- 260. Rasko DA, Myers GSA, & Ravel J (2005) Visualization of comparative genomic analyses by BLAST score ratio. *Bmc Bioinformatics* 6.
- 261. Finn RD, et al. (2014) Pfam: the protein families database. Nucleic Acids Research 42(D1):D222-D230.
- 262. Mitra S, *et al.* (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *Bmc Bioinformatics* 12.
- 263. Puente XS, Sanchez LM, Overall CM, & Lopez-Otin C (2003) Human and mouse proteases: A comparative genomic approach. *Nature Reviews Genetics* 4(7):544-558.
- 264. Jutras I & Desjardins M (2005) Phagocytosis: At the crossroads of innate and adaptive immunity. *Annual Review of Cell and Developmental Biology*, Annual Review of Cell and Developmental Biology, Vol 21, pp 511-527.
- 265. Luthringer B, *et al.* (2010) Poriferan survivin exhibits a conserved regulatory role in the interconnected pathways of cell cycle and apoptosis. *Cell Death Differ*.
- 266. WagnerHulsmann C, *et al.* (1996) A galectin links the aggregation factor to cells in the sponge (Geodia cydonium) system. *Glycobiology* 6(8):785-793.
- 267. Sato S, St-Pierre C, Bhaumik P, & Nieminen J (2009) Galectins in innate immunity: dual functions of host soluble beta-galactoside-binding lectins as damage-associated molecular patterns (DAMPs) and as receptors for pathogen-associated molecular patterns (PAMPs). *Immunological Reviews* 230:172-187.
- 268. Kumar H, Kawai T, & Akira S (2009) Toll-like receptors and innate immunity. *Biochemical and Biophysical Research Communications* 388(4):621-625.
- 269. Farhat K, *et al.* (2008) Heterodimerization of TLR2 with TLR1 or TLR6 expands the ligand spectrum but does not lead to differential signaling. *Journal of Leukocyte Biology* 83(3):692-701.
- 270. Shaw PJ, Lamkanfi M, & Kanneganti TD (2010) NOD-like receptor (NLR) signaling beyond the inflammasome. *European Journal of Immunology* 40(3):624-627.
- 271. Bosch TCG, *et al.* (2009) Uncovering the evolutionary history of innate immunity: The simple metazoan Hydra uses epithelial cells for host defence. *Developmental and Comparative Immunology* 33(4):559-569.

- 272. Stuart LM & Ezekowitz RAB (2005) Phagocytosis: Elegant complexity. *Immunity* 22(5):539-550.
- 273. Desjardins M, Houde M, & Gagnon E (2005) Phagocytosis: the convoluted way from nutrition to adaptive immunity. *Immunological Reviews* 207:158-165.
- 274. Stuart LM & Ezekowitz RA (2008) Phagocytosis and comparative innate immunity: learning on the fly. *Nature Reviews Immunology* 8(2):131-141.
- 275. Gronlien HK, Berg T, & Lovlie AM (2002) In the polymorphic ciliate Tetrahymena vorax, the non-selective phagocytosis seen in microstomes changes to a highly selective process in macrostomes. *Journal of Experimental Biology* 205(14):2089-2097.
- 276. Plattner H (2010) Membrane trafficking in Protozoa: SNARE proteins, H+-ATPase, actin, and other key players in ciliates. *International Review of Cell and Molecular Biology, Vol 280* 280:79-184.
- 277. Williams MJ (2007) Drosophila hemopoiesis and cellular immunity. *Journal of Immunology* 178(8):4711-4716.
- 278. Mydlarz LD, Jones LE, & Harvell CD (2006) Innate immunity environmental drivers and disease ecology of marine and freshwater invertebrates. *Annual Review of Ecology Evolution and Systematics* 37:251-288.
- 279. Ghosh J, et al. (2011) Invertebrate immune diversity. Developmental and Comparative Immunology 35(9):959-974.
- 280. Libro S, Kaluziak ST, & Vollmer SV (2013) RNA-seq Profiles of Immune Related Genes in the Staghorn Coral Acropora cervicornis Infected with White Band Disease. *Plos One* 8(11).
- 281. Leys SP & Eerkes-Medrano DI (2006) Feeding in a calcareous sponge: Particle uptake by pseudopodia. *Biological Bulletin* 211(2):157-171.
- 282. Garcia-Garcia E & Rosales C (2002) Signal transduction during Fc receptor-mediated phagocytosis. *Journal of Leukocyte Biology* 72(6):1092-1108.
- 283. Fujita T (2002) Evolution of the lectin-complement pathway and its role in innate immunity. *Nature Reviews Immunology* 2(5):346-353.
- 284. Vasta GR, Quesenberry M, Ahmed H, & O'Leary N (1999) C-type lectins and galectins mediate innate and adaptive immune functions: their roles in the complement activation pathway. *Developmental and Comparative Immunology* 23(4-5):401-420.
- 285. Glick D, Barth S, & Macleod KF (2010) Autophagy: cellular and molecular mechanisms. *Journal of Pathology* 221(1):3-12.
- 286. Duszenko M, et al. (2011) Autophagy in protists. Autophagy 7(2):127-158.
- 287. Suzuki K & Ohsumi Y (2007) Molecular machinery of autophagosome formation in yeast, Saccharomyces cerevisiae. *Febs Letters* 581(11):2156-2161.
- 288. Takehana A, *et al.* (2002) Overexpression of a pattern-recognition receptor, peptidoglycan-recognition protein-LE, activates imd/relish-mediated antibacterial defense and the prophenoloxidase cascade in Drosophila larvae. *Proceedings of the National Academy of Sciences of the United States of America* 99(21):13705-13710.
- 289. Bortoluci KR & Medzhitov R (2010) Control of infection by pyroptosis and autophagy: role of TLR and NLR. *Cellular and Molecular Life Sciences* 67(10):1643-1651.
- 290. Delgado M, et al. (2009) Autophagy and pattern recognition receptors in innate immunity. *Immunological Reviews* 227:189-202.

