The Genome of *Aiptasia* and the Role of MicroRNAs in Cnidarian-Dinoflagellate Endosymbiosis

Dissertation by

Sebastian Baumgarten

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology, Thuwal,

Kingdom of Saudi Arabia

EXAMINATION COMMITTEE APPROVALS FORM

Committee Chairperson: Christian R. Voolstra

Committee Member: Manuel Aranda

Committee Member: Arnab Pain

Committee Member: John R. Pringle

To my Brother in Arms

ABSTRACT

The Genome of *Aiptasia* and the Role of MicroRNAs in Cnidarian-Dinoflagellate Endosymbiosis

Sebastian Baumgarten

Coral reefs form marine-biodiversity hotspots of enormous ecological, economic, and aesthetic importance that rely energetically on a functional symbiosis between the coral animal and a photosynthetic alga. The ongoing decline of corals worldwide due to anthropogenic influences heightens the need for an experimentally tractable model system to elucidate the molecular and cellular biology underlying the symbiosis and its susceptibility or resilience to stress.

The small sea anemone *Aiptasia* is such a model organism and the main aims of this dissertation were 1) to assemble and analyze its genome as a foundational resource for research in this area and 2) to investigate the role of miRNAs in modulating gene expression during the onset and maintenance of symbiosis.

The genome analysis has revealed numerous features of interest in relation to the symbiotic lifestyle, including the evolution of transposable elements and taxonomically restricted genes, linkage of host and symbiont metabolism pathways, a novel family of putative pattern-recognition receptors that might function in host-microbe interactions and evidence for horizontal gene transfer within the animal-alga pair as well as with the associated prokaryotic microbiome.

The new genomic resource was used to annotate the *Aiptasia* miRNA repertoire to illuminate the role of post-transcriptional regulatory mechanisms in regulating endosymbiosis. *Aiptasia* encodes a majority of species-specific miRNAs and first evidence is presented that even evolutionary conserved miRNAs are undergoing recent differentiations within the *Aiptasia* genome. The analysis of miRNA expression between different states of *Symbiodinium* infection further revealed that species-specific and conserved miRNAs are symbiotically regulated. In order to detect functional miRNA-mRNA interactions and to investigate the downstream effects of such miRNA action, a protocol for cross-linking immunoprecipitations of Argonaute, the central protein of the miRNA-induced silencing complex, was developed. This method identified binding sites of miRNAs on a transcriptome-wide scale and revealed target genes of symbiotically regulated miRNAs that were identified previously to be involved in the symbiosis.

In summary, this dissertation provides novel insights into miRNA-mediated post-transcriptional modulation of the host transcriptome and by presenting a critically needed genomic resource, lays the foundation for the continued development of *Aiptasia* as a model for coral symbiosis.

ACKNOWLEDGMENTS

A PhD for sure it is a great journey workwise, and you can find the “logbook” of this work in the document below. But what makes this - and every other - trip so special are always the people that are joining you.

I’d like to express my gratitude to Dr. Christian Voolstra, who trusted my potential from the very beginning of our joint work. I greatly appreciate the ultimate freedom that he provided me to pursue my projects, offering the guidance when it was needed and especially giving me the opportunities to collaborate with great people to broaden my expertise. This holds true as much for Dr. Manuel Aranda: Thank you for the constructive criticism, your sympathetic ear and the great scientific discussions.

I would like to extent my appreciation to Dr. John Pringle for his mentoring, the endurance, sharing his immense scientific experience and finally for hosting me at Stanford. I’m also grateful to Dr. Arnab Pain for serving in my PhD committee and giving the right advice in the right moment and, finally, to Craig Michell for showing me his invaluable tricks in the lab and being so patient with me.

Also, it wouldn’t have been gone as smoothly without all the colleagues and friends: Thank you for sharing the highs and being there during the lows!

And most importantly, I would like to thank my family. Knowing about your support made me go all the way through and I’ll be forever grateful for my parents trust, faith and unconditional love, and finally to my brother: you know it all!

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

aa	Amino acid
Adi	*Acropora digitifera*
Ago	Argonaute
Aip	*Aiptasia*
Aqu	*Amphimedon queenslandica*
ASFW	Artifical, sterile-filtered seawater
ASW	Artifical seawater
Ava	*Adineta vaga*
Bfl	*Branchiostoma floridae*
BLAST	Basic local alignment search tool
bp	Basepair
cDNA	Complementary DNA
CDS	Coding DNA sequence
Cel	*Caenorhabditis elegans*
Cgi	*Crassostrea gigas*
Cin	*Ciona intestinalis*
CLIP	Crosslinking immunoprecipitation
CNIFL	Cnidarian ficolin-like proteins
CRBC	Chicken red blood cells
Cte	*Capitella teleta*
Dme	*Drosophila melanogaster*
DNA	Deoxyribonucleic acid
Dpu	*Daphnia pulex*
Dre	*Danio rerio*
FDR	False discovery rate
FPKM	Fragment per kilobase per one million mapped reads
FREP	Fibriongen related proteins

gDNA	Genomic DNA
Gga	*Gallus gallus*
GO	Gene ontology
GTR	Generalized time-reversible
HGT	Horizontal gene transfer
HITS-CLIP	High-throughput sequencing after crosslinking immunoprecipitation
Hma	*Hydra magnipapillata*
HMM	Hidden markov model
Hro	*Helobdella robusta*
Hsa	*Homo sapiens*
ILR	Interleukin-like receptor
Isc	*Ixodes scapularis*
kb	Kilobae
kcal	Kilocalories
kDa	Kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
Lgi	*Lottia gigantea*
LRR	Leucin rich repeat
Mb	Megabase
MFE	Minimum free energy
miRISC	miRNA-induced silencing complex
miRNA	microRNA
Mmu	*Mus musculus*
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NLR	NOD-like receptor
NOD	nucleotide-binding and oligomerization domain
nr	non-redundant protein database
Nve	*Nematostella vectensis*
ORF	Open reading frame

PCA	Principal component analysis
Pfu	Pinctada fucata
piRNA	PIWI-interacting RNA
PIWI	P-element induced wimpy testis
pre-miRNA	Precursor RNA
pri-miRNA	Primary RNA
PRR	Pattern recognition receptor
RNA	Ribonucleic acid
RNAi	RNA interference
RNase	Ribonuclease
rRNA	Ribosomal RNA
siRNA	Small interfering RNA
Sko	*Saccoglossus kovalevskii*
Sma	*Schistosoma mansoni*
Sme	*Schmidtea mediterranea*
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
Spu	*Strongylocentrotus purpuratus*
SRA	Sequence Read Archive
Tad	*Trichoplax adhaerens*
Tca	*Triboleum castaneum*
TE	Transposable element
TLR	Toll-like receptor
TRG	Taxonomically-restricted gene
tRNA	Transfer RNA
UTR	Untranslated region
Xtr	*Xenopus tropicalis*

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Coral Reefs and the cnidarian-dinoflagellate endosymbiosis

The symbiotic relationship between scleractinian corals and dinoflagellates of the genus *Symbiodinium* enables the build-up of the largest biogenic structures on earth – coral reefs. In this endosymbiosis, the autotrophic symbionts reside within vesicles of the host gastrodermal cells and translocate >90% of their photosynthates to the host organism in exchange for inorganic nutrients and shelter (1). This tight nutrient recycling mechanism enables the coral-symbiont 'holobiont' to overcome the oligothrophic (nutrient-depleted) nature of tropical ocean waters, making coral reefs hotspots of biomass and biodiversity with an invaluable economical output and aesthetic value (2). However, coral reefs are one of the most endangered ecosystems, mainly due to anthropogenic impacts such as overfishing (3, 4), increasing sea surface temperatures as well as ocean acidification as a result of rising atmospheric CO_2 levels (5). To understand the mechanisms and consequences of those impacts and to facilitate coral reef conservation, an increasing global research effort addresses the broad ecological functioning of coral reefs (6, 7).

In contrast to the growing knowledge on coral reef ecology, the underlying molecular mechanisms of coral susceptibility, resilience and acclimation, which are arguable tied to the symbioses of coral hosts and their symbionts, remain unclear (8). For example, while it is well known that the diversity of *Symbiodinium* residing within a coral host depends largely on the host species itself, its health

state, the geographic location and water depth (9–15). In addition, the association of coral hosts with specific endosymbiont genotypes might also confer different physiologies to the coral-dinoflagellate metaorganism, such as resistance to warmer water temperatures or high salinities (16–19). The understanding of the underlying molecular mechanisms of how this symbiosis with its seemingly high specificity is selected, established, and maintained could ultimately help in the prediction - or even alteration - of coral resilience and/or susceptibility, but is still almost unknown to date.

1.2 The molecular basis of coral symbiosis

First evidence of genes and pathways involved in the endosymbiosis between cnidarians and dinoflagellates arose from studies using coral larvae or symbiotic anthozoan anemones. Several studies focused on the analysis of differentially expressed genes between symbiotic and aposymbiotic host animals using both microarray platforms as well as RNA sequencing (20–25). In most of these studies, similar major gene classes are differentially regulated in response to *Symbiodinium* infection, maintenance of a stable endosymbiont population and also during the symbiosis breakdown.

In the early phase of host-symbiont recognition, several pattern recognition receptors (PRRs) and pathway components were suggested to be important based on their homology to known causative regulators of innate immunity in other organisms and their response to mostly endoparasites. These include the families of C-type lectins, Toll-like, NOD-like, and scavenger receptors, the

complement factor C3 and TGFβ pathway components (26–30). Beyond their changing expression patterns, evidence for their involvement in the endosymbiosis arose especially from *in situ* hybridizations and co-localization studies using immunohistochemical protein staining (29, 31, 32).

Research on the internalization of the symbiont and maturation of the symbiont-containing vesicle (i.e. the symbiosome) focused particularly on a family of Rab GTPases, proteins that were known to be involved in the infection of mammalian cells by *Trypanosomas bruzei* or *Mycobacterium tuberculosis* (33). The GTPases Rab5 and Rab7 were found to co-localize exclusively with either healthy (Rab5) (34) or heat-killed (Rab7) (35) symbiosomes in the host cell, suggesting a highly specific role of these two proteins in endosomal trafficking of the symbiosome.

During the endosymbioses maintenance, especially genes that are involved in metabolic exchange and cross-membrane transportation, presumably at the symbiosome membrane could also be identified by differential expression analysis and an proteomics approach of the symbiosome membrane itself (21, 36, 37). In those studies, especially the sterol transporters of the Niemann-Pick type C family (NPCs) appeared to be symbiotically regulated across different taxa and developmental stages, drawing increased attention to their involvement in the interaction of symbiotic cnidarians and dinoflagellates (21, 36, 38).

Although the protein families presented above are regarded contributory in endosymbioses of several cnidarian species, functional studies proofing an effective role of any of these proteins are still lacking. In addition, large differences in the genomic and transcriptomic content even among symbiotic

cnidarians became evident from the first sequencing efforts (39, 40). Those include the huge expansion of NOD-like receptors in *Acropora digitifera* (41), the seemingly recent duplications of the NPC family in *Aiptasia* and *A. digitifera* (21, 38) and the reduced set or even complete absence of canonical Toll-like receptors in the two anthozoan anemones *Anthopleura elegantissima* and *Aiptasia* (28).

In contrast, the specific change of the host transcriptome in the context of endosymbiosis also raises the question about putative (post-) transcriptional regulations that mediate, control and fine-tune these alterations of gene expression. This is also of particular interest since previous studies in coral larvae have also reported few measurable transcriptional responses during *Symbiodinium* infection, putting epigenetic, post-transcriptional and -translational regulations into the foreground (23, 42).

Changing the focus from expression changes of individual genes towards their regulatory mechanisms might not only provide new insights into basal metazoan gene regulatory mechanism in general, but also identify regulators that are orchestrating multiple transcriptional responses that are commonly contributing to the symbiotic lifestyle of cnidarians.

1.3 Small RNA mediated post-transcriptional gene regulation

In the last decades, small non-coding RNAs and the gene regulatory mechanism of RNA interference (RNAi) got well described as key players in genome and gene regulation both in plants and metazoans (43–45). Three major classes of

small RNAs are now well characterized to control a variety of biological processes: MicroRNAs (miRNAs), small-interfering RNAs (siRNAs), and PIWI-interacting RNAs (piRNAs) (46). In common to all three classes is their mode of action in which they bind complementary to their target transcripts, leading mainly to post-transcriptional gene silencing but also to DNA methylation and transposon silencing through which they control or modify a number of cellular processes (45, 47).

Investigations on the role of miRNAs in various biological contexts has revealed their crucial importance in a multitude of cellular processes, including the regulation of development and cell differentiation, the fine-tuning of signaling pathways and the control of innate immunity (48–51). The identification of miRNAs and the protein machinery involved in their biogenesis and action in animals, plants, and also alveolates suggests that miRNA-mediated gene regulation is highly evolutionary conserved (43, 52, 53) and underlines the critical importance of miRNA-mediated genome and gene regulation (26, 53, 54).

The canonical miRNA biogenesis pathways of bilaterians and cnidarians overlap in most parts of the main proteins that are involved, but a few cnidarian protein co-factors and modifications of the miRNAs itself resemble those of plant miRNAs, raising question on the independent evolution of miRNA biogenesis in plants and metazoans (46, 54).

In all metazoans, a primary miRNA transcript (pri-miRNA) is transcribed in the nucleus by the RNA polymerase II and subsequently capped and polyadenylated at the 5’ and 3’ ends, respectively (Figure 1.1). From this pri-miRNA transcript, a

first miRNA intermediate product, the miRNA precursor, is cleaved by the Drosha protein. This endonuclease cleaves a hairpin structure, the pre-miRNA, from the exceeding outer 5’ and 3’ ends of the pri-miRNA transcripts, a process which also involves the two co-factors Ars2 and Pasha. The pre-miRNA hairpin is then exported from the nucleus by Exportin-5, where a second endonuclease, Dicer, performs the further miRNA maturation. The double-stranded pre-miRNA gets bound into the Dicer protein by a PAZ domain that recognizes RNA duplex ends with short 3’ nucleotide overhangs, which got produced during the cleavage through the Drosha protein. From the PAZ domain, the pre-miRNA extends two helical turns towards two RNase III domains that cleave the hairpin stem off the pre-miRNA terminal loop, releasing the mature miRNA together with its complementary strand (miRNA*) from the Dicer protein. The miRNA duplex is characterized by incomplete base pairing and by its two 3’ 2 nt overhangs which referring to the staggered cleavage of the two Dicer RNase III domains. In bilaterians, the protein co-factors Loquacious, TRBP and PACT help in the Dicer-mediated pre-miRNA cleavage, whereas homologs of those proteins seem to be absent in cnidarians. In contrast, cnidarian genomes seem to encode a homolog of the plant Dicer-like co-factor HYL1, a double-stranded RNA binding protein which is absent in bilaterians but might be involved in cnidarian miRNA biogenesis (54). Following the last Dicer cleavage, bilaterian miRNAs get bound into the PAZ domain of the Argonaute effector protein to form the basic miRNA-induced silencing complex (miRISC), whereas the miRNA* is released and degraded. In comparison, before the final assembly of the miRISC in cnidarian,

the miRNAs are 2’ *O*-methylated at their 3’ end, a methylation pattern that is usually found only in bilaterians piRNAs and siRNAs as well as plant miRNAs.

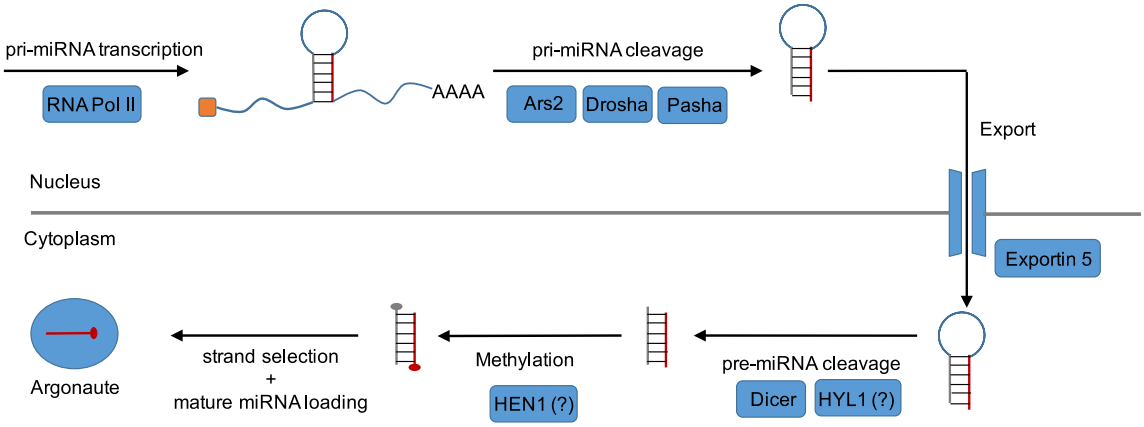

Figure 1.1: Model of proteins that are involved in the biogenesis of miRNAs in cnidarians.