- 291. Yano T & Kurata S (2011) Intracellular recognition of pathogens and autophagy as an innate immune host defence. *Journal of Biochemistry* 150(2):143-149.
- 292. Espert L, Codogno P, & Biard-Piechaczyk M (2007) Involvement of autophagy in viral infections: antiviral function and subversion by viruses. *Journal of Molecular Medicine-Jmm* 85(8):811-823.
- 293. Choy A & Roy CR (2013) Autophagy and bacterial infection: an evolving arms race. *Trends in Microbiology* 21(9):451-456.
- 294. Kawai T & Akira S (2011) Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. *Immunity* 34(5):637-650.
- 295. Kufer TA & Sansonetti PJ (2011) NLR functions beyond pathogen recognition. *Nature Immunology* 12(2):121-128.
- 296. Nakahira K, *et al.* (2011) Autophagy proteins regulate innate immune responses by inhibiting the release of mitochondrial DNA mediated by the NALP3 inflammasome. *Nature Immunology* 12(3):222-U257.
- 297. Kostic AD, Howitt MR, & Garrett WS (2013) Exploring host-microbiota interactions in animal models and humans. *Genes & Development* 27(7):701-718.
- 298. McFall-Ngai MJ & Ruby EG (2000) Developmental biology in marine invertebrate symbioses. *Current Opinion in Microbiology* 3(6):603-607.
- 299. Chandler JA, Lang JM, Bhatnagar S, Eisen JA, & Kopp A (2011) Bacterial Communities of Diverse Drosophila Species: Ecological Context of a Host-Microbe Model System. *Plos Genetics* 7(9).
- 300. Eckburg PB, *et al.* (2005) Diversity of the human intestinal microbial flora. *Science* 308(5728):1635-1638.
- Bakker MG, Schlatter DC, Otto-Hanson L, & Kinkel LL (2014) Diffuse symbioses: roles of plant-plant, plant-microbe and microbe-microbe interactions in structuring the soil microbiome. *Molecular Ecology* 23(6):1571-1583.
- 302. Wang J, *et al.* (2014) Modulation of gut microbiota during probiotic-mediated attenuation of metabolic syndrome in high fat diet-fed mice. *ISME J*.
- 303. Reigstad CS & Kashyap PC (2013) Beyond phylotyping: understanding the impact of gut microbiota on host biology. *Neurogastroenterology and Motility* 25(5):358-372.
- 304. Lim GE & Haygood MG (2004) "Candidatus endobugula glebosa," a specific bacterial symbiont of the marine bryozoan Bugula simplex. *Applied and Environmental Microbiology* 70(8):4921-4929.
- 305. Lim-Fong GE, Regali LA, & Haygood MG (2008) Evolutionary relationships of "Candidatus Endobugula" bacterial symbionts and their Bugula bryozoan hosts. *Applied and Environmental Microbiology* 74(11):3605-3609.
- 306. Sharp KH, Davidson SK, & Haygood MG (2007) Localization of 'Candidatus Endobugula sertula' and the bryostatins throughout the life cycle of the bryozoan Bugula neritina. *Isme Journal* 1(8):693-702.
- 307. McCutcheon JP & Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology* 10(1):13-26.
- 308. Moran NA, McLaughlin HJ, & Sorek R (2009) The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria. *Science* 323(5912):379-382.

- 309. Sloan DB & Moran NA (2012) Genome Reduction and Co-evolution between the Primary and Secondary Bacterial Symbionts of Psyllids. *Molecular Biology and Evolution* 29(12):3781-3792.
- 310. Otero-Gonzalez AJ, *et al.* (2010) Antimicrobial peptides from marine invertebrates as a new frontier for microbial infection control. *Faseb Journal* 24(5):1320-1334.
- 311. O'Hara AM & Shanahan F (2006) The gut flora as a forgotten organ. *Embo Reports* 7(7):688-693.
- 312. Behnam F, Vilcinskas A, Wagner M, & Stoecker K (2012) A Straightforward DOPE (Double Labeling of Oligonucleotide Probes)-FISH (Fluorescence In Situ Hybridization) Method for Simultaneous Multicolor Detection of Six Microbial Populations. *Applied and Environmental Microbiology* 78(15):5138-5142.
- 313. Chaurand P, Schwartz SA, & Caprioli RM (2002) Imaging mass spectrometry: a new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. *Current Opinion in Chemical Biology* 6(5):676-681.
- 314. Osinga R, Tramper J, & Wijffels RH (1999) Cultivation of marine sponges. *Marine Biotechnology* 1(6):509-532.
- 315. Rinkevich B (2005) Marine invertebrate cell cultures: New millennium trends. *Marine Biotechnology* 7(5):429-439.
- 316. Pomponi SA (2006) Biology of the Porifera: cell culture. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 84(2):167-174.
- 317. Rivera AS, *et al.* (2011) RNA interference in marine and freshwater sponges: actin knockdown in Tethya wilhelma and Ephydatia muelleri by ingested dsRNA expressing bacteria. *Bmc Biotechnology* 11.

Appendices

Appendix A Indicator species analysis taxonomic composition and abundance of indicator OTUs

Phylum	Terminal taxon	#	Indicator OTUs (b	est indicator IV, p)		Relative al	oundance o	f indicator	OTUs (%)
i nyium		Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Thaumarchaeoto	Candidatus Nitrosopumilus Cenarchaeum	0	1 (75, 0.0036)	1 (66.7, 0.0096)	0	0.00	0.02	0.01	0.00
Thaumarchaeota	Marine Group I	0	3 (99.6, 0.001)	0	0	0.00	2.59	0.00	0.01
	Acidimicrobiales TM214	0	1 (75, 0.002)	0	0	0.00	0.01	0.00	0.00
Actinobacteria	Acidimicrobiales	3 (99.1, 0.0002)	1 (67.4, 0.0058)	0	0	0.73	0.02	0.00	0.00
	Actinobacteria PeM15	0	1 (75, 0.002)	0	0	0.00	0.02	0.00	0.00
	Flexibacter polymorphus Fubivirga	0	2 (74 4 0 0038)	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Cytophagales TAA-5-07	1 (66.7, 0.008)	2 (74.4, 0.0058)	0	0	0.00	0.00	0.00	0.01
	Algibacter	0	1 (70.2, 0.0038)	0	0	0.00	0.01	0.00	0.00
	Lutibacter	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Flavobacteriaceae bacterium K8-13	1 (66.7, 0.0092)	0	0	0	0.04	0.00	0.00	0.00
	Merialanimaribacter Mesoflavibacter	0	0	1 (66.9, 0.0062)	0	0.00	0.00	0.01	0.00
Bacteroidetes	Flavobacteriaceae NS4 marine group	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Flavobacteriaceae NS5 marine group	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Polaribacter sp. J2-11	0	1 (75.1, 0.0064)	0	0	0.01	0.04	0.00	0.00
	Polaribacter sponge bacterium Zo9	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Psychroserpens Tenacibaculum sp. MGP-74/AN6	0	0	1 (69.5, 0.0084)	0	0.00	0.00	0.01	0.00
	Flavobacteriaceae	2 (76.1, 0.0038)	1 (75, 0.0038)	2 (71.5 0.0082)	0	0.13	0.05	0.04	0.02
	Saprospiraceae	0	0	2 (92.5, 0.0004)	0	0.00	0.00	0.03	0.00
	Parachlamydia acanthamoebae	0	1 (75, 0.0028)	0	0	0.00	0.02	0.00	0.00
Chlamydiae	Candidatus Fritschea eriococci	0	0	3 (78.7, 0.0014)	0	0.00	0.00	0.02	0.00
	Cyanobacteria Chloroplast	2 (85 6 0 0002)	1 (75 0 0036)	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
Cyanobacteria	Synechococcus	0	20 (100, 0.0002)	0	0	1.09	11.48	0.01	0.00
	Cyanobacteria SubsectionIII FamilyI	0	1 (68.8, 0.0098)	0	0	0.00	0.02	0.00	0.00
	Caminicella	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
Firmiontos	Clostridium sp. DY192	1 (83.3, 0.0022)	0	0	0	0.01	0.00	0.00	0.00
Finiticutes	Seaimentibacter Acidaminobacter	0	0	1 (66.7, 0.0096) 3 (66.7, 0.0096)	0	0.00	0.00	0.02	0.00
	Fusibacter	0	Ő	52 (75, 0.0042)	0	0.00	0.00	1.69	0.25
	Blastopirellula	0	12 (100, 0.0002)	0	0	0.01	0.29	0.01	0.01
	Planctomyces	0	3 (75, 0.0038)	0	0	0.00	0.05	0.00	0.00
Planctomycetes	Rhodopirellula	0	10 (90.3, 0.0002)	0	0	0.03	0.74	0.02	0.04
	Planctomycetaceae	0	1(70, 0.007) 11(100, 0.0002)	0	0	0.00	4 47	0.00	0.00
	Hellea	0	1 (75, 0.0038)	Ő	0	0.00	0.08	0.00	0.00
	OCS116 clade	0	2 (79.2, 0.0056)	0	0	0.25	0.93	0.00	0.08
	Filomicrobium	0	2 (91.3, 0.0002)	0	0	0.00	0.04	0.00	0.00
	Hyphomicrobiaceae Phyllohaetoriaceae hastorium AMV1	0	1 (100, 0.0002)	0	0	0.00	0.03	0.00	0.00
	Ahrensia	0	2 (75 0.0038)	0	0	0.00	0.00	0.08	0.00
	Phyllobacteriaceae	Ő	6 (75, 0.0038)	Ő	0	0.01	1.98	0.00	0.00
	Rhizobium sp. BZ3	0	0	1 (99.6, 0.0002)	0	0.00	0.00	0.51	0.00
	Rhodobium	0	3 (73.1, 0.0052)	0	0	0.00	0.07	0.00	0.00
	Dinoroseobacter Donghicola	0	1(75, 0.002) 1(815, 0.0054)	0	0	0.00	0.02	0.00	0.00
	Leisingera	0	0	1 (66 7 0 0096)	0	0.00	0.07	0.00	0.01
	Phaeobacter	0	1 (82.4, 0.002)	1 (66.7, 0.0096)	0	0.03	0.13	0.02	0.00
	Rhodobacter	0	1 (68.4, 0.0032)	0	0	0.00	0.03	0.00	0.00
	Rhodothalassium	2 (83.3, 0.0036)	2 (75, 0.0038)	0	0	0.15	0.41	0.00	0.00
(2 Roseobacter clade AS-21 lineage	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Roseobacter clade NAC11-7 lineage	1 (66 0 0098)	1 (75 0 0038)	2 (00.7, 0.0090)	1 (73 6 0 007)	0.00	0.00	0.01	0.00
	Roseobacter clade OCT lineage	0	1 (75, 0.0038)	0	0	0.00	0.01	0.00	0.00
	Roseovarius	0	1 (75, 0.0038)	0	0	0.00	0.02	0.00	0.00
	Ruegeria	0	4 (87.2, 0.0008)	0	0	0.02	0.25	0.00	0.00
	Sulfitobacter	0	1 (70, 0.0046)	0	0	0.00	0.01	0.00	0.00
	Rhodobacteraceae	0	4 (75, 0 002)	2 (66.7. 0 0096)	1 (72.0, 0.0026)	0.00	0.00	0.00	0.05
	Rhodospirillales KCM-B-15	0	1 (75, 0.0038)	0	Ő	0.00	0.09	0.00	0.00
	Pelagibius	0	2 (75, 0.0038)	0	0	0.00	0.41	0.00	0.00
	Rhodospirillaceae	0	5 (95.4, 0.0052)	1 (100, 0.0006)	0	0.00	0.49	0.03	0.01
	Sneathiella	0	1 (75, 0.0038)	0	0	0.00	0.07	0.00	0.00
	Altererythrobacter Sphingomonas	0	U 1 (68 2 0 0086)	1 (00.7, 0.0096) 0	0	0.00	0.00	0.01	0.00
	Sphingopyxis	0	1 (75, 0.0038)	0	0	0.00	0.32	0.00	0.00
	Comamonadaceae BAL58 marine group	0	0	1 (91.1, 0.0008)	Õ	0.00	0.00	0.06	0.00
	Comamonas	0	1 (65.1, 0.0098)	0	0	0.00	0.02	0.00	0.00
	Pelomonas	0	4 (69.4, 0.0054)	0	0	0.04	0.17	0.01	0.02
1	Nitrosomonas	0	U 2 (75 0 0028)	1 (86.6, 0.0024)	0	0.00	0.00	0.03	0.00
	1111 050monus	U	2 (75, 0.0058)	0	0	0.00	0.43	0.00	0.00

Dhulum	Touriseltours	i	# Indicator OTUs (be	st indicator IV, p)		Relative at	oundance of	indicator	OTUs (%)
Phylum	Terminal taxon	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
	P Nitrosomonadaceae	1 (83.3, 0.0034)	6 (75, 0.0038)	0	0	0.03	1.06	0.00	0.00
	^P Betaproteobacteria oca12	0	1 (72.5, 0.0078)	0	0	0.00	0.08	0.00	0.00
	Bdellovibrio	0	2 (75, 0.0038)	0	0	0.00	0.05	0.00	0.00
	δ Desulfopila	1 (66.7, 0.0078)	0	0	0	0.05	0.00	0.00	0.00
	Haliangium	0	0	0	1 (66.7, 0.0092)	0.00	0.00	0.00	0.01
	Candidatus Endobugula	25 (100, 0.0004)	0	0	0	3.83	0.02	0.00	0.00
	Alteromonadaceae	1 (85.7, 0.0004)	0	0	0	0.03	0.01	0.00	0.00
	Colwellia	0	0	6 (79.6, 0.0002)	0	0.00	0.00	0.17	0.04
	Colwellia sp. KMD002	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.02	0.00
	Pseudoalteromonas	0	1 (75, 0.0038)	0	0	0.00	0.02	0.00	0.00
	Nitrosococcus	1 (66.7, 0.0074)	0	0	0	0.01	0.00	0.00	0.00
	Thiolamprovum	0	1 (75, 0.0038)	0	0	0.00	0.06	0.00	0.00
	Thioalkalispira	1 (66.7, 0.0068)	0	0	0	0.02	0.00	0.00	0.00
	Granulosicoccus	0	1 (75, 0.0036)	0	0	0.00	0.02	0.00	0.00
	Gammaproteobacteria E01-9C-26 marine	12 (83.3, 0.0022)	1 (75, 0.0038)	0	0	2.89	0.09	0.00	0.00
	Gammaproteobacteria EC3	0	1 (75, 0.0028)	0	0	0.00	0.01	0.00	0.00
	Gammaproteobacteria HOC36	0	1 (75, 0.0038)	0	0	0.00	0.19	0.00	0.00
	Coxiella	0	1 (93.3, 0.0008)	3 (66.7, 0.0096)	0	0.00	0.03	0.02	0.00
	Legionella	0	5 (99.2, 0.0002)	2 (66.7, 0.0096)	0	0.02	1.40	0.01	0.00
	Legionella adelaidensis	0	1 (72.3, 0.0036)	0	0	0.00	0.05	0.00	0.00
D (1 ()	Legionellaceae	0	0	1 (70.0, 0.0062)	0	0.00	0.00	0.01	0.00
Proteobacteria	Methylococcales IheB2-23	0	1 (75, 0.0036)	0	0	0.00	0.03	0.00	0.00
	Methylomonas	0	1 (74.8, 0.0052)	0	0	0.00	0.28	0.00	0.00
	γ Methylococcales pItb-vmat-59	0	1 (71.9, 0.0048)	0	0	0.00	0.05	0.00	0.00
	Endozoicomonas	0	0	2 (93.8, 0.0002)	0	0.00	0.00	0.23	0.02
	Oceanospirillales J8P41000-1F04	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Oceanospirillales MBAE14	1 (100, 0.0004)	0	0	0	0.01	0.00	0.00	0.00
	Amphritea sp. MEBiC05461T	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Neptuniibacter	0	0	1 (100, 0.0006)	0	0.00	0.00	0.01	0.00
	Neptunomonas sp. 0536	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Pseudospirillum	3 (100, 0.0004)	0	0	0	0.24	0.01	0.00	0.00
	Oceanospirillales OM182 clade	0	1 (75, 0.0038)	0	0	0.00	0.02	0.00	0.00
	SAR86 clade	0	1 (69.9, 0.006)	0	0	0.00	0.08	0.00	0.00
	Arenicella	0	0	1 (88.1, 0.0006)	0	0.00	0.00	0.01	0.00
	Gammaproteobacteria OXIC-003	0	3 (74.6, 0.007)	0	0	0.01	1.77	0.00	0.00
	Acinetobacter baumannii	0	1 (79.4, 0.0034)	0	0	0.00	0.02	0.00	0.00
	Salinisphaeraceae	0	1 (75, 0.0038)	0	0	0.00	0.11	0.00	0.00
	Piscirickettsiaceae	0	0	4 (83.2, 0.0018)	0	0.00	0.00	0.08	0.01
	Leucothrix	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Photobacterium	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
	Stenotrophomonas	0	1 (100, 0.0002)	0	0	0.00	0.02	0.00	0.00
	ε Arcobacter	0	0	1 (85, 0.0008)	0	0.00	0.00	0.02	0.00
	Turneriella	0	2 (75, 0.0038)	0	0	0.00	0.87	0.00	0.00
0 . 1 .	Leptospiraceae	2 (100, 0.0004)	1 (75, 0.0038)	0	0	0.07	0.55	0.00	0.00
Spirochaetes	Spirochaeta	3 (100, 0.0004)	5 (75, 0.0038)	0	0	0.35	1.13	0.00	0.00
	Spirochaetaceae	0	2 (75, 0.0038)	0	0	0.00	0.04	0.00	0.00
TM6	TM6	0	0	1 (66.7. 0.0096)	0	0.00	0.00	0.01	0.00
	Opitutae vadinHA64	0	0	1 (66.7, 0.0096)	0	0.00	0.00	0.01	0.00
Verrucomicrobi	a Rubritalea spongiae	0	1 (70.4, 0.0092)	0	0	0.08	0.15	0.00	0.00
	Roseibacillus	0	1 (65.9. 0.0092)	0	0	0.00	0.02	0.00	0.00
	Rhytidocystis	13 (100, 0,0004)	0	0	0	0.93	0.14	0.00	0.00
SAR supergrou	P Syndiniales Group I	1 (76, 0,0016)	2 (93,2,0,0004)	0	0	0.03	0.03	0.00	0.00
1 8	Thraustochytriaceae E170	0	0	1 (84.9, 0.0004)	0	0.00	0.00	0.02	0.00
Porifera	Demospongiae*	113 (100, 0.0004)	57 (100, 0.0002)	88 (100, 0.0006)	0	30.16	8.32	38.07	0.46