In the miRNA-mediated gene silencing process, the miRNA guides the miRISC to a mRNA target where it binds to partial complementary bindings sites. The specific binding pattern of the miRNA to its target mRNA is often incomplete and includes GU-wobbles and bulges and is also distinct among different metazoan phyla. For example, it was shown that ‘higher’ metazoans require the binding of the miRNA seed, corresponding to the 5’ nucleotides 2 to 7, for efficient target repression (55). In contrast, for cnidarians it was suggested that the miRNA frequently features almost perfect complementarity to the mRNA target and regulates the mRNA target by cleavage, a mechanism that is known predominantly from plant miRNAs and endogenous siRNAs (56).

1.4 The role of miRNA-mediated gene regulation in eukaryotic endosymbioses

One miRNA often regulates several genes within one pathway leading to overall broad transcriptional changes. One well-studied example of such miRNA regulation is the vertebrate miR-122, which is almost exclusively expressed in liver cells at overall very high expression levels (57). Several genes involved in fatty-acid and cholesterol metabolism are modulated by miRNA-122 and inhibition of miR-122 action leads to a reduced plasma cholesterol level and increased membrane fluidity. Even though overall little is known about the role of miRNAs in eukaryotic endosymbiosis, one example of parasite interference in the miRNA-mediated gene regulation of the host is known from the mammalian eukaryotic endoparasites *Leishamina donovani* (58). During the host invasion phase, *L. donovani* excretes exosomoes containing surface virulence factors such as the metalloprotease gp63 that leads to the inhibition of the Dicer endonuclease in the host organism, followed by and interruption of the miRNA biogenesis pathway and eventually to decreased levels of miR-122. Down-regulation of miR-122 causes lowered membrane cholesterol levels and eventually leads to a decreased membrane stability and easier cell invasion of *L. donovani*. This parasite-mediated interference in the host miRNA biogenesis highlights the importance of both, the broad transcriptional changes that are mediated by one individual miRNA as well as the putative effective role of miRNAs in a eukaryote-eukaryote endosymbiosis.

This example also raises the possibility that the broad transcriptional changes observed in the mutualistic relationship between cnidarians and their dinoflagellate endosymbionts might also be modulated by miRNAs.

1.5 *Aiptasia* as a model organism for coral symbiosis

The overall lack of knowledge in regard to the molecular basis of the cnidarian-endosymbiosis mainly refers to several drawbacks related to the nature of scleractinian corals that make lab-based, molecular work difficult; the susceptibility of corals to changes of their environmental conditions asks for sophisticated and environmentally stable culturing setups for laboratory experiments. Further, sexual reproduction of corals in nature occurs only a few times per year and is triggered by external environmental factors. Till now, it is not possible to imitate those environmental triggers in culturing setups and to induce sexual reproduction of coral cultures to fully close their life cycle. This also prevents basic molecular and genetic methods to be applicable to coral research, such as stably inducing cross-generational somatic mutations and gene knockouts. Another weakness of corals as laboratory model systems is the fact that they occur only for a very short period in a symbiont-free (aposymbiotic) state during their life history, namely in the larval phase, making reliable comparisons of symbiotic and aposymbiotic states on a physiological and genetical basis unfeasible.

The small sea anemone *Aiptasia* lives in the same symbiotic relationship with dinoflagellates of the genus *Symbiodinium* as tropical scleractinian corals.

However, while scleractinian corals are dependent on their dinoflagellate symbiont and thus live in an obligate symbiosis, *Aiptasia* anemones also thrive in an aposymbiotic (symbiont-free) state, featuring a facultative symbiosis. The aposymbiotic state of adult *Aiptasia* anemones allows to comparatively study symbiont acquisition and maintenance on a molecular, cellular, tissue-specific, and organismal level. Further, *Aiptasia* incorporates several advantages that qualify it as a model organism not only for eco-physiological studies; most importantly, *Aiptasia* is undemanding in regard of its culturing conditions, making it possible to grow whole populations in stagnant water tanks and even allowing the regular induction of gametogenesis and spawning (59).

Aiptasia is globally distributed with two major genotypes being described to date; one global genotype is found in Japan, Mexico, Hawaii, the Mediterranean and Australia and a second, local genotype occurs exclusively in Florida (51). Interestingly, the two genotypes are distinguished in their endosymbiotic relationship with *Symbiodinium* species. The global genotype seems to occur exclusively in endosymbiosis with *Symbiodinium minutum* from clade B. In contrast, the local genotype from Florida occurs in combinations of *S. minutum* and other species of clade A and C, but never exclusively with either species of clade A or C. These distinct patterns of distribution and the specificity to *S. minutum* suggests a certain degree of coevolution, but also provides a setup to study host-symbiont adaptation and evolution in the more ‘flexible’ relationship of the ‘local’ *Aiptasia* genotype (61).

1.6 Genomic research in the cnidarian phylum

First molecular resources such as the *Aiptasia* transcriptome of the local genotype were made available only recently (21, 40, 62), providing first opportunities for functional genetic and genomic studies focusing on differentially expressed genes between different symbiotic states (21) as well as bleaching (59) and nutrient translocation (64). Until today, only three cnidarian genomes are fully sequenced, including the freshwater cnidarian *Hydra magnipapillata* (65), the starlet sea anemone *Nematostella vectensis* (66) and the scleractinian coral *A. digitifera* (39). Each of these sequencing allowed for the first time comparative studies among basal metazoans as well as genomic comparisons of cnidarians as the sister lineage to bilaterian metazoans (66, 67). But more importantly, these genome assemblies also overcame for the first time the lack of baseline information on the organisms' genetic content, which greatly pushed forward and facilitated research using these organisms as novel model systems. Today, especially *H. magnipapillata* and *N. vectensis* are established models organisms and frequently used for the study of aging, regeneration, stem cell activity (68–70), axis formation and patterning (71, 72) as well as innate immunity (73) and even the interplay of host organisms with their prokaryotic microbiome (74). Thus, to also fully establish *Aiptasia* as the model system for cnidarian-dinoflagellate endosymbiosis, further genomic resources are urgently needed.

The presence of a complete and annotated reference genome assembly will greatly facilitate the advance of new technical approaches (i.e. CRISPR-Cas9 mediated genome editing) as well as easier analysis of data on a transcriptome

or even genome wide scale such as DNA methylation patterns or the role of small RNA-mediated, post-transcriptional gene regulation (8, 75).

1.7 Project Summary

This dissertation aimed to elucidate the role of miRNAs in the cnidarian-dinoflagellate endosymbiosis. For this, first the *Aiptasia* genome was assembled and annotated to (1) allow the identification of functional small RNAs and (2) provide a foundational resource to facilitate future studies on the molecular basis of endosymbiosis using a model-system approach. Chapter 2 describes the technical approach of the *Aiptasia* genome analysis and portrays the genome and gene content of *Aiptasia*, including the evolution of transposable elements, the potential role of *Aiptasia* taxonomically-restricted genes in the endosymbiosis, a novel family of phylum-specific, putative pattern-recognition receptors and evidence for extensive horizontal gene transfer between the *Aiptasia* host and the prokaryotic microbiome as well as the dinoflagellate symbiont. Chapter 3 presents evidence for the involvement of miRNA-mediated gene regulation in the onset and maintenance of the endosymbiosis by annotating the *Aiptasia* miRNA repertoire and analyzing their expression pattern between different states of *Symbiodinium* infection. By using UV-crosslinking co-immunoprecipitation (CLIP) of the central miRISC protein - Argonaute - functional molecular interactions of miRNAs and their cognate target mRNAs were identified. The integration of miRNA and mRNA expression data, together with the analysis of molecular miRNA-mRNA interactions, points towards specific roles of miRNAs in modulating the host transcriptome in the context of the cnidarian-dinoflagellate endosymbiosis in *Aiptasia.* This work also represents the

first approach of using CLIP in a non-bilaterian organism and highlights its powerful use in pinpointing towards specific RNA-protein interactions.

1.8 References

1. Falkowski P, Dubinsky Z, Muscatine L, Porter J (1984) Light and Bioenergetics of a Symbiotic Coral. *Bioscience* 34(11):705–709.
2. McCook LJ, et al. (2010) Adaptive management of the Great Barrier Reef: a globally significant demonstration of the benefits of networks of marine reserves. *Proc Natl Acad Sci U S A* 107(43):18278–18285.
3. Hughes TP, et al. (2007) Phase shifts, herbivory, and the resilience of coral reefs to climate change. *Curr Biol* 17(4):360–365.
4. Hughes TP, et al. (2003) Climate change, human impacts, and the resilience of coral reefs. *Science (80-)* 301(5635):929–933.
5. Hoegh-Guldberg O, et al. (2007) Coral reefs under rapid climate change and ocean acidification. *Science (80-)* 318(5857):1737–1742.
6. Hoegh-Guldberg O (2006) Ecology. Complexities of coral reef recovery. *Science (80-)* 311(5757):42–43.
7. Berumen ML, et al. (2013) The status of coral reef ecology research in the Red Sea. *Coral Reefs* 32(3):737–748.
8. Weis VM, Davy SK, Hoegh-Guldberg O, Rodriguez-Lanetty M, Pringle JR (2008) Cell biology in model systems as the key to understanding corals. *Trends Ecol Evol* 23(7):369–376.
9. Toller WW, Rowan R, Knowlton N (2001) Zooxanthellae of the *Montastraea annularis* species complex: patterns of distribution of four taxa of *Symbiodinium* on different reefs and across depths. *Biol Bull* 201(3):348–359.
10. Tonk L, Bongaerts P, Sampayo EM, Hoegh-Guldberg O (2013) SymbioGBR: a web-based database of *Symbiodinium* associated with cnidarian hosts on the Great Barrier Reef. *BMC Ecol* 13:7.
11. Finney JC, et al. (2010) The Relative Significance of Host–Habitat, Depth, and Geography on the Ecology, Endemism, and Speciation of Coral Endosymbionts in the Genus *Symbiodinium*. *Microb Ecol* 60(1):250–263.

12. Macdonald AH, Sampayo E, Ridgway T, Schleyer M (2008) Latitudinal symbiont zonation in *Stylophora pistillata* from southeast Africa. *Mar Biol* 154(2):209–217.
13. LaJeunesse T, et al. (2004) High diversity and host specificity observed among symbiotic dinoflagellates in reef coral communities from Hawaii. *Coral Reefs* 23(4):596–603.
14. LaJeunesse TC, et al. (2010) Long-standing environmental conditions, geographic isolation and host–symbiont specificity influence the relative ecological dominance and genetic diversification of coral endosymbionts in the genus Symbiodinium. *J Biogeogr* 37(5):785–800.
15. Rodriguez-Lanetty M, Loh W, Carter D, Hoegh-Guldberg O (2001) Latitudinal variability in symbiont specificity within the widespread scleractinian coral *Plesiastrea versipora*. *Mar Biol* 138(6):1175–1181.
16. Hume B, et al. (2013) Corals from the Persian/Arabian Gulf as models for thermotolerant reef-builders: Prevalence of clade C3 *Symbiodinium*, host fluorescence and ex situ temperature tolerance. *Mar Pollut Bull* 72(2):313–322.
17. Hume BCC, et al. (2015) *Symbiodinium thermophilum* sp. nov., a thermotolerant symbiotic alga prevalent in corals of the world's hottest sea, the Persian/Arabian Gulf. *Sci Rep* 5:8562.
18. Baker AC (2001) Reef corals bleach to survive change. *Nature* 411(6839):765–766.
19. Berkelmans R, Oppen M (2006) The role of zooxanthellae in the thermal tolerance of corals: a 'nugget of hope' for coral reefs in an era of climate change. *Proc R Soc B* 273:2305–2312.
20. Moya A, Ganot P, Furla P, Sabourault C (2012) The transcriptomic response to thermal stress is immediate, transient and potentiated by ultraviolet radiation in the sea anemone *Anemonia viridis*. *Mol Ecol* 21(5):1158–1174.
21. Lehnert EM, et al. (2014) Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3* 4(2):277–295.

22. Voolstra CR, et al. (2009) Effects of temperature on gene expression in embryos of the coral *Montastraea faveolata*. *BMC Genomics* 10:627.
23. Voolstra CR, et al. (2009) The host transcriptome remains unaltered during the establishment of coral-algal symbioses. *Mol Ecol* 18(9):1823–1833.
24. Rodriguez-Lanetty M, Wood-Charlson EM, Hollingsworth LL, Krupp DA, Weis VM (2006) Temporal and spatial infection dynamics indicate recognition events in the early hours of a dinoflagellate/coral symbiosis. *Mar Biol* 149(4):713–719.
25. Rodriguez-Lanetty M, Phillips WS, Weis VM (2006) Transcriptome analysis of a cnidarian-dinoflagellate mutualism reveals complex modulation of host gene expression. *BMC Genomics* 7:23.
26. Davy SK, Allemand D, Weis VM (2012) Cell biology of cnidarian-dinoflagellate symbiosis. *Microbiol Mol Biol Rev* 76(2):229–261.
27. Wood-Charlson EM, Hollingsworth LL, Krupp DA, Weis VM (2006) Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. *Cell Microbiol* 8(12):1985–1993.
28. Poole AZ, Weis VM (2014) TIR-domain-containing protein repertoire of nine anthozoan species reveals coral-specific expansions and uncharacterized proteins. *Dev Comp Immunol* 46(2):480–488.
29. Detournay O, Schnitzler CE, Poole A, Weis VM (2012) Regulation of cnidarian-dinoflagellate mutualisms: Evidence that activation of a host TGFbeta innate immune pathway promotes tolerance of the symbiont. *Dev Comp Immunol* 38(4):525–537.
30. Hambleton EA (2013) *Symbiosis specificity and innate immunity in Aiptasia, a model system for cnidarian-dinoflagellate symbiosis* (Stanford University).
31. Kvennefors EC, Leggat W, Hoegh-Guldberg O, Degnan BM, Barnes AC (2008) An ancient and variable mannose-binding lectin from the coral *Acropora millepora* binds both pathogens and symbionts. *Dev Comp Immunol* 32(12):1582–1592.

32. Kvennefors EC, et al. (2010) Analysis of evolutionarily conserved innate immune components in coral links immunity and symbiosis. *Dev Comp Immunol* 34(11):1219–1229.
33. Gruenberg J, van der Goot FG (2006) Mechanisms of pathogen entry through the endosomal compartments. *Nat Rev Mol Cell Biol* 7(7):495–504.
34. Chen MC, Cheng YM, Hong MC, Fang LS (2004) Molecular cloning of Rab5 (ApRab5) in *Aiptasia pulchella* and its retention in phagosomes harboring live zooxanthellae. *Biochem Biophys Res Commun* 324(3):1024–1033.
35. Chen M-C, Cheng Y-M, Sung P-J, Kuo C-E, Fang L-S (2003) Molecular identification of Rab7 (ApRab7) in Aiptasia pulchella and its exclusion from phagosomes harboring zooxanthellae. *Biochem Biophys Res Commun* 308(3):586–595.
36. Ganot P, et al. (2011) Adaptations to endosymbiosis in a cnidarian-dinoflagellate association: differential gene expression and specific gene duplications. *PLoS Genet* 7(7):e1002187.
37. Peng S-E, et al. (2010) Proteomic analysis of symbiosome membranes in Cnidaria-dinoflagellate endosymbiosis. *Proteomics*:NA–NA.
38. Dani V, Ganot P, Priouzeau F, Furla P, Sabourault C (2014) Are Niemann-Pick type C proteins key players in cnidarian-dinoflagellate endosymbioses? *Mol Ecol* 23(18):4527–4540.
39. Shinzato C, et al. (2011) Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* 476(7360):320–323.
40. Lehnert EM, Burriesci MS, Pringle JR (2012) Developing the anemone *Aiptasia* as a tractable model for cnidarian-dinoflagellate symbiosis: the transcriptome of aposymbiotic *A. pallida*. *BMC Genomics* 13:271.
41. Hamada M, et al. (2013) The complex NOD-like receptor repertoire of the coral *Acropora digitifera* includes novel domain combinations. *Mol Biol Evol* 30(1):167–176.