* Sponge taxonomy collapsed to Class level due to missing Order to Genus levels of classification in Silva taxonomy

Appendix B

B.1 Comparison of "A" and "B" type genomic sequences

To compare the genomes of different *C. symbiosum* populations, 97 "B" type fosmids were sequenced and assembled. The assemblies were recruited to the existing genome, revealing gaps between assemblies, which could represent regions of divergence.



Nucmer recruitment of "B" type fosmids to "A" type genome



Sponge ID	C. symbiosum serpin gene copies/g (SE)		
SB1	$1.18 \times 10^{10} (5.30 \times 10^8)$		
SB2	$7.98 \times 10^9 (2.34 \times 10^8)$		
SB3	Not Detected (NA)		
SB4	Not Detected (NA)		
SB5	8.26x10 ⁸ (1.13x10 ⁸)		
SB6	$3.26 \times 10^9 (1.31 \times 10^8)$		
SB7	3.28x10 ⁹ (3.15x10 ⁸)		
SB8	3.21x10 ⁹ (2.07x10 ⁸)		
SB10	Not Detected (NA)		
SB11	Not Detected (NA)		
BC1	Not Detected (NA)		
BC2	Not Detected (NA)		
BC3	Not Detected (NA)		
BC4	Not Detected (NA)		
BC5	Not Detected (NA)		
BC6	Not Detected (NA)		
BC7	Not Detected (NA)		
BC8	$3.54 \times 10^8 (1.00 \times 10^8)$		
BC9	Not Detected (NA)		
BC10	Not Detected (NA)		
BC11	Not Detected (NA)		
BC12	Not Detected (NA)		

B.2 Quantification of *C. symbiosum*-encoded serpin genes by qPCR

B.3 Identities of proteins pulled down by a synthetic peptide based on C. symbiosum

CENSYa_1605 serpin C-terminus.

Huh7.5.1		НЕК293		
CENSYa_1605WT	CENSYa_1605mut	CENSYa_1605WT	CENSYa_1605mut	
ACAT1	АСТВ	ACAT1	DDX41	
ACSL1	ALDOA	ACSL1	EEF2	
АСТВ	ATP5B	AFG3L2	HSP90AB1	
AGPS	СКВ	ATP5A1	HSPA1A	
AK4	DPP3	ATP5B	HSPA1B	
ATP5O	EEF2	CAMSAP3	HSPA5	
CAMSAP3	EIF4A2	CCT3	HSPA8	
СКВ	ENO1	CDK1	HSPA9	
DPM1	EZR	DDX41	HSPD1	
DYNC1LI1	FASN	EEF1A1	NUDT21	
EEF1A1	GANAB	FADS1	PHGDH	
ENO1	GLUD1	FECH	PIP	
FADS2	HNRNPH1	GPD1L	SUB1	
HIST1H4A	HSD17B4	HSD17B12	TUBA1B	
HSP90AB1	HSP90AB1	HSPA1A	TUBB4B	
HSPA5	HSP90B1	HSPA1B	UBA1	
HSPA8	HSPA5	HSPD1		
HSPA9	HSPA8	IQCB1		
HSPD1	HSPA9	IQGAP3		
IDH1	HSPD1	KPNB1		
KHSRP	KHSRP	KRT6B		
LRPPRC	KRT4	LRPPRC		
MCM7	PABPC1	MCM7		
MYL6	PDIA6	METTL15		
NAMPT	PKM2	MTPAP		
NCL	PPIA	NAMPT		
NDUFS3	PRDX3	NIPSNAP1		
NIPSNAP1	TPI1	NSF		
NSF	TUBA1B	NTPCR		
NSUN5	TUBB2A	NUP153		
NTPCR	UBC	PCYT1A		
PFKL		PDK3		
PFKM		PFKL		
PGAM2		PFKP		

Huh	Huh7.5.1		K293
CENSYa_1605mut	CENSYa_1605WT	CENSYa_1605mut	
PLIN2		PHGDH	
PRDX1		POLD1	
PSMC2		POLD3	
PSMC6		PRKDC	
PSMD9		PSMC2	
PYCR1		PSMC3	
RARS2		PSMC6	
RFC4		RARS2	
RPL12		RDH13	
RPL23		RDH14	
RPL38		RFC3	
RPLP2		RFC4	
RPS14		RPL12	
RPS16		RPL23	
RPS18		RPLP0	
RPS19		RPS10	
RPS25		RPS14	
RPS3		RPS17L	
RPS4X		RPS18	
SCCPDH		RPS19	
SERPINH1		RPS20	
SLC25A5		RPS3	
STOML2		RPS5	
TBC1D5		SERPINH1	
TIMM44		SGPL1	
TRIM28		SLC25A13	
TUBB3		SLC25A5	
TUBB6		TPD52L2	
TUBG1		TPI1	
TUFM		TUBA1B	
UBC		TUBB	
UMPS		TUBB4B	
VAT1		TUBB6	
		TUFM	
		UMPS	
		VAT1	

Appendix C

C.1 Unique KEGG orthologs in *D. mexicanum* and *T. caiifornaian*

Dragmacidon mexicanum	Tethya californiana
K05757 actin related protein 2/3 complex, subunit 1A/1B	K10214 3alpha,7alpha,12alpha-trihydroxy-5beta-cholestanoyl-CoA 24-hydroxylase
K05756 actin related protein 2/3 complex, subunit 3	K08048 adenylate cyclase 8
K05754 actin related protein 2/3 complex, subunit 5	K04135 adrenergic receptor alpha-1A
K07941 ADP-ribosylation factor 6	K01371 cathepsin K
K07368 B-cell CLL/lymphoma 10	K04950 cyclic nucleotide gated channel alpha 3
K08060 class II, major histocompatibility complex, transactivator	K04952 cyclic nucleotide gated channel beta 1
K08009 cytochrome b-245, alpha polypeptide	K02089 cyclin-dependent kinase 4
K04536 guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1	K12366 engulfment and cell motility protein 1
K06483 integrin alpha 4	K05107 Eph receptor A6
K06590 integrin beta 7	K12796 erbb2-interacting protein
K04733 interleukin-1 receptor-associated kinase 4	K05092 fms-related tyrosine kinase 3
K07361 lymphocyte cytosolic protein 2	K09408 forkhead box protein O3
K04432 mitogen-activated protein kinase kinase 3	K02833 GTPase HRas
K12798 NACHT, LRR and PYD domains-containing protein 1	K07209 inhibitor of nuclear factor kappa-B kinase subunit beta
K04446 nuclear factor of activated T-cells, cytoplasmic 1	K04960 inositol 1,4,5-triphosphate receptor type 3
K08064 nuclear transcription factor Y, alpha	K04730 interleukin-1 receptor-associated kinase 1
K05734 p21-activated kinase 4	K11218 Janus kinase 3
K06698 proteasome activator subunit 3 (PA28 gamma)	K05744 LIM domain kinase 2
K05402 toll-interacting protein	K08536 liver X receptor alpha
K10169 toll-like receptor 6	K04369 mitogen-activated protein kinase kinase 2
	K04433 mitogen-activated protein kinase kinase 6
	K13358 Na(+)/H(+) exchange regulatory cofactor NHE-RF2
	K06088 occludin
	K04410 p21-activated kinase 2
	K01324 plasma kallikrein
	K04350 RAS guanyl-releasing protein 1
	K07530 Ras homolog gene family, member D
	K06840 semaphorin 3
	K12459 SH2B adaptor protein 1/3
	K05398 toll-like receptor 1
	K05703 tyrosine-protein kinase Fyn
	K05854 tyrosine-protein kinase Lyn
	K04861 voltage-dependent calcium channel alpha-2/delta-4
	K04851 voltage-dependent calcium channel L type alpha-1D
	K01384 wingless-type MMTV integration site family, member 11
	K00182 wingless-type MMTV integration site family, member 2

C.2 Unique protein families expressed by *D. mexicanum* and *T. californiana*

	Dragmacidon mexicanum		Tethva californiana
PF137461	4Fe-4S dicluster domain	PF13534-1	4Fe-4S dicluster domain
PF04739 10	5'-AMP-activated protein kinase beta subunit interation domain	PF13394.1	4Fe-4S single cluster domain
PF13481 1	A A A domain	PF13173 1	A A A domain
PF12689.2	Acid Phosphatase	PF04572 7	Alpha 1 4-glycosyltransferase conserved region
PF13772 1	AIG2-like family	PF03229.8	Alphavirus glycoprotein I
PE04864.8	Alliinase	PF05586.6	Anthray recentor C-terminus region
DE08521.5	Alnha L. rhamnosidasa N. terminal domain	DE05587.8	Anthrax receptor e-terminus region
DE07244.6	Amastin surface glycoprotain	DE05701.6	Pagillus haemolytic enterotoxin (UPL)
DE04906 7	Ammonia monocyugonoso/mothono monocyugonoso, gubunit C	DE08021 7	Datharing and hathering like
PF12950.7	Ammonia monooxygenase/methane monooxygenase, subunit C	PF15020.1	Cation abound another entries and another and another
PF12859.2	Anapnase-promoting complex subunit 1	PF15020.1	Cation channel sperm-associated protein subunit delta
PF06396.6	Angiotensin II, type I receptor-associated protein (AGTRAP)	PF15510.1	Centromere kinetocnore component w
PF03079.9	ARD/ARD' family	PF13884.1	Chaperone of endosialidase
PF04062.9	ARP2/3 complex ARPC3 (21 kDa) subunit	PF09295.5	ChAPs (Chs5p-Arf1p-binding proteins)
PF04045.9	Arp2/3 complex, 34 kD subunit p34-Arc	PF122/3.3	Chitin synthesis regulation, resistance to Congo red
PF01992.11	ATP synthase (C/AC39) subunit	PF03174.8	Chitobiase/beta-hexosaminidase C-terminal domain
PF01990.12	ATP synthase (F/14-kDa) subunit	PF05966.7	Chordopoxvirus A33R protein
PF05873.7	ATP synthase D chain, mitochondrial (ATP5H)	PF02861.15	Clp amino terminal domain
PF00213.13	ATP synthase delta (OSCP) subunit	PF06172.6	Cupin superfamily (DUF985)
PF14960.1	ATP synthase regulation	PF10170.4	Cysteine-rich domain
PF09486.5	Bacterial type III secretion protein (HrpB7)	PF06144.8	DNA polymerase III, delta subunit
PF02961.9	Barrier to autointegration factor	PF14966.1	DNA repair REX1-B
PF07716.10	Basic region leucine zipper	PF06327.9	Domain of Unknown Function (DUF1053)
PF02892.10	BED zinc finger	PF08014.6	Domain of unknown function (DUF1704)
PF10515.4	beta-amyloid precursor protein C-terminus	PF09350.5	Domain of unknown function (DUF1992)
PF12215.3	beta-Glucocerebrosidase 2 N terminal	PF11958.3	Domain of unknown function (DUF3472)
PF0063413	BRCA2 repeat	PF143261	Domain of unknown function (DUF4384)
PF138651	C-terminal duplication domain of Friend of PRMT1	PF151011	Domain of unknown function (DUF4557
PF139121	C2H2-type zinc finger	PF15141 1	Domain of unknown function (DUF4574)
PE07888.6	Calcium hinding and coiled-coil domain (CALCOCO1) like	PF15158 1	Domain of unknown function (DUF4579)
DE00088 17	Carbamoyl phosphate synthese small chain (CALCOCOT) like	DE15162 1	Domain of unknown function (DUF4570)
DE02787 14	Carbamoyi-phosphate syntheses large shain, cli Sase dollarin	DE15270 1	Domain of unknown function (DUF4580)
PF02/8/.14	Carbahoyi-phosphate synthetase large chain, ongoinensation domain	PF155/9.1	Domain of unknown function (DOF4000)
PF03422.10	Carbonydrate binding module (family 6)	PF15017.1	Drug resistance and apoptosis regulator
PF14915.1	CCDC144C protein colled-coll region	PF04300.8	F-box associated region
PF04103.10	CD20-like family	PF12831.2	FAD dependent oxidoreductase
PF08205.7	CD80-like C2-set immunoglobulin domain	PF14904.1	Family of unknown function
PF08174.6	Cell division protein anillin	PF05400.8	Flagellar protein FliT
PF13097.1	CENP-A nucleosome associated complex (NAC) subunit	PF07504.8	Fungalysin/Thermolysin Propeptide Motif
PF12416.3	Cep120 protein	PF05637.7	galactosyl transferase GMA12/MNN10 family
PF05495.7	CHY zinc finger	PF13522.1	Glutamine amidotransferase domain
PF02017.10	CIDE-N domain	PF01102.13	Glycophorin A
PF01086.12	Clathrin light chain	PF03663.9	Glycosyl hydrolase family 76
PF10534.4	Connector enhancer of kinase suppressor of ras	PF01697.22	Glycosyltransferase family 92
PF12243.3	CTD kinase subunit gamma CTK3	PF04488.10	Glycosyltransferase sugar-binding region containing DXD motif
PF03091.10	CutA1 divalent ion tolerance protein	PF00372.14	Hemocyanin, copper containing domain
PF02936.9	Cytochrome c oxidase subunit IV	PF03723.9	Hemocyanin, ig-like domain
PF02284.11	Cytochrome c oxidase subunit Va	PF00353.14	Hemolysin-type calcium-binding repeat (2 copies)
PF02238.10	Cytochrome c oxidase subunit VIIa	PF03486.9	HI0933-like protein
PF05038.8	Cytochrome Cytochrome b558 alpha-subunit	PF02183-13	Homeobox associated leucine zipper
PF148801	Cytochrome oxidase c assembly	PF14696 1	Hydroxyphenylpyruvate dioxygenase, HPPD, N-terminal
PF02297 12	Cytochrome oxidase c subunit VIb	PF00218 16	Indole-3-glycerol-phosphate synthese
PE110203	DAZ associated protein 2 (DAZAP2)	PF03030 11	Inorganic H+ pyrophosphatase
PE02701.12	DDT domain	PF14755 1	Intracellular membrane remodeller
DE01678 14	Diaminonimalata animarasa	DE00462 16	Isocitrate lusse
DE01728 12	Dianalactona hydrolasa family	DE11747.2	Killing trait
PF01/36.15	DMDK soiled soil demain like	PF021(9.9	Kining uau
PF08820.5	DNPK colled coll domain like	PF03108.8	Late embryogenesis abundant protein
PF08599.5	DNA damage repair protein Nbs1	PF13306.1	Leucine rich repeats (6 copies)
PF12213.3	DNA polymerases epsilon N terminal	PF04991.8	LicD family
PF06469.6	Domain of Unknown Function (DUF1088)	PF117/4.3	Lsr2
PF11864.3	Domain of unknown function (DUF3384)	PF14521.1	Lysine-specific metallo-endopeptidase
PF13320.1	Domain of unknown function (DUF4091)	PF15502.1	M-phase-specific PLK1-interacting protein
PF13660.1	Domain of unknown function (DUF4147)	PF07961.6	MBA1-like protein
PF13904.1	Domain of unknown function (DUF4207)	PF08631.5	Meiosis protein SPO22/ZIP4 like
PF13910.1	Domain of unknown function (DUF4209)	PF13455.1	Meiotically up-regulated gene 113
PF13960.1	Domain of unknown function (DUF4218)	PF13583.1	Metallo-peptidase family M12B Reprolysin-like
PF14124.1	Domain of unknown function (DUF4291)	PF09203.6	MspA
PF15012.1	Domain of unknown function (DUF4519)	PF05283.6	Multi-glycosylated core protein 24 (MGC-24)
PF15074.1	Domain of unknown function (DUF4541)	PF02875.16	Mur ligase family, glutamate ligase domain
PF06012.7	Domain of Unknown Function (DUF908)	PF13887.1	Myelin gene regulatory factor -C-terminal domain 1
			-