42. Schnitzler CE, Weis VM (2010) Coral larvae exhibit few measurable transcriptional changes during the onset of coral-dinoflagellate endosymbiosis. *Mar Genomics* 3(2):107–116.
43. Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell* 136(4):669–687.
44. Berezikov E (2011) Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* 12(12):846–860.
45. Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell* 136(4):642–655.
46. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10(2):94–108.
47. Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. *Cell* 136(4):656–668.
48. Carlsbecker A, et al. (2010) Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature* 465(7296):316–321.
49. Inui M, Martello G, Piccolo S (2010) MicroRNA control of signal transduction. *Nat Rev Mol Cell Biol* 11(4):252–263.
50. Kloosterman WP, et al. (2006) Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Res* 34(9):2558–2569.
51. O'Neill LA, Sheedy FJ, McCoy CE (2011) MicroRNAs: the fine-tuners of Toll-like receptor signalling. *Nat Rev Immunol* 11(3):163–175.
52. Baumgarten S, et al. (2013) Integrating microRNA and mRNA expression profiling in *Symbiodinium microadriaticum*, a dinoflagellate symbiont of reef-building corals. *BMC Genomics* 14(1):704.
53. Grimson A, et al. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455(7217):1193–1197.
54. Moran Y, Praher D, Fredman D, Technau U (2013) The Evolution of MicroRNA Pathway Protein Components in Cnidaria. *Mol Biol Evol* 30(12):2541–2552.
55. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136(2):215–233.

56. Moran Y, et al. (2014) Cnidarian microRNAs frequently regulate targets by cleavage. *Genome Res*. doi:10.1101/gr.162503.113.
57. Esau C, et al. (2006) miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab* 3(2):87–98.
58. Ghosh J, Bose M, Roy S, Bhattacharyya SN (2013) *Leishmania donovani* targets Dicer1 to downregulate miR-122, lower serum cholesterol, and facilitate murine liver infection. *Cell Host Microbe* 13(3):277–288.
59. Grawunder D, et al. (2015) Induction of Gametogenesis in the Cnidarian Endosymbiosis Model *Aiptasia sp. Sci Rep* 5:15677.
60. Thornhill DJ, Xiang Y, Pettay DT, Zhong M, Santos SR (2013) Population genetic data of a model symbiotic cnidarian system reveal remarkable symbiotic specificity and vectored introductions across ocean basins. *Mol Ecol* 22(17):4499–4515.
61. Voolstra CR (2013) A journey into the wild of the cnidarian model system *Aiptasia* and its symbionts. *Mol Ecol* 22(17):4366–4368.
62. Sunagawa S, et al. (2009) Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its dinoflagellate endosymbiont. *BMC Genomics* 10:258.
63. Tolleter D, et al. (2013) Coral bleaching independent of photosynthetic activity. *Curr Biol* 23(18):1782–1786.
64. Burriesci MS, Raab TK, Pringle JR (2012) Evidence that glucose is the major transferred metabolite in dinoflagellate-cnidarian symbiosis. *J Exp Biol* 215(19):3467–77.
65. Chapman JA, et al. (2010) The dynamic genome of *Hydra*. *Nature* 464(7288):592–596.
66. Putnam NH, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science (80-)* 317(5834):86–94.
67. Chourrout D, et al. (2006) Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* 442(7103):684–687.

68. Bosch TCG, David CN (1987) Stem cells of *Hydra magnipapillata* can differentiate into somatic cells and germ line cells. *Dev Biol* 121(1):182–191.
69. Boehm AM, Rosenstiel P, Bosch TCG (2013) Stem cells and aging from a quasi-immortal point of view. *BioEssays* 35(11):994–1003.
70. Sánchez Alvarado A, Yamanaka S (2014) Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell* 157(1):110–9.
71. Watanabe H, et al. (2014) Nodal signalling determines biradial asymmetry in Hydra. *Nature* 515(7525):112–5.
72. Leclère L, Rentzsch F (2014) RGM Regulates BMP-Mediated Secondary Axis Formation in the Sea Anemone *Nematostella vectensis*. *Cell Rep* 9(5):1921–30.
73. Bosch TC (2013) Cnidarian-microbe interactions and the origin of innate immunity in metazoans. *Annu Rev Microbiol* 67:499–518.
74. Bosch TC, et al. (2009) Uncovering the evolutionary history of innate immunity: the simple metazoan Hydra uses epithelial cells for host defence. *Dev Comp Immunol* 33(4):559–569.
75. Meyer E, Weis VM (2012) Study of cnidarian-algal symbiosis in the “omics” age. *Biol Bull* 223(1):44–65.

Chapter 2: The genome of *Aiptasia,* a sea anemone model for coral symbiosis

Sebastian Baumgarten[a], Oleg Simakov[b+], Lisl Y. Esherick[c], Yi Jin Liew[a], Erik M. Lehnert[c], Craig T. Michell[a], Yong Li[a], Elizabeth A. Hambleton[b], Annika Guse[b], Matt E. Oates[d], Julian Gough[d], Virginia M. Weis[e], Manuel Aranda[a], John R. Pringle[c], Christian R. Voolstra[a]

[a] Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[b] Centre for Organismal Studies, Heidelberg University, 69120 Heidelberg, Germany

[c] Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[d] Department of Computer Science, University of Bristol, Bristol BS8 1UB, United Kingdom

[e] Department of Integrative Biology, Oregon State University, Corvallis, Oregon 97331, USA

[+] Current address: Okinawa Institute of Science and Technology, Okinawa 904-0495, Japan

2.1 Abstract

The most diverse marine ecosystems, coral reefs, depend upon a functional symbiosis between a cnidarian animal host (the coral) and intracellular photosynthetic dinoflagellate algae. The molecular and cellular mechanisms underlying this endosymbiosis are not well understood, in part because of the difficulties of experimental work with corals. The small sea anemone *Aiptasia* provides a tractable laboratory model for investigating these mechanisms. Here we report on the assembly and analysis of the *Aiptasia* genome, which will provide a foundation for future studies and has revealed several features that may be key to understanding the evolution and function of the endosymbiosis. These features include genomic rearrangements and taxonomically restricted genes that may be functionally related to the symbiosis, aspects of host dependence on alga-derived nutrients, a novel and expanded cnidarian-specific family of putative pattern-recognition receptors that might be involved in the animal-algal interactions, and extensive lineage-specific horizontal gene transfer. Extensive integration of genes of prokaryotic origin, including genes for antimicrobial peptides, presumably reflects an intimate association of the animal-algal pair also with its prokaryotic microbiome.

2.2 Introduction

Coral reefs form marine-biodiversity hotspots that are of enormous ecological, economic, and aesthetic importance. Coral growth and reef deposition are based energetically on the endosymbiosis between the cnidarian animal hosts and photosynthetic dinoflagellate algae of the genus *Symbiodinium*, which live in vesicles within the gastrodermal (gut) cells of the animal and typically supply ≥90% of its total energy, while the host provides the algae with a sheltered environment and the inorganic nutrients needed for photosynthesis and growth (1). This tight metabolic coupling allows the holobiont (i.e., the animal host and its microbial symbionts) to thrive in nutrient-poor waters. Although the ecology of coral reefs has been studied intensively, the molecular and cellular mechanisms underlying the critical endosymbiosis remain poorly understood (2). As coral reefs face an ongoing and increasing threat from anthropogenic environmental change (3), new insights into these mechanisms are of critical importance to understanding the resilience and adaptability of coral reefs and thus to the planning of conservation strategies (4).

Aiptasia is a globally distributed sea anemone that harbors endosymbiotic *Symbiodinium* like its Class Anthozoa relatives the stony corals (Figs. 2.1 and 2.2) (4, 5). *Aiptasia* has a range of polyp sizes convenient for experimentation and is easily grown in laboratory culture, where it reproduces both asexually (so that large clonal populations can be obtained) and sexually (allowing experiments on larvae and potentially genetic studies), and it can be maintained indefinitely in an aposymbiotic (dinoflagellate-free) state and reinfected by a variety of *Symbiodinium* strains (6, 7). These characteristics make *Aiptasia* a highly

attractive model system for studies of the molecular and cellular basis of the cnidarian-dinoflagellate endosymbiosis (2, 4). To provide a solid platform for research on *Aiptasia*, we have sequenced and analyzed its genome. The results have already provided important insights into several aspects of the evolution and function of the symbiotic system.

2.3 Results and Discussion

2.3.1 Genome size and assembly

We used flow cytometry to estimate the haploid genome size of *Aiptasia*, obtaining a value of ~260 Mb (Fig. 2.3A, 1-3). This value is smaller than those reported previously for three other cnidarians (Table 2.1), including two other anthozoans, the anemone *Nematostella vectensis* (450 Mb (8)) and the stony coral *Acropora digitifera* (420 Mb (9)). However, we estimated a genome size of ~329 Mb for *N. vectensis* by flow cytometry (Fig. 2.3A, 4-6), and a previous independent estimate by densitometry of Feulgen-stained embryonic nuclei gave an even lower value of ~220 Mb (10).

To generate a genome sequence, we used DNA from a clonal population of aposymbiotic anemones to obtain ~140x coverage by Illumina short-read sequencing plus additional sequence from long-insert mate-pair libraries (Table 2.2). From these sequences we assembled a draft genome of ~258 Mb comprised of 5,065 scaffolds and 29,410 contigs, with half of the genome found in scaffolds longer than 440 kb and in contigs longer than 14.9 kb (Tables 2.3 and 2.4). Using *ab initio* prediction informed by a reference transcriptome

covering several developmental and symbiotic states (Tables 2.5 and 2.6; Fig. 2.2C), we identified 29,269 gene models, of which 26,658 appear to be complete (Table 2.3) and 27,469 (including 25,370 of the seemingly complete genes) were supported by RNA evidence (Table 2.7). Of the 29,269 gene models, 26,162 (89%) have identifiable homologues in other eukaryotes and 22,993 have annotations in the UniProt database (Table 2.7; Dataset S2.1). The *Aiptasia* genome assembly and gene predictions compare well to those in the other sequenced cnidarians (Table 2.1), and this comparison also suggests that the smaller genome size of *Aiptasia* (see above) reflects largely the smaller sizes of introns and their lower frequency (as indicated by a larger mean exon size) in *Aiptasia* (Table 2.1). Indeed, as expected, the diversity of molecular functions represented in the predicted *Aiptasia* protein set is comparable to those in other cnidarians and in other 'basal' metazoans (Fig. 2.3B-C) (11).

2.3.2 Repeat content and evolution, synteny, and fast-evolving genes

The *Aiptasia* genome contains ~26% repeated sequences, intermediate between what has been reported for other cnidarians (Table 2.1) and similar to the estimates for other 'basal' metazoans such as *Capitella teleta* (31%) and *Lottia gigantea* (20%) (11, 12). ~36% of the repeated sequences can be assigned to previously known repetitive elements (Table 2.8) that are in multiple categories including a variety of known transposable elements (TEs). We found at least four classes of TEs that showed an expansion at a Jukes-Cantor distance of 0.4-0.5 (Figs. 2.4A and 2.5A,B). This ancient TE expansion aligns with the time of

divergence of *Aiptasia* from *Verrilactis paguri* and *Sagartia elegans*, two anemone species that are not symbiotic with dinoflagellates, based on Nei-Gojobori synonymous-substitution rates in the *COX3* gene (Fig. 2.4A). Such burst-like expansions of TEs have been reported to reflect ancient population bottlenecks (13) and/or speciation events, as well as major increases in genome size, such as that found in *Hydra* (14). Remarkably, an ancient expansion of TEs in the stony coral *A. digitifera* resembles that in *Aiptasia*, whereas the more closely related *N. vectensis* (which is not symbiotic with dinoflagellates) appears not to have had such an expansion (Fig. 2.5). The existence of similar repeat expansions in *Aiptasia* and *A. digitifera* is intriguing, and it remains to be determined if these genomic rearrangements were functionally associated with some common event affecting their symbiotic lifestyles.

Despite the potential genome-rearranging effects of such TE activity (11), a total of 3,377 *Aiptasia* genes are found in synteny blocks of 3 to 33 genes (mean of 4.8) that are shared with one or more other metazoans (Materials and Methods). This is similar to previously reported findings with other animals (11). About half of these synteny blocks, containing 1,727 genes, are shared only with one or more other cnidarians. Interestingly, although many of the *Aiptasia* HOX genes are in clusters, as in other metazoans, the detailed organization of the clusters is distinct even from that in other anthozoans, leaving the ancestral organization of the HOX cluster uncertain (15, 16) (Fig. 2.4B, Dataset S2.2).

To seek further insights into *Aiptasia* evolution and symbiosis, we identified the gene families with elevated rates of amino-acid substitution relative to other

cnidarians (Materials and Methods). Among 2,478 gene families, 143 (containing 165 genes) were found to show accelerated evolution (Dataset S2.3), but analysis of the functions ascribed to these families did not reveal obvious clues to the molecular basis of the endosymbiotic relationship.

2.3.3 Taxonomically restricted genes

We found 3,107 genes whose predicted products had no discernible homologues in other organisms (Table 2.7). Of these, 1,946 were supported by transcriptomic evidence. Such taxonomically restricted genes (TRGs) are thought to play important roles in the creation of evolutionary novelties and morphological diversity within the taxonomic group (17, 18). Although the TRG products lack similarity to known proteins from which putative functions could be derived, we found that many of the TRGs were expressed differentially in different states of symbiosis (Figs. 2.6A and 2.7A) and/or between adults and larvae (Fig. 2.7A), suggesting that at least some are real genes with specific functions, possibly in *Aiptasia*-specific aspects of endosymbiosis or development.

To search further for homologies of the TRGs, a Hidden-Markov-Model analysis was conducted (Materials and Methods), identifying 52 putatively new protein domains. These domains were found in 122 of the putative TRGs, including 14 cases in which the two genes encoding the same seemingly new domain were immediately adjacent to each other or separated by a single gene, suggesting recent gene duplication. In addition, these putatively new domains were found in 145 *Aiptasia* proteins that were not classified as TRG products (Dataset S2.4).

Also of interest was that the putative TRGs were not randomly distributed in the genome, with about one third being present in clusters of two to seven with no interspersed genes (Fig. 2.6B). The genes in these clusters were not co-expressed (Fig. 2.7), and the biological reason for the non-random genomic arrangement remains unclear.

2.3.4 Metabolic exchanges between the partners

The cnidarian-dinoflagellate symbiosis involves an intimate metabolic exchange between the partners, but many of the details remain obscure. In particular, the respective contributions to animal metabolism from ingested food, synthesis by *Symbiodinium,* synthesis by the prokaryotic microbiome, and synthesis by the animal cells themselves remain unclear in many cases. The *Aiptasia* genome sequence should help to elucidate these matters, as the following examples illustrate. First, BLASTP searches of the gene models confirm previous transcriptomic evidence (19) that anthozoans, like other animals, lack the ability to synthesize 12 of the 20 common amino acids from central-pathway intermediates (Table 2.9). The missing enzymes include two that are essential for the synthesis of the sulfur-containing amino acids, but the genome also confirms that *Aiptasia*, like other animals including *N. vectensis* (Table 2.9) and several corals (9), contains a gene for a cystathionine-β-synthase, so that it should be able to synthesize cysteine from methionine (19). Thus, the reported absence of such a gene in several *Acropora* species (9) remains a puzzle. Although labeling studies have indicated that various essential amino acids are transported from

the dinoflagellate to the host (20), it remains unclear if this supply is adequate for the animals' needs or whether additional supplies from ingested food are also necessary.