	Dragmacidon mexicanum	Tethya californiana
PF05160.8	DSS1/SEM1 family	PF01275.14 Myelin proteolipid protein (PLP or lipophilin)
PF01912.13	eIF-6 family	PF12578.3 Myotubularin-associated protein
PF14578.1	Elongation factor Tu domain 4	PF04666.8 N-Acetylglucosaminyltransferase-IV (GnT-IV) conserved region
PF03735.9	ENT domain	PF08347.6 N-terminal CTNNB1 binding
PF01287.15	Eukaryotic elongation factor 5A hypusine, DNA-binding OB fold	PF03553.9 Na+/H+ antiporter family
PF08555.5	Eukaryotic family of unknown function (DUF1754)	PF00662.15 NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus
PF03332.8	Eukaryotic phosphomannomutase	PF05224.7 NDT80 / PhoG like DNA-binding family
PF01115.12	F-actin capping protein	PF07562.9 Nine Cysteines Domain of family 3 GPCR
PF01267.12	F-actin capping protein alpha subunit	PF02898.10 Nitric oxide synthase, oxygenase domain
PF14675.1	FANCI solenoid I	PF14815.1 NUDIX domain
PF11107.5 PE00532.5	Fanconi anemia group F protein (FANCF)	PF10129.4 OpgC protein PF08447.6 PAS fold
PF08165.6	For domain Fer A (NUC095) domain	PE01364-13 Pentidase family C25
PF06473.7	FGF binding protein 1 (FGF-BP1)	PF02917.9 Pertussis toxin subunit 1
PF14853.1	Fis1 C-terminal tetratricopeptide repeat	PF01503.12 Phosphoribosyl-ATP pyrophosphohydrolase
PF14852.1	Fis1 N-terminal tetratricopeptide repeat	PF06694.6 Plant nuclear matrix protein 1 (NMP1)
PF07474.7	G2F domain	PF06101.6 Plant protein of unknown function (DUF946)
PF14227.1	gag-polypeptide of LTR copia-type	PF04230.8 Polysaccharide pyruvyl transferase
PF13976.1	GAG-pre-integrase domain	PF13243.1 Prenyltransferase-like
PF00337.17	Galactoside-binding lectin	PF07786.7 Protein of unknown function (DUF1624)
PF03227.11	Gamma interferon inducible lysosomal thiol reductase (GILT)	PF07787.7 Protein of unknown function (DUF1625)
PF03321.8	GH3 auxin-responsive promoter	PF10356.4 Protein of unknown function (DUF2034)
PF02800.15	Giveraidenyde 3-phosphate denydrogenase, C-terminal domain	PF10998.5 Protein of unknown function (DUF2838)
PF13/26 1	Glycine-zipper containing OmpA-like membrane domain	PE11595.3 Protein of unknown function (DUE2852)
PF113450.1	Glycoprotein III.132	PF15047 1 Protein of unknown function (DUF4533)
PF02015 11	Glycosyl hydrolase family 45	PF04862.7 Protein of unknown function (DUF642)
PF02057.10	Glycosyl hydrolase family 59	PF06637.6 PV-1 protein (PLVAP)
PF01229.12	Glycosyl hydrolases family 39	PF05202.7 Recombinase Flp protein
PF04616.9	Glycosyl hydrolases family 43	PF00468.12 Ribosomal protein L34
PF13896.1	Glycosyl-transferase for dystroglycan	PF00978.16 RNA dependent RNA polymerase
PF10181.4	GPI-GlcNAc transferase complex, PIG-H component	PF05001.8 RNA polymerase Rpb1 C-terminal repeat
PF13167.1	GTP-binding GTPase N-terminal	PF03579.8 Small hydrophobic protein
PF15003.1	HAUS augmin-like complex subunit 2	PF15497.1 snRNA-activating protein complex subunit 19, SNAPc subunit 19
PF12856.2	Helix-nairpin-helix motif	Pr04852.7 SOUL neme-binding protein
PF13013.1 PF12210.2	Henatocyte growth factor-regulated tyrosing kingse substrate	PE00686.14 Starch binding domain
PF15313 1	Hexamethylene bis-acetamide-inducible protein	PF04069.7 Substrate binding domain of ABC-type glucine betaine transport system
PF09453.5	HIRA B motif	PF06653.6 Tight junction protein, Claudin-like
PF04774.10	Hyaluronan / mRNA binding family	PF01609.16 Transposase DDE domain
PF01630.13	Hyaluronidase	PF01060.18 Transthyretin-like family
PF15244.1	Hydroxy-steroid dehydrogenase	PF04820.9 Tryptophan halogenase
PF01294.13	ibosomal protein L13e	PF12381.3 Tungro spherical virus-type peptidase
PF11711.3	Inner membrane protein import complex subunit Tim54	PF08581.5 Tup N-terminal
PF00219.13	Insulin-like growth factor binding protein	PF04406.9 Type IIB DNA topoisomerase
PF11261.3	Interferon regulatory factor 2-binding protein zinc finger	PF15544.1 Type IV pillin N-term methylation site GFxxxE
PF04836./	Internet on-related protein conserved region	PF02594.11 Uncharacterised AUK, YggU family CUG18/2 DE12264.2 Waikavirus consid protoin 1
PF05439 7	Jumping translocation breakpoint protein (JTR)	PF13115.1 VtkA-like
PF10282.4	Lactonase. 7-bladed beta-propeller	PF13240.1 zinc-ribbon domain
PF00052.13	Laminin B (Domain IV)	
PF00055.12	Laminin N-terminal (Domain VI)	
PF15454.1	Late endosomal/lysosomal adaptor and MAPK and MTOR activator	
PF01613.13	lavin reductase like domain	
PF00538.14	linker histone H1 and H5 family	
PF10242.4	Lipoma HMGIC fusion partner-like protein	
PF01299.12	Lysosome-associated membrane glycoprotein (Lamp)	
PF14918.1 DE15162.1	MDM2-0Inding	
PF05850 7	Mis12 protein	
PF05511.6	Mitochondrial ATP synthase coupling factor 6	
PF04718.10	Mitochondrial ATP synthase coupling factor of	
PF08923.5	Mitogen-activated protein kinase kinase 1 interacting	
PF12554.3	Mitotic-spindle organizing gamma-tubulin ring associated	
PF02536.9	mTERF	
PF08523.5	Multiprotein bridging factor 1	
PF08245.7	Mur ligase middle domain	
PF07994.7	Myo-inositol-1-phosphate synthase	
PF12632.2	Mysoin-binding motif of peroxisomes	
PF0/657.8	N terminus of Notch ligand	