Second, analysis of the genome sequence confirmed the presence of five genes (19) encoding Npc2-type sterol-transport proteins and identified a sixth such gene. Interestingly, five of these genes occur in two clusters that probably reflect their evolutionary origins by tandem duplication (Fig. 2.8A, Dataset S2.5). The two genes encoding proteins with typical cholesterol-binding sites (*NPC2A* and the newly identified *NPC2F*) are closely linked and have three introns apiece, whereas three of the four genes encoding proteins that presumably cannot bind cholesterol (19) are present in a second cluster and have either zero or one intron. The latter genes may represent an anthozoan-specific expansion of the family (19), and both differential expression (19, 21) and protein-localization (22) data suggest that the product of one of these genes (*NPC2D*) is involved in transport of dinoflagellate-synthesized sterols to the host cytosol in symbiotic anemones. Such transport may provide the host with the bulk sterols that it needs for membrane formation, and indeed analysis of the predicted protein sets of *Aiptasia, N. vectensis,* and *A. digitifera* indicates that anthozoans (like *Drosophila*, but unlike both mammals and *Symbiodinium*) are unable to synthesize sterols from glycolytic intermediates (Fig. 2.8B; Dataset S2.6). However, most dinoflagellate-synthesized sterols are structurally distinct from the bulk of those found in the cnidarian hosts (23–26), and it remains unclear whether the dinoflagellate and host can transport and biochemically convert the

dinoflagellate-produced sterols in sufficient quantity to fulfill the host's needs in the absence of a supply also from ingested food. As suggested previously (19), an intriguing alternative would be that the transported sterols serve a role in host recognition of the algal symbionts.

2.3.5 Interactions of the host with algal symbionts and other microbes

A cnidarian host must distinguish among potential symbionts, pathogens, and particles of food. Moreover, a given host can establish symbiosis with some strains of *Symbiodinium* but not others (6, 7), and there is good evidence from *Hydra* that cnidarian hosts also actively shape the assembly of a specific and beneficial prokaryotic microbiome (27, 28). Although understanding of microbiome assembly in corals and anemones is much less complete, recent studies suggest that processes similar to those in *Hydra* are involved (29–32). Such discriminations must be accomplished in the absence of an adaptive immune system and presumably depend on innate-immunity mechanisms that involve the recognition of microbial cell-surface molecules by host pattern-recognition receptors (PRRs) (2, 33). Indeed, there is evidence that the recognition of compatible *Symbiodinium* types depends on the specific binding of algal cell-surface glycans by host lectins (34).

Thus, it was of interest that among the significantly enriched protein domains in the *Aiptasia* genome (Fig. 2.9A) was the fibrinogen domain, which acts as a carbohydrate-binding moiety in secreted vertebrate PRRs that trigger the innate-immunity cascade of the lectin-complement pathway (35). Searching the Pfam

annotation (see Materials and Methods) revealed 116 *Aiptasia* proteins containing fibrinogen domains, of which 13 also contain an N-terminal collagen domain as in the bilaterian ficolins (35). Strikingly, not only has this protein family undergone apparently independent expansions in the symbiotic anemone and stony-coral lineages (Fig. 2.10A), but most family members (including all identified to date in *Aiptasia*) also contain two or three immunoglobulin (Ig) domains lying between the collagen and fibrinogen domains (Fig. 2.10A). As proteins with this tripartite domain structure appear to be absent from bilaterians and have not previously been described, we have named them Cnidarian ficolin-like proteins (CniFLs), and we speculate that the Ig domains may contribute to their ability to recognize a variety of microbial surface patterns with high specificity. Interestingly, a similar expansion of proteins containing both Ig-superfamily and fibrinogen domains (FREPs) has been identified in the snail *Biomphalaria glabrata*, in which the Ig domains are not only expanded in the genome but further diversified through point mutations and somatic recombination of the germline source sequences (36). Even though the Ig-domain diversity in CniFLs and FREPs falls far short of that in the antibody system of vertebrates, our findings further support the hypothesis that the pattern-recognition capabilities of invertebrate innate-immunity systems are more flexible than once thought.

Both analogy to the function of vertebrate PRRs and previous studies of complement components in cnidarians (37, 38) suggest that the *Aiptasia* CniFLs might function through the complement pathway, and indeed we were able to

identify putative orthologs of most components of this pathway in the *Aiptasia* predicted protein set (Fig. 2.10B, Dataset S2.8). Accordingly, a CniFL-complement pathway might be involved in the recognition of compatible *Symbiodinium* types, in shaping other aspects of the host microbiome, or both. Consistent with the former possibility, we note (i) that the CniFLs have so far been found only in symbiotic cnidarians (*Aiptasia* and two corals, but not *N. vectensis*) and (ii) that the *Aiptasia* CniFLs were almost all upregulated – in some cases dramatically so – in aposymbiotic relative to symbiotic anemones (Table 2.10). This upregulation might reflect a role in recognition and uptake of compatible *Symbiodinium* cells by animals who lack them, with downregulation following the successful establishment of symbiosis.

Toll-like receptors (TLRs) and Interleukin-like receptors (ILRs) are additional PRR classes; both recognize extracellular microbial patterns and signal through the intracellular Toll/Interleukin-receptor (TIR) domain. Consistent with a previous transcriptome study (39), the current *Aiptasia* genome assembly has not revealed any canonical TLR with both TIR and extracellular leucine-rich-repeat (LRR) domains, but it does contain four ILR homologues with both a TIR domain and one to three extracellular Ig domains, as well as two proteins with a TIR domain only (Table 2.11; Fig. 2.9B), which may partner with an unknown extracellular PRR. Consistent with previous studies (39, 40), many other predicted components of the TLR/ILR signaling pathways were identified using the KEGG-annotated protein sets of *Aiptasia, N. vectensis,* and *A. digitifera* (Fig. 2.9C, Dataset S2.9), suggesting that these pathways are functional in

anthozoans. The close similarity between the *Aiptasia* and *N. vectensis* TLR/ILR protein repertoires and the apparently lineage-specific expansion of these proteins in *A. digitifera* (Table 2.11; Fig. 2.9B), but not in *Aiptasia,* suggest that these pathways are not involved in the interaction between host and *Symbiodinium* strain. Instead, as suggested by studies in *Hydra* (41), these proteins might function in shaping the host-associated prokaryotic microbiome. The microbiome might be more complex in corals than in anemones, or coral and anemone hosts might rely on different sets of mechanisms to shape their microbiomes.

Microbes that evade the extracellular recognition systems and invade the animal cell can be recognized by intracellular "nucleotide-binding and oligomerization domain" (NOD)-like receptors (NLRs), which trigger defensive responses overlapping with those of the TLRs and ILRs (27) (Fig. 2.9C,D). NLRs are characterized by a central NACHT or NB-ARC domain and are greatly expanded and diversified in cnidarians (42, 43). A search of the *Aiptasia* Pfam annotation revealed 86 proteins containing NACHT domains and 22 proteins containing NB-ARC domains, numbers similar to those reported for *N. vectensis* but much smaller than those reported for both *H. magnipapillata* and *A. digitifera*. It is not clear what could explain these substantial lineage-specific differences. As reported previously for *H. magnipapillata* (42), KEGG-based analysis revealed the presence in *Aiptasia* and other anthozoans of homologs of many components of the vertebrate NLR-triggered pathways but, surprisingly, not of the centrally important "receptor-interacting protein kinases" RIPK1 and RIPK2 (Fig. 2.9D;

Dataset S2.10). This result suggests that cnidarians have a novel mechanism for the processing of signals resulting from NLR recognition of microbial patterns.

2.3.6 Evidence for extensive horizontal gene transfer

The associations between cnidarians and their endosymbiotic dinoflagellates and prokaryotic microbiomes have evolved over millions of years, making it likely that horizontal gene transfer (HGT) has occurred (44). To explore this possibility, we searched the predicted *Aiptasia* protein set for cases in which the best alignment was to a non-metazoan rather than a metazoan sequence (Materials and Methods). This search identified 275 HGT candidates that are specific to *Aiptasia* (Fig. 2.11A, left) and an additional 548 "cnidarian-specific" cases in which the *Aiptasia* protein has a hit in one or more other cnidarians, but not in other metazoans, so that the gene is likely to have been transferred from a non-metazoan source to a basal cnidarian (Fig. 2.11A, right, Dataset S2.11). Although the HGT candidates have a variety of apparent sources, genes of putative bacterial origin predominate (Fig. 2.11A). Consistent with origins in bacterial sequences (and/or elements arising by reverse transcription) and a subsequent gradual accumulation of introns, both the *Aiptasia*-specific and cnidarian-specific HGT candidates have, on average, fewer introns than the overall *Aiptasia* gene set (Fig. 2.12A). Similar patterns have been observed for HGT candidates in *H. magnipapillata*, *N. vectensis*, and the rotifer *Adineta vaga* (14, 45, 46). As the *Aiptasia*-specific HGT candidates also have, on average, fewer introns than the cnidarian-specific HGT candidates (Fig. 2.12A), the

Aiptasia lineage has presumably continued to acquire new genes by horizontal transfer since its split from the other cnidarian lineages examined here (rather than these genes' having been lost in the other lineages).

Among the HGT candidates were 29 (17 *Aiptasia*-specific and 12 cnidarian-specific) whose best alignment was to a *Symbiodinium* protein (Fig. 2.11A), suggesting that they were transferred from *Symbiodinium* or another dinoflagellate(s) into cnidarian hosts (or possibly the reverse). [Because the databases used (Materials and Methods) contained sequences from *Symbiodinium* but not other dinoflagellates, there is ambiguity about the dinoflagellate(s) involved.] To test this possibility further, we constructed maximum-likelihood phylogenetic trees for four of the 12 cnidarian-specific proteins and sequences representing the full phylogenetic diversity in the UniProt database (Materials and Methods). In each case, the *Aiptasia* protein and its homologs in other cnidarians and *Symbiodinium* form a distinct clade that is itself embedded in a larger clade containing almost entirely proteins from non-metazoans (Fig. 2.11B), supporting the hypothesis of a dinoflagellate-to-cnidarian HGT (or *vice versa*).

One of the 12 cnidarian-specific genes has already been identified in previous studies as a probable case of HGT (9, 47). In *N. vectensis, A. digitifera,* and several dinoflagellates, a single gene was found to encode a 3-dehydroquinate synthase (3-DHQS)-like domain fused to an *O*-methyltransferase (O-MT)-like domain; these domains appear to catalyze consecutive steps in the synthesis of UV-protective mycosporine amino acids (48). Both domains from these

dinoflagellate and cnidarian fusion proteins cluster in phylogenetic trees with each other and with the corresponding domains found in several cyanobacteria (9, 47) (Fig. 2.11C,D), although the domains are encoded in distinct but adjacent genes in those cyanobacteria (49). Indeed, our search of 89 cyanobacterial genomes (https://img.jgi.doe.gov/) for gene models containing a 3-DHQS domain (as annotated by Pfam) revealed 119 genes, none of which contained a fused O-MT domain. We also found that both domains of the predicted *Aiptasia* fusion protein cluster in phylogenetic trees as described previously for the *N. vectensis* and *A. digitifera* proteins, and that *Symbiodinium microadriaticum* also contains a gene encoding both domains with similar sequences (Fig. 2.11C,D). As the same fusion-protein gene is present in both dinoflagellates and cnidarians, and the cnidarian sequences of both domains are more similar to those in dinoflagellates than to those in cyanobacteria, it seems likely that the gene fusion occurred just once and simultaneously with the transfer from a cyanobacterium into a basal dinoflagellate (49), with a subsequent transfer into one or more cnidarian ancestors from a dinoflagellate(s) that was potentially symbiotic (providing the intimate association that would facilitate the gene transfer). On this model, the presence of the fusion gene in *N. vectensis* can be explained most readily by the hypothesis that the modern species had an ancestor with a symbiotic dinoflagellate, a possibility that is supported by the fact that at least nine of the 11 additional cnidarian-specific HGT candidates whose best alignment was to *Symbiodinium* (see above) are present in *N. vectensis.*

Among the 17 *Aiptasia*-specific HGT candidates whose best alignment was to a *Symbiodinium* protein, seven had domains that were annotated through the Pfam database. Three of these annotations were to the bacterial Tox-Art-HYD1 domain of ~100 amino acids, which is known to act as an ADP-ribosyltransferase toxin in several bacterial pathogens (50, 51) but has not previously been described in eukaryotic genomes (although one crustacean sequence is annotated in the Pfam database). In addition to the three genes found in the HGT screen, we found three more *Aiptasia* genes that contain Tox-Art-HYD1 domains based on Pfam annotation. Moreover, three of these six Tox-Art-HYD1 genes are present in one genomic region of ~64 kb without intervening genes, and these three encode proteins that are particularly closely related in sequence (Fig. 2.11E). Thus, it appears that Tox-Art-HYD1 domain-containing genes have undergone intra-specific duplications in *Aiptasia* after the initial HGT event.

Key residues of the Tox-Art-HYD1 domain are conserved in *Aiptasia* and *Symbiodinium* (Fig. 2.11E), and in a phylogenetic analysis, the *Aiptasia* and *Symbiodinium* domains formed a distinct clade embedded within a large clade of bacterial sequences (Fig. 2.12B), supporting the hypothesis of a *Symbiodinium*-*Aiptasia* HGT event following an initial transfer of a bacterial gene into one partner or the other. Bacterial toxins containing Tox-Art-HYD1 domains are thought to function in interspecific conflicts (51), and the *Aiptasia* and *Symbiodinium* proteins may be used similarly to combat pathogens and/or shape the composition of the host prokaryotic microbiome, much as species-specific antimicrobial peptides in *Hydra* appear to help shape its microbiome (52).

2.4 Conclusions

Recent years have brought dramatically increased appreciation of the importance of understanding the molecular, cellular, and organismal bases of mutualistic host-microbe symbioses. Here we have reported on the assembly and analysis of the genome sequence and predicted protein sets of the sea anemone *Aiptasia*. Although differences are likely to exist between different symbiotic anthozoans, our analyses have revealed a variety of conserved features that should help to illuminate the evolution of the symbiotic lifestyle and provide the basis for the continued development of *Aiptasia* as a critically needed model for coral-dinoflagellate endosymbiosis, which underlies one of the most important marine ecosystems, coral reefs.

2.5 Materials & Methods

2.5.1 Organisms

Anemones of the clonal *Aiptasia* strains CC7 (53) and H2 (54) were used in this study, with H2 animals used only in the crosses to produce larvae for transcriptome analyses (see below). Although CC7 and H2 have been described previously as *A. pallida* and *A. pulchella*, respectively, recent global genotyping has found two genetically distinct *Aiptasia* populations that do not conform to previous species descriptions (5). Thus, we refer to both strains simply as *Aiptasia* in this paper.

Anemones were grown as described previously (19). Briefly, animals were raised in a circulating artificial seawater (ASW) system at ~25ºC with 20-40 µmol photons m^{-2} s^{-1} of photosynthetically active radiation on a 12 h:12 h light:dark cycle and fed freshly hatched *Artemia* (brine-shrimp) nauplii approximately twice per week. To generate aposymbiotic anemones (*i.e.,* animals without dinoflagellate symbionts), anemones were placed in a polycarbonate tub and subjected to multiple repetitions of the following cycle: cold-shocking by addition of 4ºC ASW and incubation at 4ºC for 4 h, followed by 1-2 d at ~25ºC in ASW containing the photosynthesis inhibitor diuron (Sigma-Aldrich #D2425) at 50 µM (lighting approximately as above). The putatively aposymbiotic anemones were then maintained for ≥1 month in ASW at ~25ºC in the light (as above) with feeding (as above, with water changes on the days following feeding) to test for possible repopulation of the animals by any residual dinoflagellates. The

anemones were then inspected individually by fluorescence stereomicroscopy to confirm the complete absence of dinoflagellates, whose bright chlorophyll autofluorescence is conspicuous when they are present (Fig. 2.1D). The anemones used as sources of genomic DNA were from tubs in which all of the animals were lacking dinoflagellates as assessed in this way.

To generate symbiotic anemones for transcriptome analysis (see below), adult aposymbiotic CC7 animals were infected with the compatible Clade B *Symbiodinium* strain SSB01, as originally isolated from *Aiptasia* strain H2 (54). To clarify the taxonomic position of SSB01, the microsatellite marker *Sym15* was amplified and sequenced as described previously (55). Using BLASTN, the SSB01 *Sym15* was found to align best to *Sym15* from *S. minutum* (Clade B) genotypes FLAp2 (5) and rt002 (56) (E-values, 5e-100 and 4e-96, respectively), and less well to Sym15 from the sister species *S. psygmophilum* (56) (E-value, 9e-63). The sequence of the *cp23S* marker gene from SSB01 (54) is also highly similar to that of *S. minutum* (56). Maximum parsimony trees of the *Sym15* and *cp23S* markers from various strains confirm that SSB01 represents a genotype of the species *S. minutum* (Fig. 2.2B).

2.5.2 Estimations of genome size

To estimate the genome size of *Aiptasia* and compare it to that of the related (but not symbiotic with dinoflagellates) anemone *Nematostella vectensis* (Fig. 2.1A), nuclear DNA contents were measured using a fluorescence-activated cell sorter with chicken red-blood cells (CRBC: Innovative Research #50-176-866) as a well

characterized reference of known genome size (2C DNA content of 2.33 pg or 2,279 Mb). Extraction and staining of nuclei were performed using the Partec CyStainPI absolute T kit (Partec #05-5023) following the manufacturer's protocol. Briefly, cell lysates of seven aposymbiotic *Aiptasia* anemones (~0.5 cm long) and two *N. vectensis* anemones (~3 cm long) were prepared in 1 ml apiece of nuclei-isolation buffer using a plastic mortar and pestle. To isolate CRBC nuclei, 70 µl of cells were added to 1 ml of nuclei-isolation buffer without further mechanical grinding by mortar and pestle. The resulting cell lysates were incubated for 15 min at 22ºC and filtered through a 40-µm mesh. After staining of both separate and mixed (*Aiptasia* plus CRBC; *Aiptasia* plus *N. vecte*nsis) cell lysates, their fluorescence signals were measured with a BD FACSCanto II cell analyzer (BD Bioscience).

2.5.3 Isolation of genomic DNA, library preparation, and sequencing

To isolate high-molecular-weight genomic DNA (gDNA), aposymbiotic animals (see above) were processed using a Qiagen Genomic-tip 100/G kit (Qiagen #10243) following the instructions for extraction of gDNA from tissues. Briefly, ~10 medium-sized animals (~50-60 mg total wet weight) were homogenized in 9.5 ml of buffer G2 containing 400 µg ml^{-1} RNase A for ~10 s in a PowerGen 125 rotor stator (Fisher Scientific) at 30,000 rpm. Proteinase K (Qiagen #19133) was added to a final concentration of 1 mg/ml, and the homogenate was incubated at 50°C for 2 h, centrifuged at 5,000 x *g* for 10 min at 4°C to remove particulates, and loaded onto an equilibrated Qiagen Genomic-tip 100/G. The sample was

then washed and eluted following Qiagen's's instructions. The quality of the isolated DNA was assessed by agarose-gel electrophoresis, and the DNA was quantified using a Qubit 2.0 Fluorometer (Life Technologies).

Two libraries designed to yield overlapping paired-end sequences were prepared using the NEBNext Ultra DNA Library Prep Kit (NEB #E7370) with insert sizes of ~180 and ~550 bp. These libraries were sequenced on one lane of an Illumina HiSeq2000 sequencer (180-bp library) and one lane of an Illumina MiSeq sequencer (550-bp library) with read lengths of 2 x 101 bp and 2 x 300 bp, respectively. In addition, two mate-pair libraries with insert sizes of 10 and 12 kb were prepared using the Nextera Mate Pair Sample Preparation Kit (gel-plus protocol) (Illumina #FC-132-1001) and sequenced on one lane of an Illumina MiSeq sequencer at a read length of 2 x 101 bp. Finally, two mate-pair libraries with insert sizes of 2 and 5 kb were prepared at the Beijing Genome Institute using DNA that we supplied and sequenced on one lane of an Illumina GAIIx sequencer with a read length of 2 x 90 bp. The libraries used and sequences obtained are summarized in Table 2.2. All raw sequences have been deposited in the NCBI Sequence-Read Archive (SRA) as FASTQ files under the accession number SRS742511.

Additional high-quality genomic sequence was obtained but not used in the final assembly because it was found not to improve the assembly. These sequences have also been deposited in SRA as FASTQ and SFF raw-sequence files. First, one paired-end library of ~250-bp insert size was prepared with the TruSeq DNA Sample Prep Kit (Illumina #FC-121-2003) and sequenced both in one lane on an

Illumina GAIIx sequencer at 2 x 101-bp read length (accession number SRR576330) and in one lane on an Illumina HiSeq2000 at 2 x 101-bp read length (accession number SRR576326). Second, two libraries of ~250-bp insert size were prepared with the Nextera DNA Sample Prep Kit (Illumina #FC-121-1031); each library was then sequenced in two lanes on the HiSeq2000 at 2 x 101-bp read length (accession numbers SRR606428, SRR646473, and SRR646474; SRR646474 contains the reads from two lanes). Third, three reduced-representation libraries (57) were produced by digesting with a restriction enzyme (*Dra*I, *Bam*HI, or *Pvu*II), selecting for fragments of 1-5 kb, preparing libraries (with separate barcodes) with the Nextera Kit, pooling, and sequencing in a single lane on the HiSeq2000 at 2 x 101-bp read length (accession number SRR609200). Finally, four libraries were sequenced on a 454FLX sequencer (accession number SRX757524).

2.5.4 Genome assembly and removal of contaminating sequences

Sequences from the libraries used for assembly were preprocessed with Trimmomatic (58) to trim sequencing adaptors and filter out low-quality reads. The genome assembly was then performed using ALLPATH-LG (59) with default settings and including the HAPLOIDIFY=True setting. This *de novo* assembly incorporated error correction of reads, contig assembly from the paired-end reads, and final scaffolding using the mate-pair information. Gaps in the resulting genome assembly were closed using 10 iterations of GapFiller (60). A super-scaffolding step was then performed on the gap-filled assembly using SSPACE

(61), followed by 10 additional iterations of GapFiller to close gaps introduced during the super-scaffolding step. These assembly-refinement steps used the error-corrected paired-end and mate-pair reads that were also used in the initial assembly.

To identify and remove scaffolds that were likely to have originated from dinoflagellate, bacterial, or viral contaminants, we used a custom script to conduct BLASTN searches against six databases: the genomes of *S. minutum* (62) and *S. microadriaticum* (http://reefgenomics.org/blast); the NCBI complete-bacterial-genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz), draft-bacterial-genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/), and complete-viral-genomes (ftp://ftp.ncbi.nih.gov/genomes/Viruses/all.fna.tar.gz) databases; and the viral database PhAnToMe (phantome.org). As the lengths of the query and hit sequences ranged up to hundreds of kilobases, a combination of cutoffs (total bit score >1000, E-value ≤1e-20) was used to identify contigs with significant similarities to sequences in the databases. Five scaffolds that displayed significant similarity over ≥50% of their non-N sequences were considered to have originated from bacterial contaminants (Table 2.12) and were thus removed from the final assembly.

The basic statistics for contigs and scaffolds were assessed as in (63) using the perl script provided by the Korf laboratory (http://korflab.ucdavis.edu/datasets/Assemblathon/Assemblathon2/Basic_metrics/assemblathon_stats.pl). Briefly, assembled scaffolds were broken into contigs if

they were separated by one or more stretches of >25 N's, and statistics were calculated separately for scaffolds and contigs.

2.5.5 Identification of repeats and analyses of transposable-element activity

RepeatScout (64) was used to conduct *de novo* identification of repeat regions in the genome assembly with an l-mer size of l = 15 bp. Using the default settings, 4,295 distinct repeat motifs were identified that occurred ≥10 times. Annotation of these repeats was performed as described previously (11) using three different methods: (i) RepeatMasker (65) using RepBase version 19.04; (ii) TBLASTX against RepBase version 19.04; and (iii) BLASTX against a custom-made non-redundant database of proteins encoded by transposable elements (TEs; NCBI keywords: retrotransposon, transposase, reverse transcriptase, gypsy, copia). The best annotation among the three methods was chosen based on alignment coverage and score. Both the repeat motifs identified in this way and the set of known eukaryotic TEs from RepBase (May 2014 release) were then used to locate and annotate the repeat elements in the assembled genome using RepeatMasker (65). Repeat annotation (Table 2.8) is available as a track on the *Aiptasia* Genome Browser (http://aiptasia.reefgenomics.org/jbrowse).

From the RepeatMasker output, the timing of TE activity was computed as described previously (11). Briefly, the ages of insertions were determined for repeats ≥100 bp in length by determining the total number of divergent positions (excluding gaps) of the genomic sequence compared to the consensus sequence for that repeat family. This number was subsequently adjusted using the Jukes-

Cantor formula (K=(3/4)*ln[1-i*(4/3)]) to account for multiple substitutions. For comparison, the same repeat annotation and analysis of TE activity were performed for the genomes of *N. vectensis* and *A. digitifera*.

To relate the timing of TE activity to speciation events of *Aiptasia* and its sister species, the Nei-Gojobori distances (*dS*) of the *COX3* genes from *Aiptasia* (Accession No. KJ482979 (66)) and the related anemones *Bartholomea annulata, Verrillactis paguri*, and *Sagartia elegans* [Accession Nos. FJ489483, FJ489503 (67), and JF833012 (68), respectively] were assessed. Protein sequences were aligned using ClustalW (69) and the corresponding nucleotide alignments were extracted using PAL2NAL (70). The *dS* values were further calculated using codeml from the PAML package (71).

2.5.6 Reference transcriptome sequencing, assembly, and annotation

To help identify protein-coding genes in the assembled *Aiptasia* genome, a reference transcriptome was assembled using RNA derived from different developmental states in order to maximize the diversity of expressed and assembled transcripts. First, adult strain CC7 anemones were sampled over a 30-d time course after infecting with strain SSB01 *Symbiodinium* cells as follows: day 1, algae were added at ~10^5 cells/ml to wells containing aposymbiotic anemones in autoclaved and sterile-filtered seawater (AFSW); day 2, brine shrimp were added as food in the usual way (see above) without changing the AFSW or adding additional algae; day 3, the AFSW was changed and algae were again added at ~10^5 cells/ml; day 11, the AFSW was changed without

adding additional algae, and the incubation was continued until final sampling. The anemones were held throughout under the routine culture conditions described above (see 'Organisms'). Total RNA was isolated in four biological replicates from aposymbiotic, partially populated (12 days), and fully populated (30 days) anemones using TRIzol (Life Technologies #15596-026) following the manufacturer's instructions; all samples were taken at the mid-point of the 12-h light period to control for possible circadian changes in gene expression.

Second, larvae were obtained as described previously (7) from crosses between a single CC7 anemone (male) and a single H2 anemone (female). Larvae from each of two such crosses were rinsed on 70-µm-mesh filters (BD Falcon) with sterile-filtered artificial seawater (AFSW) and then split into duplicate infection and no-infection treatments (cross 1: 2 x 6,500 larvae; cross 2: 2 x 8,400 larvae). Each set of larvae was transferred to a clean glass bowl in a final volume of ~20 ml of FASW. For the infection treatments, strain SSB01 *Symbiodinium* cells were added twice, at 3 and 4 d post-fertilization, at ~2.5-5 x 10^4 algal cells/ml. Both infected and uninfected larvae were then harvested at 10 d post-fertilization. For RNA extraction, larvae were rinsed with AFSW on a 70-µm-mesh filter (as above), transferred to a 50-ml Falcon tube in a final volume of 7.5 ml, and centrifuged for 3 min at 4,000 rpm at 4°C. The supernatant was removed, and the larvae were resuspended in the remaining small volume of liquid before adding 400-500 µl of TRIzol and storing at -20°C. RNA was then extracted according to the manufacturer's instructions with the inclusion of an additional

cleaning step with chloroform. Total RNA was resuspended in 30 µl of RNAse-free water.

mRNA was isolated from all samples using Dynabeads oligo$(dT)_{25}$ (Ambion #61002). The quantity and quality of RNA were assessed using a Bioanalyzer 2100 (Agilent Technologies, RNA Nano/Pico Chip) both after total RNA extraction and after mRNA purification. Libraries were then prepared using the NEBNext Ultra Directional RNA Library Prep Kit (NEB #E7420) with 180-bp insert sizes and sequenced together on one lane of an Illumina HiSeq2000 sequencer with read lengths of 2 x 101 bp (Table 2.5). The sequence reads were first trimmed by removing sequencing adaptor sequences and filtered for low-quality reads using Trimmomatic (58), and the error correction module of ALLPATH-LG (59) was then used to identify and remove sequencing errors. To remove read pairs that originated from *Symbiodinium* in the libraries prepared from symbiotic anemones and larvae, the sequences were mapped using bowtie2 (72) to a reference transcriptome that was assembled using four different genotypes of *S. minutum* (NCBI Project Accession No. PRJNA274852), including strain rt002, which is closely related to strain SSB01 (see above). Read pairs that aligned using default parameters were removed from the dataset before proceeding with assembly.

The remaining sequences were combined and assembled using the Trinity transcriptome-assembly tool (73) using the strand-specificity information of the paired-end reads. The default settings were used except for the minimum kmer coverage to be included in the de Bruijn graph, which was set to 3. The final

reference transcriptome contained 43,770 genes with 70,499 transcripts of ≥200 bp (both "genes" and "transcripts" as defined by the assembler); the mean transcript length was 1,142 bp (Table 2.6). The genes were annotated successively against the SwissProt and TrEMBL databases (74) using BLASTX (75) with a E-value cut-off of 1e-5. The final annotation contained 16,588 genes annotated by SwissProt and an additional 5,283 genes annotated by TrEMBL (but not by SwissProt). Gene ontology (GO) terms, including the ancestral GO terms, were assigned to the UniProt identifiers of annotated genes through the UniProt-GOA database (76).

2.5.7 Development of gene models

We identified genes by *ab initio* prediction based on selected gene models that were obtained by aligning the reference transcriptome to the genome assembly and further refinement (Fig. 2.2C). First, the genome assembly and all 70,499 transcripts of the reference transcriptome were used for gene-structure annotation using PASA (77). This yielded 39,628 gene-structure models of which 13,887 featured seemingly full-length coding regions (with both plausible start and stop codons). To generate a high-confidence training set for *ab initio* gene prediction, these complete gene models were subjected to several filters: (i) gene models with <3 exons were excluded (to allow confident definition of intron-exon boundaries); (ii) where the nucleotide sequences of gene models overlapped, only the longest one was retained; (iii) where the predicted protein sequences of ≥2 gene models matched by BLASTP with an E-value of ≤1e-10, only the longest

model was retained; (iv) only gene models with unambiguous 5’ and 3’ untranslated regions (UTRs), as assigned by PASA, were retained; and (v) only gene models that lacked repeat regions, as indicated by BLASTN similarity searches between the annotated repeats (see above) and the gene models, were retained. These steps yielded 2,260 gene-structure models, on which Augustus 3.0. (78) was then trained in order to predict genes (i.e., coding sequences plus UTRs) in the genome assembly using the default Augustus training pipeline.

To refine the *ab initio* predictions, we also mapped all 43,770 genes from the reference transcriptome to the genome assembly using BLAT (79) to generate "hints" (i.e., supplementary evidence of gene presence and location). In total, 38,205 genes from the transcriptome mapped to the genome with an identity of ≥97%. Incorporating these hints, gene predictions by Augustus resulted in 27,998 gene models, of which 25,352 contained a predicted complete coding region. Finally, we used PASA to compare the 39,628 transcript-based gene-structure models that it had produced (see above) to the predicted genomic gene set from Augustus. This step added some gene models and refined others, yielding a final set of 29,269 gene models, of which 26,658 appeared to contain the complete coding regions.

To evaluate the success of the gene-model predictions, we mapped all of the mRNA libraries used in the reference transcriptome assembly back to the predicted gene set and evaluated other evidence for the validity of the minority of

gene predictions that were not supported by such mRNA evidence (see Table 2.7).