	Dragmacidon mexicanum	Tethya californiana
PF09764.4	N-terminal glutamine amidase	
PF102004	NADH:ubiquinone oxidoreductase NDUES5-15kDa	
DE02259 10	NADPH dependent EMN reduction	
PF05558.10	NADPH-dependent FMIN reductase	
PF05/41.8	Nanos KNA binding domain	
PF05536.6	Neurochondrin	
PF01106.12	NifU-like domain	
PF12922.2	non-SMC mitotic condensation complex subunit 1. N-term	
PE08163 7	NUC194 domain	
DE002706	Nuclease related domain	
PF083/8.6	Nuclease-related domain	
PF03066.10	Nucleoplasmin	
PF13634.1	Nucleoporin FG repeat region	
PF02101.10	Ocular albinism type 1 protein	
PF05708 7	Orthonoxyirus protein of unknown function (DUF830)	
PE00024 21	PAN domain	
FF00024.21	PAN domain	
PF14295.1	PAN domain	
PF15364.1	PAXIP1-associated-protein-1 C term PTIP binding protein	
PF12708.2	Pectate lyase superfamily protein	
PF13812.1	Pentatricopeptide repeat domain	
PE08127.8	Pentidase family C1 propentide	
DE01625.16	Pontida mathianina gulfavida raduatasa	
FF01023.10	replide metholine sunoxide reductase	
PF09262.6	Peroxisome biogenesis factor 1, N-terminal	
PF15473.1	PEST, proteolytic signal-containing nuclear protein family	
PF07819.8	PGAP1-like protein	
PF02567.11	Phenazine biosynthesis-like protein	
PE03660.9	PHE5-like protein	
DE04607.9	ninin/SDV concerned region	
PF04097.8	pinin/SDK conserved region	
PF03840.9	Preprotein translocase SecG subunit	
PF15388.1	Protein Family FAM117	
PF06918.9	Protein of unknown function (DUF1280)	
PF07713.8	Protein of unknown function (DUF1604)	
DE07804 7	Protein of unknown function (DUF1660)	
PF00(40.7	Protein of unknown function (DUF1707)	
PF08648./	Protein of unknown function (DUF1///)	
PF08894.6	Protein of unknown function (DUF1838)	
PF10176.4	Protein of unknown function (DUF2370)	
PF10309.4	Protein of unknown function (DUF2414)	
PF10961 3	Protein of unknown function (DUE2763)	
DE11770.2	Distain of unknown function (DUE2217)	
PF11//9.5	Protein of unknown function (DUF3517)	
PF12341.3	Protein of unknown function (DUF3639)	
PF12530.3	Protein of unknown function (DUF3730)	
PF12903.2	Protein of unknown function (DUF3830)	
PF03385.12	Protein of unknown function, DUF288	
DE0/685.8	Protein of unknown function DUE608	
DE07020.0		
PF0/830.8	Protein serine/threonine phosphatase 2C, C-terminal domain	
PF10350.4	Putative death-receptor fusion protein (DUF2428)	
PF06508.8	Queuosine biosynthesis protein QueC	
PF13902.1	R3H-associated N-terminal domain	
PF09072 5	ranslation machinery associated TMA7	
DE10262.4	Rdy family	
FF10202.4	Rux failing	
PF03398.9	Regulator of Vps4 activity in the MVB pathway	
PF04471.7	Restriction endonuclease	1
PF02453.12	Reticulon	1
PF00077.15	Retroviral aspartyl protease	
PF07727 9	Reverse transcriptase (RNA-dependent DNA polymerase)	
DE08012.6	Pho Binding	
DE02115.15	Nilo munig	
PF02115.12	KHO protein GDP dissociation inhibitor	
PF00545.15	Ribonuclease	
PF01776.12	Ribosomal L22e protein family	
PF01777.13	Ribosomal L27e protein family	
PF01780 14	Ribosomal L 37ae protein family	
DE00672 14	ribosomal I 5D family C terminus	1
FF000/3.10	noosoniai Lor family C-terminus	
PF03939.8	Ribosomal protein L23, N-terminal domain	
PF01198.14	Ribosomal protein L31e	1
PF01158.13	Ribosomal protein L36e	
PF00281 14	Ribosomal protein L.5	
DE03849 10	Ribosomal protein L6 N terminal domain	
DE01202.14	Diharana protein 20, iv-terminar dollalli	
PF01283.14	Kibosomai protein S26e	
PF01200.13	Ribosomal protein S28e	1
PF00189.15	Ribosomal protein S3, C-terminal domain	
PF04758.9	Ribosomal protein S30	
PF00410 14	Ribosomal protein S8	
	··· r ··· ·	1

	Dragmacidon mexicanum	Tethya californiana
PF08069.7	Ribosomal S13/S15 N-terminal domain	
PF10501.4	Ribosomal subunit 398	
PF11707.3	Ribosome 60S biogenesis N-terminal	
PF14200.1	Ricin-type beta-trefoil lectin domain-like	
PF08675.6	RNA binding domain	
PF05183.7	RNA dependent RNA polymerase	
PF10347.4	RNA pol II promoter Fmp27 protein domain	
PF01192.17	RNA polymerase Rpb6	
PF04699.9	RP2/3 complex 16 kDa subunit (p16-Arc)	
PF08621.5	RPAP1-like. N-terminal	
PF12328.3	Rpp20 subunit of nuclear RNase MRP and P	
PF08167.7	rRNA processing/ribosome biogenesis	
PF08071.7	RS4NT (NUC023) domain	
PF02026.11	Rvanodine receptors	
PF02199.10	Saposin A-type domain	
PF12701.2	Scd6-like Sm domain	
PF03911.11	Sec61beta family	
PF02978.14	Signal peptide binding domain	
PF01466.14	Skp1 family dimerisation domain	
PF12680.2	SnoaL-like domain	
PF08557.5	Sphingolipid Delta4-desaturase (DES)	
PF05032.7	Spo12 family	
PF01922.12	SRP19 protein	
PF07304.6	Steroid receptor RNA activator (SRA1)	
PF01127.17	Succinate dehvdrogenase/Fumarate reductase transmembrane subunit	
PF09177.6	Syntaxin 6, N-terminal	
PF09247.6	TATA box-binding protein binding	
PF11640.3	Telomere-length maintenance and DNA damage repair	
PF05485.7	THAP domain	
PF04821.9	Timeless protein	
PF15122.1	TMEM206 protein family	
PF00923.14	Transaldolase	
PF09748.4	Transcription factor subunit Med10 of Mediator complex	
PF03847.8	Transcription initiation factor TFIID subunit A	
PF00838.12	Translationally controlled tumour protein	
PF14995.1	Transmembrane protein	
PF04201.10	Tumour protein D52 family	
PF14617.1	U3-containing 90S pre-ribosomal complex subunit	
PF02320.11	Ubiquinol-cytochrome C reductase hinge protein	
PF05365.7	Ubiquinol-cytochrome C reductase, UQCRX/QCR9 like	
PF03671.9	Ubiquitin fold modifier 1 protein	
PF03650.8	Uncharacterised protein family (UPF0041)	
PF03669.8	Uncharacterised protein family (UPF0139)	
PF03670.8	Uncharacterised protein family (UPF0184)	
PF05251.7	Uncharacterised protein family (UPF0197)	
PF05255.6	Uncharacterised protein family (UPF0220)	
PF01980.11	Uncharacterised protein family UPF0066	
PF15369.1	Uncharacterised protein KIAA1328	
PF09848.4	Uncharacterized conserved protein (DUF2075)	
PF09803.4	Uncharacterized conserved protein (DUF2346)	
PF02151.14	UvrB/uvrC motif	
PF13538.1	UvrD-like helicase C-terminal domain	
PF03179.10	Vacuolar (H+)-ATPase G subunit	
PF05827.7	Vacuolar ATP synthase subunit S1 (ATP6S1)	
PF09967.4	VWA-like domain (DUF2201)	
PF10349.4	WW-domain ligand protein	
PF02542.11	YgbB family	
PF08892.6	Yqcl/Y cgG tamily	
PF14369.1	zinc-tinger	<u> </u>

C.3 Proteases expressed at the mRNA level by *D. mexicanum*

(BLASTp, E-value <1E-6, bsr >0.4)

Aspartic	Metallo	Cysteine	Serine	Threonine
cathepsin D	ADAM10	calpain 5	acylaminoacyl-peptidase	proteasome β-3 subunit
gamma-secretase subunit Aph- 1b-like	AFG3-like protein 2	calpain 9	coagulation factor VII	γ-glutamyltransferase
presenilin-2-like	aminoacylase	caspase 10	coagulation factor XI	
	aminopeptidase 3	caspase 3	coagulation factor XIII	
	aminopeptidase A	caspase 7	corin	
	aminopeptidase N	caspase 8	dipeptidyl-peptidase 9	
	aminopeptidase Y	caspase 9	dipeptidyl-peptidase II	
	aspartate carbamoyltransferase dihydroorotase	cathepsin B	furin	
	aspartyl aminopeptidase	cathepsin F	heat shock 90kDa protein 1	
	carboxypeptidase A1-like	cathepsin H	hepatocyte growth factor	
	carboxypeptidase A4-like	cathepsin L	HTRA2	
	carboxypeptidase B2-like	cathepsin Z	lysosomal Pro-X carboxypeptidase	
	carboxypeptidase C	legumain	matriptase	
	carboxypeptidase D	pyroglutamyl-peptidase	neurotrypsin	
	carboxypeptidase E	testin	prolyl oligopeptidase	
	COPS6	ubiquitin carboxyl-terminal hydrolase 10	proprotein convertase subtilisin/kexin type 2	
	cytosol aminopeptidase	ubiquitin carboxyl-terminal hydrolase 11	proprotein convertase subtilisin/kexin type 5-like	
	cytosol aminopeptidase-like	ubiquitin carboxyl-terminal hydrolase 12	proprotein convertase subtilisin/kexin type 6 precursor	
	cytosolic carboxypeptidase 1	ubiquitin carboxyl-terminal hydrolase 14	proprotein convertase subtilisin/kexin type 7-like	
	dihydroorotase-like	ubiquitin carboxyl-terminal hydrolase 15	proprotein convertase subtilisin/kexin type 9-like	
	dihydropyrimidinase	ubiquitin carboxyl-terminal hydrolase 16	protein C	
	dipeptidyl-peptidase III	ubiquitin carboxyl-terminal hydrolase 19	rhomboid-like protein 1	
	endothelin-converting enzyme	ubiquitin carboxyl-terminal hydrolase 2	serine carboxypeptidase 1	
	glutamyl aminopeptidase-like	ubiquitin carboxyl-terminal hydrolase 22	site-1 protease	
	insulysin	ubiquitin carboxyl-terminal hydrolase 24	thrombin	
	leishmanolysin	ubiquitin carboxyl-terminal hydrolase 25	tripeptidyl-peptidase II	
	leukotriene A4 hydrolase	ubiquitin carboxyl-terminal hydrolase 28	tumor rejection antigen gp96-like	
	membrane dipeptidase	ubiquitin carboxyl-terminal hydrolase 3	β-lactamase	
	methionyl aminopeptidase	ubiquitin carboxyl-terminal hydrolase 31		
	mitochondrial intermediate peptidase	ubiquitin carboxyl-terminal hydrolase 32		
	mitochondrial processing peptidase	ubiquitin carboxyl-terminal hydrolase 4		

Aspartic	Metallo	Cysteine	Serine	Threonine
	nardilysin	ubiquitin carboxyl-terminal hydrolase 46		
	neprilysin	ubiquitin carboxyl-terminal hydrolase 47		
	O-sialoglycoprotein endopeptidase	ubiquitin carboxyl-terminal hydrolase 5		
	paraplegin	ubiquitin carboxyl-terminal hydrolase 6		
	pitrilysin metallepetidase 1-like	ubiquitin carboxyl-terminal hydrolase 7		
	plasma glutamate carboxypeptidase	ubiquitin carboxyl-terminal hydrolase BAP1		
	prolyl aminopeptidase			
	prolyl aminopeptidase serine peptidase merops fa	mily s33		
	PRPF8			
	PSMD7			
	puromycin-sensitive aminopeptidase			
	serine carboxypeptidase CPVL			
	serine carboxypeptidase S10 family member 1			
	X-Pro aminopeptidase			
	X-Pro dipeptidase			