2.5.8 Annotation of gene models, identification of taxonomically restricted genes, and identification of specific protein families

The final set of predicted proteins was first annotated using the SwissProt, TrEMBL, and NCBI nr databases, and GO terms were assigned using a pipeline similar to that described previously (80). Briefly, BLASTP searches of all genomic protein models were carried out against the SwissProt database (June 2014 release). The GO terms associated with the hits were then obtained from UniProt-GOA (July 2014 release) (76). If the best-scoring hit of the BLASTP search did not yield any GO annotation, further hits (up to 20, or an E-value > 1e-5, whichever was reached first) were considered, and the best-scoring hit with an available GO annotation was used. If there was no hit in SwissProt with an associated GO term(s), the TrEMBL database (June 2014 release) was queried using the same approach (Table 2.7). Proteins that had no matches to either database were searched against the NCBI nr database (E-value cutoff of 1e-5). Finally, the predicted proteins with no hits to any of the three public databases were searched against the *A. digitifera* (http://marinegenomics.oist.jp) and *Stylophora pistillata* (http://reefgenomics.org/blast/) databases, and the proteins that still had no hits were tentatively classified as taxonomically restricted (*Aiptasia*-specific).

In addition, to obtain domain-based annotations of the predicted *Aiptasia* proteins, the entire set of gene models was also annotated using PANTHER v7.0 (81) and Pfam v27 (82). To identify ficolins and ficolin-like proteins, the Pfam annotation was searched for proteins that had fibrinogen and collagen domains, with the former domain C-terminal to the latter. The related proteins in other cnidarians were identified similarly. Similarly, Toll-like and Interleukin-like receptors were identified by searching the Pfam annotation for TIR domains, essentially as described previously for *N. vectensis* and *A. digitifera* (39), and NOD-like receptors were sought by searching for NACHT and NB-ARC domains (42).

2.5.9 Completeness of the gene set and diversity of molecular functions

As one approach to assessing the completeness of the predicted genomic gene set, we searched for the presence of 458 evolutionarily conserved "core eukaryotic genes" (CEGs) from six model organisms (*Homo sapiens, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Schizosaccharomyces pombe*, and *Saccharomyces cerevisiae*) as proposed by CEGMA (the Core Eukaryotic Genes Mapping Approach) (83). Instead of searching our gene models *de novo* by the CEGMA pipeline, we used BLASTP (75) with an E-value cut-off of 1e-5 to search for homologues of the 458 CEGs in the predicted *Aiptasia* proteome, as well as in the proteomes of *N. vectensis* and *A. digitifera* (Fig. 2.3B).

In addition, to analyze the diversity of molecular functions, PANTHER and Pfam domain counts for 23 metazoan species (Table 2.13) were combined in a single table and a PCA was conducted using the *R* prcomp function [Component loadings %] (84).

Finally, to help determine the presence or absence of *Aiptasia* proteins involved in key biosynthetic and signaling processes, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) (85). First, the complete set of *Aiptasia* gene models was annotated using the KEGG database, yielding an initial list of 13,156 genes with KEGG annotations. Because these annotations were not always reliable (some matches were not optimal as judged by reciprocal BLAST), we then proceeded as follows for each pathway of interest. A list was produced of the KEGG ID numbers for each gene in the pathway, and this list was used to search the previously generated list of KEGG annotations. For each assignment, we asked if it were meaningful and (in cases where there was more than one gene assigned to a given KEGG ID number) the most appropriate gene by using BLASTP to ask if the best hit for the *Aiptasia* sequence in SwissProt (see above) was indeed a protein of the expected type (requiring an E-value of ≤1e-5). In cases in which these steps found no *Aiptasia* gene corresponding to a particular pathway component, we asked if the apparent absence were real by using the human (or other appropriate) sequence with that KEGG annotation to search the predicted *Aiptasia* protein set using BLASTP with an E-value cut-off of 1e-5; the best hits meeting that criterion (up to five) were then used to search SwissProt to test for matches with an E-value cut-off of 1e-5. Several additional pathway

components were found in this way. To compare the *Aiptasia* gene set with those for *N. vectensis* and *A. digitifera* for the same pathways, we used the appropriate annotated gene sets from the KEGG database. In cases where a gene appeared to be present in *Aiptasia* but absent in *N. vectensis* and/or *A. digitifera*, we asked if this difference were real by using BLASTP to search the *N. vectensis* (http://genome.jgi-psf.org/Nemve1/Nemve1.home.html) or *A. digitifera* (http://marinegenomics.oist.jp) gene sets with the *Aiptasia* protein. In cases where a gene appeared to be absent in *Aiptasia* but present in *N. vectensis* and/or *A. digitifera*, the reliability of the latter assignments was tested by using BLASTP against SwissProt as described above. For visualization of KEGG pathways, a reference KEGG list for submission to the KEGG Search&Color tool (http://www.genome.jp/kegg/tool/map_pathway2.html) is provided (Dataset S2.12).

2.5.10 Analyses of differential gene expression

The error-corrected paired-end mRNA reads used as input for the reference transcriptome assembly (see above; Table 2.5) were mapped to the genomic gene set using bowtie2 (72) with settings “-a -t -X 500 --no-unal --rdg 6,5 --rfg 6,5 --score-min L,-.6,-.4 --no-discordant --no-mixed --phred33”. The mapping output was piped into SAMtools (86) for generation of a bam file and sorting. The read-count abundances were then calculated using eXpress (87) with settings “--rf-stranded”. Count tables for "fragments per kilobase of exon per million mapped reads" (FPKM) values and unnormalized read counts were generated using a

custom perl script, and expression levels were subsequently calculated in R (84) using the unnormalized read counts and the package edgeR (88)

. Heat maps were generated using FPKM values imported to the Multiple Experiment Viewer (MeV) and clustered using Euclidean distances (89). MDS plots were created in edgeR using "plotMDS" where distances between pairs of RNA samples correspond to the leading $\log_2$-fold-changes [i.e., average (root-mean-square) of the largest absolute $\log_2$-fold-changes] (88).

2.5.11 Comparative genomic analyses

For all comparative genomic analyses, the same genome assemblies were used as described previously (11) (Table 2.13).

2.5.12 Construction of phylogenetic tree

For construction of a phylogenetic tree, gene families for 14 species were built using a previously described orthology method (12). Only gene families with at least one ortholog in each species were considered, which resulted in 952 gene families. Gene families with multiple paralogs in a single species were reduced to one gene with the best cumulative BLASTP score to all out-group species (to retain only the slowest evolving and/or best-predicted gene models). Alignments were done using MUSCLE (90), and poorly aligned regions were trimmed using GBlocks (91). The final alignment had 150,747 amino-acid positions and no gaps. Phylogenetic inference was conducted with MrBayes (92) using the GTR model (93) with four chains and 1,000,000 generations.

2.5.13 Analysis of synteny and of HOX genes

Synteny was analyzed as described previously (11) requiring blocks of ≥3 genes with ≤10 intervening genes. We compared gene-family clustering (see above and Fig. 2.2A) and PANTHER orthology assignments (v7.0 database and annotation pipeline) (81) and found that marginally more syntenic blocks could be detected with PANTHER, which was thus used for all subsequent synteny analyses. All synteny blocks are available as tracks on the genome browser (http://aiptasia.reefgenomics.org/jbrowse) with species information.

Aiptasia HOX-like genes were identified by BLASTP searches of the *N. vectensis HOX* genes (15) against the *Aiptasia* protein set (E-value cutoff 1e-5). The gene identities were confirmed by protein alignments of the homeobox domains using ClustalW (69).

2.5.14 Analyses of gene-function expansions and fast-evolving gene families

To identify families of protein domains that are specifically expanded in the *Aiptasia* lineage, we conducted Fisher's exact test (94) for PANTHER categories for genomic gene sets, comparing in-group counts (*Aiptasia*) to average counts in the out-groups (all other species in the analysis). This test was iterated over all domains, and the *P*-values obtained were corrected with the Bonferroni correction (95) to identify the significantly expanded domain families. To visualize these expansions, counts were normalized by the total domain-family count in

each species, and significantly expanded gene families were plotted using R's heatmap.2 function (from package gplots) (84).

To check for *Aiptasia* gene families that show elevated substitution rates in their amino-acid sequences, we used the gene-family clustering (see above) and required *Aiptasia, N. vectensis, A. digitifera*, *H. magnipapillata,* and human (as an out-group) sequences to be present in each cluster. This allowed assessment of 2,478 gene families. We constructed alignments using MUSCLE (90) and ran FastTree (96) maximum-likelihood phylogeny inferences on them using the default settings and the generalized time-reversible (GTR) model (93). We used a stringent cutoff to identify fast-evolving gene families in which the *Aiptasia* sequences have a longer branch length relative to the root node of all other cnidarian and human sequences.

2.5.15 Analysis of the taxonomically restricted genes

To seek evidence for TRG function and to ask if clusters of TRGs (see below) were expressed coordinately, we examined their expression as described above in the different developmental and symbiotic states used for transcriptome analysis (Table 2.5). To search for distant homologies of the TRGs, we computed Hidden Markov Models (HMMs) using hmmer (http://hmmer.org) based on the *Aiptasia* paralogs identified by the gene-family clustering (see above). In total, 126 TRGs could be assigned to 54 HMMs (*i.e.,* protein motifs). Identification of distant hits to proteins in the UNIREF90 database (97) and to other *Aiptasia* proteins was performed using hmmsearch. Two of the newly identified motifs had

a distant hit to sequences in UNIREF90 (97), and the remaining 52 (from 122 TRGs) were found in a total of 145 other *Aiptasia* proteins. Finally, to ask if the TRGs were nonrandomly organized in the genome, we compared their observed frequency of clustering to a model of randomness generated by using a custom script to shuffle the order of all genes in the genome 100 times.

2.5.16 Analyses of protein phylogenies

ClustalW (69) was used with default settings for amino-acid-sequence alignments of ficolins and ficolin-like proteins and of Toll-like and Interleukin-like receptors. In both cases, the corresponding mouse and human sequences were downloaded from UniProt and included as the out-group. Alignment gaps were removed with trimAl (98), and the best-fitting substitution model was identified using prottest3 (99). Maximum-likelihood phylogenetic trees were constructed using MEGA v6 (100) with 100 bootstrap replicates, and the final trees were visualized with FigTree (v1.4.2, http://tree.bio.ed.ac.uk/software/figtree/). For the phylogenies of the 3-DHQS and *O*-MT protein domains, the 3-DHQS and *O*-MT domains were identified from Pfam annotations (see above), and the corresponding protein sequences were trimmed to the edges of the annotated domains, and aligned using ClustalW (69) with default settings. Neighbor-joining trees were calculated in Geneious R8 (101) using the JC-model and visualized in FigTree (v1.4.2, http://tree.bio.ed.ac.uk/software/figtree/).

For the phylogeny of the Tox-Art-HYD1 domains, we used MAFFT (102) with default settings and included the eight *Aiptasia* and *Symbiodinium* domains

together with numerous bacterial reference sequences (50). Gaps in the domain alignment were removed with trimAl (98), and the best fitting amino-acid-substitution model was computed using prottest3 (99). The maximum-likelihood tree was calculated and visualized as described above for the ficolin-like proteins.

2.5.17 Genome-wide analysis of horizontal gene transfer (HGT)

To search for *Aiptasia* genes that may have been derived by horizontal transfer from non-metazoans, we first constructed 12 protein-sequence databases that did not overlap among themselves. Ten represented five metazoan (human; rodents; non-human, non-rodent mammals; non-mammal vertebrates; and non-cnidarian invertebrates) and five non-metazoan (plants, fungi, bacterial, archaea, and viruses) protein sets. These databases were downloaded from UniProt, retaining their original taxonomic divisions (see ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/README), and filtered for organisms with nominally complete proteomes to reduce the search space (103). In addition, a cnidarian database was constructed by downloading the protein sets available from the public databases (*N. vectensis* from TrEMBL and *H. magnipapillata* from nr) and supplementing with those for *A. digitifera* (http://marinegenomics.oist.jp) and *S. pistillata* (http://reefgenomics.org/blast). Finally, a *Symbiodinium* database contained the predicted protein sets from *S. minutum* (clade B1) (62) and *S. microadriaticum* (clade A1) (http://reefgenomics.org/blast). In total, the 12 filtered

databases covered 5,570 species. We then conducted BLASTP alignments of all predicted *Aiptasia* proteins to the 12 databases and saved the results in XML format for post-processing. Only hits with e-values ≤1e-5 were retained for further analysis to reduce spurious hits.

We then identified HGT candidates as described previously (103). Briefly, for each gene, we calculated the HGT index (h_u) as the bitscore difference between the highest-scoring non-metazoan hit and the highest-scoring metazoan hit. As previous results indicate that the ratio of true non-metazoan-origin hits to true metazoan-origin hits reaches a maximum at h_u = 30 and then plateaus (103), we considered genes with h_u ≥30 as candidate HGT genes. To identify HGT candidates that were present in other cnidarians as well as *Aiptasia*, we repeated the bitscore calculations with the cnidarian database excluded. To determine the distributions of intron numbers in the *Aiptasia*-specific and cnidarian-specific HGT candidates and compare them to the overall distribution for *Aiptasia* genes, exon information was parsed from the genome-feature file by summing up the number of exons for each predicted gene model. For the *Aiptasia*-specific and cnidarian-specific genes, the 95% binomial-proportion confidence intervals (CIs) were calculated using the formula $CI = p \pm 1.96\sqrt{(1/n)p(1-p)}$, where p is the proportion of genes with ≤N exons and n is the total number of *Aiptasia*-specific or cnidarian-specific genes.

To further test the possibility that HGTs have occurred between *Symbiodinium* and *Aiptasia*, we constructed maximum-likelihood phylogenetic trees. BLASTP searches were conducted against both the complete UniProt database and the

cnidarian and *Symbiodinium* databases (see above), and the best 1,000 hits (E-value ≤1e-5) from each database were retained while allowing a maximum of two hits from any one species. Sequences were then aligned using MUSCLE (90) with default parameters, and regions of low alignment quality were trimmed using trimAl (98), with the in-built "gappyout" parameter. Maximum-likelihood trees were then computed with 100 bootstraps using RAxML with the following parameters: -m PROTGAMMAJTT -x 12345 -p 12345 -N 100 -f a -T 3.

Although we attempted to remove all alien scaffolds during genome assembly (see above and Table 2.12), it seemed possible that the Tox-Art-HYD1-domain-containing genes apparently present in *Aiptasia* could actually have been derived from contaminating bacterial sequences. To address this possibility, we examined the UniProt and nr annotations of the five genes lying upstream and the five genes lying downstream of each Tox-Art-HYD1-domain-containing gene (or gene cluster). As expected for bona fide *Aiptasia* genes, the predicted products of 35 of the 40 upstream and downstream genes have their closest homologues in bilaterians (n = 25), *N. vectensis* (n = 6), a sponge (n = 1), a cellular slime mold (n = 2), or a plant (n = 1). Of the remaining five genes, four had no annotation, and just one had its best match to a bacterial (*Clostridium*) sequence.

2.6 Tables and Figures

Table 2.1: Overview of the *Aiptasia* genome assembly and comparison to other published cnidarian genomes

Parameter	*Aiptasia* [a]	*N. vectensis* [b]	*A. digitifera* [c]	*H. magnipapillata*
Genome size (Mb)	260	329 [e]	420	1,300
Assembly size (Mb)	258	356	419	852
Total contig size (Mb)	213	297	365	785
Total contig size as % of assembly size	82.5	83.4	87.0	92.2
Contig N50 (kb)	14.9	19.8	10.9	9.7
Scaffold N50 (kb)	440	472	191	92.5
Number of gene models	29,269	27,273	23,668	31,452
Number of complete gene models [f]	26,658	13,343	16,434	NA [g]
Mean exon length (bp) [h]	354	208	230	NA [g]
Mean intron length (bp) [h]	638	800	952	NA [g]
Mean protein length (number of amino acids) [h]	517	331	424	NA [g]
Per cent repetitive DNA [i]	~26	26	13	57

[a] This study.

[b] Putnam et al. (8)

[c] Shinzato et al. (9)

[d] Chapman et al. (14)

[e] Value obtained in this study using a fluorescence-activated cell sorter and comparison to the values obtained for *Aiptasia* and chicken red blood cells (see Fig. 2.3D-F). The previously reported genome size was 450 Mb (8)

[f] Gene models with both predicted start and stop codons.

[g] Data not provided in the original publication (14).

[h] Statistics were calculated as in the Assemblathon2 study (63).

[i] As a per cent of the total contig size; see Materials and Methods, Table 2.8, and the references cited for details.

Table 2.2: Overview of genomic libraries and of the sequences obtained [a]

Library type [b]	Insert size (bp)	Library size (Gb)	Sequencer used	Read length (bp)	Read pairs (X 10^{-6})	Cover-age	SRA accession number
Paired-end	180	25	HiSeq 2000	101	123.5	96X	SRX757328
Paired-end	550	11	MiSeq	300	18.7	43X	SRX757521
Mate-pair	2,000	9	GAIIX	90	48.0	33X	SRX757522
Mate-pair	5,000	6	GAIIX	90	33.1	23X	SRX757523
Mate-pair	10,000	3	MiSeq	101	12.8	10X	SRX757529
Mate-pair	12,000	3	MiSeq	101	12.4	10X	SRX757530

[a] All sequences have been deposited in the NCBI Sequence Read Archive (SRA) under the accession numbers indicated.

[b] All DNA was obtained from aposymbiotic animals of clonal anemone strain CC7 (see Materials and Methods).

Table 2.3. Statistics for the *Aiptasia* genome assembly and gene annotation [a]

Parameter	Value
Genome size (by measurement of DNA content [b])	260 Mb
Total size of genome assembly (in 5,065 scaffolds)	258 Mb
Total contig size (in 29,410 contigs)	213 Mb
Scaffold N50	440 kb
Longest scaffold	2,344 kb
Contig N50	14.9 kb
Longest Contig	228 kb
Number of gene models	29,269
Number of apparently complete gene models [c]	26,658
Number of predicted proteins with recognizable (E-value ≤1e-5) homologues in other eukaryotes [d]	26,162
Mean predicted protein length	517 amino acids
Repeat content	26%

[a] See also Tables 2.1, 2.4, and 2.7. Mb, megabase pairs; kb, kilobase pairs; N50, 50% of the assembly is in scaffolds or contigs longer than this value.

[b] See Fig. 2.3A.

[c] Gene models with both predicted start and stop codons.

[d] See Table 2.7 for details.

Table 2.4: Assembly statistics for the *Aiptasia* genome assembly

Overview statistics		
	Genome size	260,000,000 bp
	Total size of scaffolds	258,189,365 bp
	Number of scaffolds	5,065
	N50 scaffold length	440,280 bp
	Total size of contigs	212,903,857 bp
	Number of contigs	29,410
	N50 contig length	14,850 bp
	Mean number of contigs per scaffold	5.8
	Average length of gaps (for gaps of >25 Ns)	1,860
Scaffold statistics		
	Longest scaffold	2,343,502 bp
	Shortest scaffold	901 bp
	Number of scaffolds >500 bp	5,065
	Number of scaffolds >1 kb	5,018
	Number of scaffolds >10 kb	1,333
	Number of scaffolds >100 kb	589
	Number of scaffolds >1 Mb	27
	Mean scaffold size	50,975 bp
	Median scaffold size	2,593 bp
Contig statistics		
	Longest contig	228,246 bp
	Number of contigs >500 bp	28,922
	Number of contigs >1 kb	27,452
	Number of contigs >10 kb	6,263
	Number of contigs >100 kb	24
	Number of contigs >1 Mb	0
	Mean contig size	7,239 bp
	Median contig size	3,370 bp

Table 2.5: Overview of cDNA libraries and sequencing type used to generate the reference transcriptome for gene modeling

Sample [a] (no. of biological replicates)	Insert size	Sequencer used	Read length	Read pairs (X 10^{-6})	SRA accession number
Aposymbiotic larvae (2)	180	HiSeq 2000	2 x 101	22.1	SRX757531
Symbiotic larvae (2)	180	HiSeq 2000	2 x 101	16.8	SRX757532
Aposymbiotic adults (4)	180	HiSeq 2000	2 x 101	46.0	SRX757525
Symbiotic adults - partially populated (4)	180	HiSeq 2000	2 x 101	51.4	SRX757526
Symbiotic adults (4)	180	HiSeq 2000	2 x 101	50.2	SRX757528

[a] See Materials and Methods for further description of samples.

Table 2.6: Assembly statistics for the *Aiptasia* reference transcriptome

Minimum *k*-mer coverage	3
Number of read pairs used for the assembly (i.e. after preprocessing)	143,660,138
Number of genes	43,770
Number of transcripts	70,499
Total length of assembled sequence	80,542,396 bp
Maximum transcript length	37,202 bp
Mean transcript length	1,142 bp
N50 transcript length	2,024 bp
Number of annotated genes (SwissProt)	16,588
Number of annotated genes (TrEMBL but not SwissProt)	5,283
Number of annotated genes (total)	21,871

[a] See Materials and Methods for further description of the analysis.

Table 2.7: Annotation statistics for the *Aiptasia* gene models

Total number of gene models	29,269	100%
Number of apparently complete gene models [a]	26,658	91%
Number of gene models supported by mRNA evidence [b]	27,469	94%
Number of gene models without support from mRNA evidence but with other evidence that they are bona fide genes [c]	1,665	5.7%
Number of predicted proteins with associated GO terms based on SwissProt annotations	18,876 [d]	65%
Number of predicted proteins without SwissProt annotations but with associated GO terms based on TrEMBL annotations	4,117 [d]	14%
Number of predicted proteins without SwissProt or TrEMBL annotations but annotated against nr database	2,052	7%
Number of predicted proteins with no annotations against SwissProt, TrEMBL, or nr	4,224	14%
Number of predicted proteins with no annotations against SwissProt, TrEMBL, or nr that have hits (E-value ≤1e-5) in the *A. digitifera* or *S. pistillata* predicted protein sets [e]	1,117	3.8%
Number of potentially species-specific genes [f]	3,107	11%
Number of potentially species-specific genes found in the reference transcriptome [g]	1,946	6.6%
Number of potentially species-specific genes with ≥1 domain annotations by PANTHER or Pfam	113 [h]	0.004%
Total number of predicted proteins with PANTHER annotations	21,047	72%
Total number of predicted proteins with Pfam domains	18,716	64%
Total number of predicted proteins with KEGG annotations	13,156	45%

[a] Gene models with both predicted start and stop codons.

[b] Represented by ≥1 read pair in the reference transcriptome (see Materials and Methods).

[c] 1,404 had a hit (E-value ≤1e-5) in SwissProt, TrEMBL, or nr. Of the gene models without such a hit, 257 were seemingly complete (with both predicted start and stop codons), and 20 of these also had a hit in PANTHER or Pfam.

Finally, another four gene models were not classified as complete but did have a hit in PANTHER or Pfam. The identification of seemingly bona fide genes without mRNA evidence indicates that annotation using both transcript-based and *ab initio* gene prediction yielded a more comprehensive genomic gene set than would have been obtained with either method alone.

[d] Of these 22,993 hits, 20,437 had E-values <1e-9, and 16,648 had E-values <1e-19, indicating that the majority of the annotations were based on high-confidence alignments to the SwissProt or TrEMBL database.

[e] The predicted protein sets for *A. digitifera* and *S. pistillata* are available at http://marinegenomics.oist.jp and http://reefgenomics.org/blast, respectively.

[f] That is, gene models whose predicted products have no hits (E-value ≤1e-5) in SwissProt, TrEMBL, nr, or the *A. digitifera* or *S. pistillata* predicted protein sets.

[g] Identified by BLASTN alignment of the reference transcriptome to the potentially species-specific genes (E-value ≤1e-100).

[h] Of these, 68 were also among the 1,946 found in the reference transcriptome.

Table 2.8: *Aiptasia* repeat content

Element [a]	Number of occurrences [b]	Number of bp covered [b]
DNA Transposon	**18,478**	**2,937,550**
Polinton	1,968	506,870
EnSpm	4,359	488,942
Helitron	1,245	294,919
Mariner	2,060	238,758
hAT	2,064	234,782
Merlin	1,259	201,578
Sola	789	188,394
Harbinger	1,806	186,924
PiggyBac	656	169,934
ISL2EU	523	125,458
Kolobok	399	93,247
IS4	277	74,196
Crypton	328	70,067
MuDR	448	27,782
P	99	19,884
Chapaev	120	11,038
Transib	34	2,238
Tc1	12	800
Rehavkus	16	737
Zator	6	376
Charlie	7	283
Novosib	2	185
Tc2	1	158
Tc3	1	66
LTR Retrotransposon	**11,751**	**2,409,853**
Gypsy	5,417	1,225,797
BEL	2,108	515,343

DIRS	2,002	450,280
Copia	2,162	196,873
Troyka	30	19,883
ROO	32	1,677
Non-LTR Retrotransposon	**15,549**	**2,707,245**
CR1	3,387	755,083
Penelope	5,145	688,295
RTE	1,420	307,343
L1	984	172,956
Crack	510	106,681
L2	655	94,609
Rex	454	89,327
CRE	262	85,140
RTEX	446	79,067
Daphne	321	78,030
SINE	669	74,191
Tx1	276	36,240
Perere	183	30,515
Nematis	109	25,202
R4	85	25,185
Poseidon	223	19,339
Jockey	200	11,498
Neptune	15	6,442
I	56	5,202
Hero	17	5,044
R2	27	4,890
R1	28	1,975
Tad1	27	1,428
Ingi	15	1,008
Loa	15	990
7SL	2	598
Proto	11	596

Outcast	4	225
Nimbus	1	65
Rtetp	1	42
RandI	1	39
Endogenous Retrovirus	**103**	**5,491**
Low complexity	**34,001**	**5,847,241**
Simple repeat	29,362	4,949,952
Other low complexity	4,639	897,289
Satellite	**390**	**28,153**
Other [c]		**5,817,614**
Not previously described		**36,110,606** [d]
Total [a]		**55,863,753**

[a] Categories of elements are indicated in bold face. The repeated genes for the rRNAs, tRNAs, and snRNAs (probably another 2-3Mb of repeated sequences) are not included in this tabulation.

[b] Bold face indicates subtotals for the respective categories.

[c] Previously described but poorly defined repetitive elements.

[d] The most abundant individual repeat motif is an almost perfectly palindromic 220-bp sequence that covers a total of 607 kb. BLASTN search of the NCBI 'nt' database produced a hit (E-value 1e-30) to a neurotoxin-precursor gene (Accession No. ABW97360) from the anthozoan *Actinia equina* (104). The 220-bp sequence aligns over its entire length to the single intron separating the two exons of this gene. No such repeat is found in the genome of *N. vectensis*, suggesting a recent expansion of this repeat within the Enthemonae (66).

Table 2.9: Presence or absence of genes for amino-acid-biosynthetic enzymes in *Aiptasia, N. vectensis,* and *A. digitifera*

Amino acid	Enzyme	Query sequence	*Aiptasia* [a]	*N. vectensis* [a]	*A. digitifera* [a]
Cys	Cystathionine β-synthase	P32582	AIPGENE 25097	Nemve1|204029	N
Cys	Cystathionine γ-lyase	P31373	AIPGENE 510	Nemve1|184363	adi_v1.09810
Met/Cys/Thr/Ile/Lys	Bifunctional aspartokinase/homoserine dehydrogenase	Q9SA18	N	Nemve1|224979 [b]	N
Met/Cys/Thr/Ile/Lys	Aspartokinase	P10869	N	Nemve1|224979 [b]	N
Met/Cys/Thr/Ile	Homoserine dehydrogenase	P31116	N	Nemve1|225948 [b]	N
Met/Cys	Homoserine *O*-acetyltransferase	P08465	AIPGENE 5556	Nemve1|224940	adi_v1.06190
Met/Cys	Cystathionine γ-synthase [c]	P47164	AIPGENE 11428 AIPGENE 11377	Nemve1|242678	N
Met	Cystathionine β-lyase [c]	P43623	AIPGENE 11428 AIPGENE 11377	Nemve1|242678	adi_v1.17863
Arg	Bifunctional acetylglutamate kinase	Q01217	N	N	N
Arg	Acetylornithine transaminase	P18544	N	N	N
Arg	Ornithine acetyltransferase	Q04728	N	N	N
Phe/Trp/Tyr	Pentafunctional AROM polypeptide	P08566	N	N	
Phe/Trp/Tyr	Chorismate synthase	P28777	N	Nemve1|223774 [b]	N
Trp	Tryptophan synthase	Q42529	N	Nemve1|152188 [b]	N
His	ATP Phosphoribosyl-transferase	P00498	N	N	N
His	Histidine biosynthesis trifunctional protein	P00815	N	N	N

His	Histidinol dehydrogenase	Q9C5U8	N	N	N
Val/Leu/ Ile	Ketol-acid reductoisomerase	P06168	N	N	N
Ile/Leu	3-isopropylmalate dehydrogenase	P04173	N	Nemve1| 157396 [b]	N
Lys	Homocitrate synthase	Q12122	N	N	N
Lys	Homoaconitate hydratase	P49367	N	N	N

[a] Systematic names as used in the corresponding gene-model databases.

[b] These gene models are likely to be derived from contaminating bacterial DNA, as they are present in the assembly as single-exon genes at the edges of scaffolds (*i.e.,* not surrounded by sequences of cnidarian origin), and the predicted proteins show high sequence similarity to bacterial proteins upon reciprocal BLAST to SwissProt (E-values 1e-13, 7e-12, 0, 0, 3e-67, 2e-44).

[c] The genes for these two enzymes appear difficult to distinguish on the basis of this type of sequence analysis alone. Thus, both of the query sequences (derived from *Saccharomyces cerevisiae*) aligned with approximately equivalent E-values to the same two distinct *Aiptasia* gene models and to the same single *N. vectensis* gene model, as shown. These genes could apparently encode either cystathionine γ-synthases or cystathionine β-lyases, and further studies will be required to distinguish these possibilities. Thus, the puzzle presented by our previous transcriptome study (19), in which only AIPGENE11428 was found and identified as a cystathionine γ-synthase – leaving *Aiptasia* apparently without a cystathionine β-lyases to encode the presumed next step in a biosynthetic pathway for methionine – may have been apparent rather than real.

Table 2.10: Differential expression of *CniFL* genes in aposymbiotic and symbiotic anemones

Gene [a]	Systematic gene name [b]	Log_2 fold-change [c]
CniFL1	AIPGENE15899	4.3
CniFL2	AIPGENE18899	6.8
CniFL3	AIPGENE4936	2.4
CniFL4	AIPGENE11648	0.0
CniFL5	AIPGENE13073	4.6
CniFL6	AIPGENE13580	1.0
CniFL7	AIPGENE19581	3.5
CniFL8	AIPGENE7322	-2.5
CniFL9	AIPGENE6313	4.2
CniFL10	AIPGENE15890	1.7
CniFL11	AIPGENE15897	6.2
CniFL12	AIPGENE27171	4.5
CniFL13	AIPGENE27771	3.0

[a] See Fig. 2.10.

[b] Gene names as they can be accessed at the genome browser (http://aiptasia.reefgenomics.org/jbrowse).

[c] For aposymbiotic relative to symbiotic anemones.

Table 2.11: Anthozoan proteins containing TIR-domains

Protein type	*Aiptasia* [a]	*N. vectensis* [a]	*A. digitifera* [a]
TLR	0	1	4
ILR	4	4	7
TIR only [b]	2	2	13

[a] Shown are the numbers of genes encoding proteins of each type that were found in the currently available genomic gene sets by searching the Pfam annotations for TIR domains (see Materials and Methods and Poole & Weis (39) for details).

[b] Proteins that have a TIR domain but lack extracellular leucine-rich-repeats (like TLRs) or immunoglobulin domains (like ILRs).

Table 2.12: Scaffolds from alien sequences that were removed from the final assembly

Query	% Identity	% Coverage	E-value
Scaffold1789|size5486	81.1	79.4	0 [a]
Scaffold2011|size4422	78.7	86.1	0 [a]
Scaffold2772|size2202	86.3	99.8	0 [a]
Scaffold4460|size1158	79.3	99.1	0 [a]
Scaffold5003|size1008	91.1	100	0 [a]

[a] Hit was to the draft-bacterial-genomes database ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/).

Table 2.13: Genome assemblies used for comparative analyses

Organismal group and abbreviation	Species	Source of gene models	Used in analyses
Vertebrates			
Hsa	*Homo sapiens* (human)	NCBI 36 from Ensembl 41	a, b, c, d
Mmu	*Mus musculus* (mouse)	NCBI m36 from Ensembl 41	c, d
Gga	*Gallus gallus* (chicken)	Ensembl v55 (August 2009) on WASHUC2 assembly	c, d
Xtr	*Xenopus tropicalis* (frog)	Annotation v4.2 on assembly v4.1	c, d
Dre	*Danio rerio* (zebrafish)	Zv 6 from Ensembl 41	a, c, d
Non-vertebrate deuterostomes			
Bfl	*Branchiostoma floridae* (lancelet)	Brafl1 JGI annotation (April 2006)	a, c, d
Cin	*Ciona intestinalis* (ascidian)	JGI finalized models (December 2005), REPLACES proteome 16	c, d
Sko	*Saccoglossus kovalevskii* (acorn worm)	JGI Annotation of *Saccoglossus kovalevskii* v3	d
Spu	*Strongylocentrotus purpuratus* (sea urchin)	NCBI gene build 2 version 1 on Baylor's assembly Spur_v2.1	c, d
Ecdysozoa			
Dme	*Drosophila melanogaster* (fruit fly)	BDGP 4 from Ensembl 41	a, c, d
Tca	*Tribolium castaneum* (flour beetle)	NCBI gene models build 1 v1 based on the assembly Tcas_2.0 (September 2005)	c, d
Pinctada	*Ixodes scapularis* (tick)	v1.1 from ftp://ftp.vectorbase.org/public_data/organism_data/iscapularis/	c, d
Dpu	*Daphnia pulex* (water flea)	FilteredModels8 from JGI (September 2007)	a, c, d
Cel	*Caenorhabditis elegans* (nematode worm)	Wormbase release WS164	a, c, d
Lophotrochozoa			

Cgi	*Crassostrea gigas* (Pacific oyster)	Genome Assembly from Zhang et al. (105)	c, d
Pfu	*Pinctada fucata* (pearl oyster)	Genome Assembly from Takeuchi et al. (106)	c, d
Lgi	*Lottia gigantea* (owl limpet)	Limpet proteome from JGI (Lotgi1@shake, FilteredModels1 table) (May 2007)	a, c, d
Hro	*Helobdella robusta* (leech)	FilteredModels3 from JGI (September 2007)	c, d
Cte	*Capitella teleta* (polychaete annelid worm)	FilteredModels2 track from Capca1 on shake, pre-release of v1.0 (June 2007)	a, c, d
Ava	*Adineta vaga* (rotifer)	Genome Assembly from Flot et al. (107)	c, d
Cnidaria			
Aip	*Aiptasia*	This study (http://aiptasia.reefgenomics.org)	a, b, c, d
Nve	*Nematostella vectensis* (starlet sea anemone)	JGI Annotation of *Nematostella* genome v1	a, b, c, d
Adi	*Acropora digitifera* (stony coral)	Assembly v1.1 from OIST	a, b, c, d
Hma	*Hydra magnipapillata* (freshwater cnidarian)	Chosen models from PASA+Augustus+Hma1 (July 2008)	a, b, c, d
Others			
Sma	*Schistosoma mansoni* (trematode worm)	v080508 from ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/	d
Tad	*Trichoplax adhaerens* (placozoan)	FilteredModels2 from JGI (August 2007)	a,d
Aqu	*Amphimedon queenslandica* (sponge)	Aqu1 models (August 2008)	a,d
Sme	*Schmidtea mediterranea* (amoeba)	Mk4 models from http://smedgd.neuro.utah.edu	d

[a] Construction of phylogenetic tree and analysis of molecular-function enrichment.

[b] Evaluation of fast evolving genes.

[c] Principal-component analysis of molecular function.

[d] Analysis of synteny.

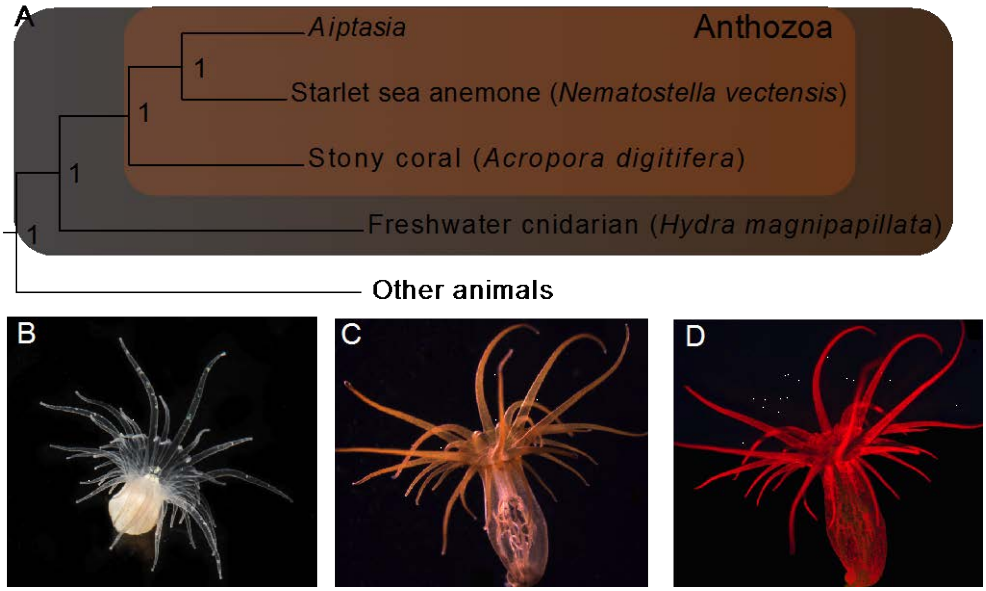

Fig. 2.1. Phylogenetic position and different symbiotic states of *Aiptasia*. (*A*) Partial phylogenetic tree (Materials and Methods and Fig. 2.2A for details) shows *Aiptasia* grouped with other anthozoans among the cnidarians. Numbers on nodes denote bootstrap values. (*B-D*) An aposymbiotic *Aiptasia* polyp (*B*) and symbiotic polyps viewed under white light (*C*) or by fluorescence microscopy to visualize the red chlorophyll autofluorescence of the endosymbiotic *Symbiodinium* algae (*D*).

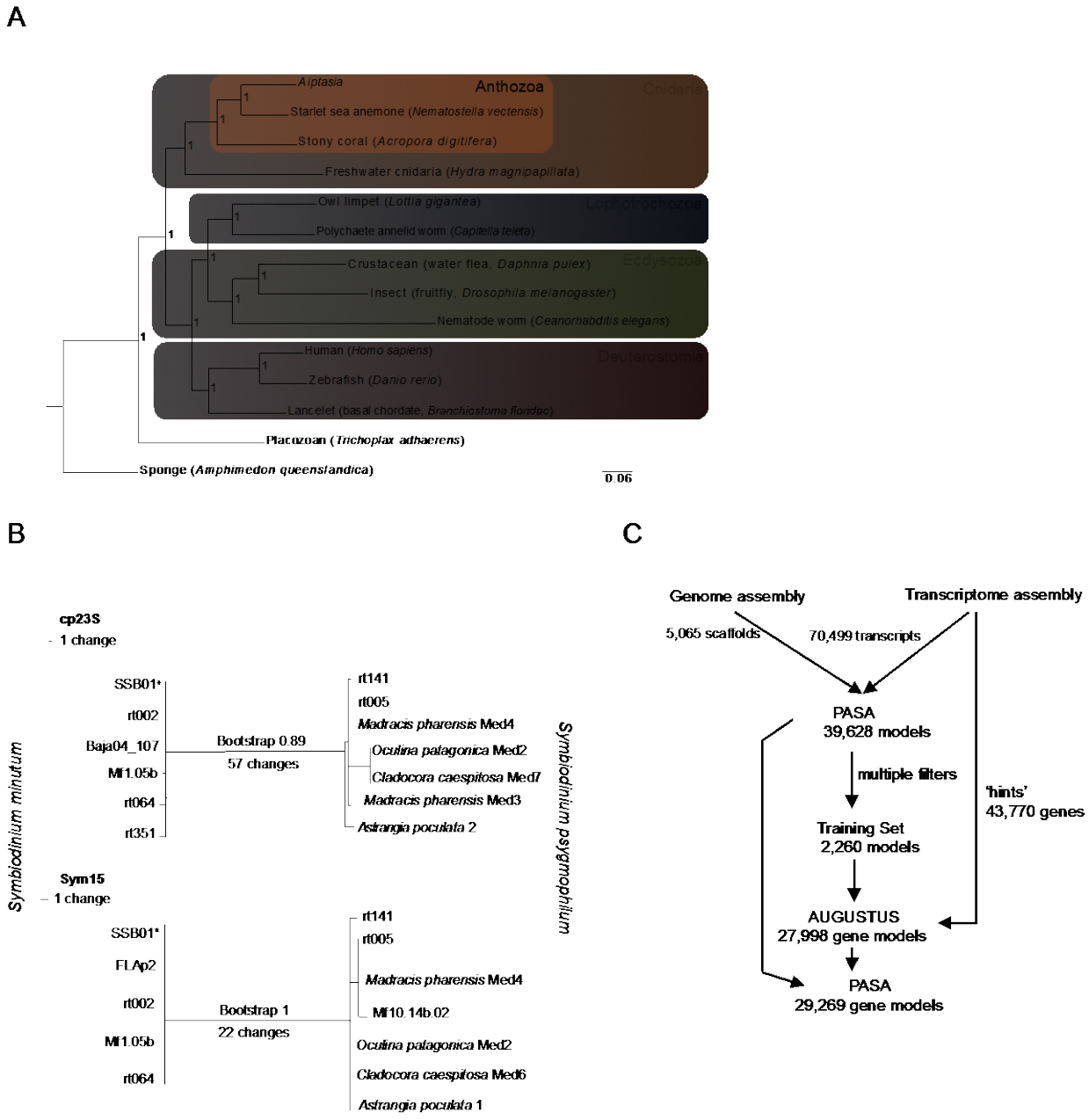

Fig. 2.2. Phylogenetic positions of the organisms used and pipeline for developing gene models. (*A*) Animal phylogenetic tree inferred from 150,747 positions of 952 orthologous protein families (see Materials and Methods) shows *Aiptasia* strain CC7 grouped with other anthozoans. Numbers on nodes denote

bootstrap values based on 1,000,000 generations and four chains; branch length (see scale bar) indicates numbers of amino-acid changes. (*B*) Maximum-parsimony trees of markers *cp23S* and *Sym15* from different strains of *Symbiodinium* show that strain SSB01 (see Materials and Methods) groups with other genotypes of *S. minutum* (left) and not with strains of *S. psygmophilum* (right). Scale bars indicate numbers of nucleotide differences. $^{+}$, Accession No. JX221048 (54); *, this study. Bootstrap values are based on 1,000 iterations. More details of this analysis (for strains other than SSB01) are provided by LaJeunesse *et al.* (56) (*C*) Pipeline for the development of genomic gene models. See Materials and Methods for details.

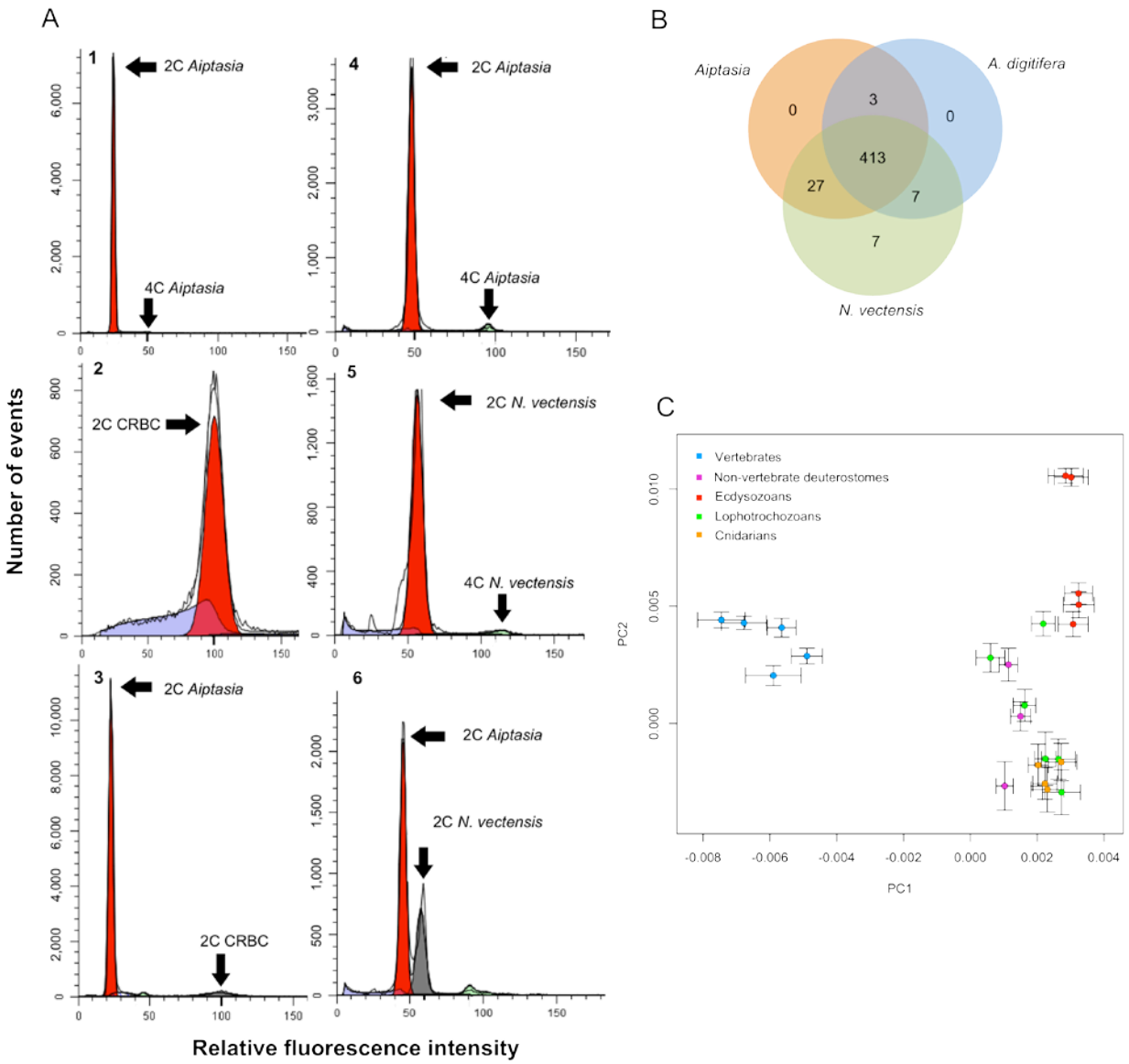

Fig. 2.3. Genome size and diversity of molecular functions. (*A*) Fluorescence-activated-cell-sorter profiles of propidium-iodide-stained nuclei of *Aiptasia,* chicken red blood cells (CRBC), and *Nematostella vectensis* (see SI Materials and Methods). Experiment 1 (panels 1-3) included nuclei from *Aiptasia* only (1), CRBC only (2), and both types of nuclei that were mixed and then stained (3). Experiment 2 (panels 4-6) was similar except that *N. vectensis* nuclei were used instead of CRBCs. Arrows indicate the peaks of interest. The different peak

heights within an experiment (note different ordinate scales) reflect the different numbers of nuclei measured; the different positions of peak channels in the two experiments reflect differences in the staining conditions. Given the CRBC 2C DNA content of 2,279 Mb (see Materials and Methods), the haploid DNA contents of *Aiptasia* and *N. vectensis* are estimated to be ~260 and ~329 Mb, respectively. (*B*) Presence of most of the 458 "core eukaryotic genes" (83) in *Aiptasia* and other anthozoans (see SI Materials and Methods). Venn diagram shows the numbers of such genes present in one, two, or all three of the gene sets. (*C*) Principal component analysis (PCA) of molecular functions (see Materials and Methods). As described previously (11), the first component separates the genomes of vertebrates (light blue) from the invertebrates, including non-vertebrate deuterostomes (purple), ecdysozoans (red), lophotrochozoans (green), and cnidarians (orange).

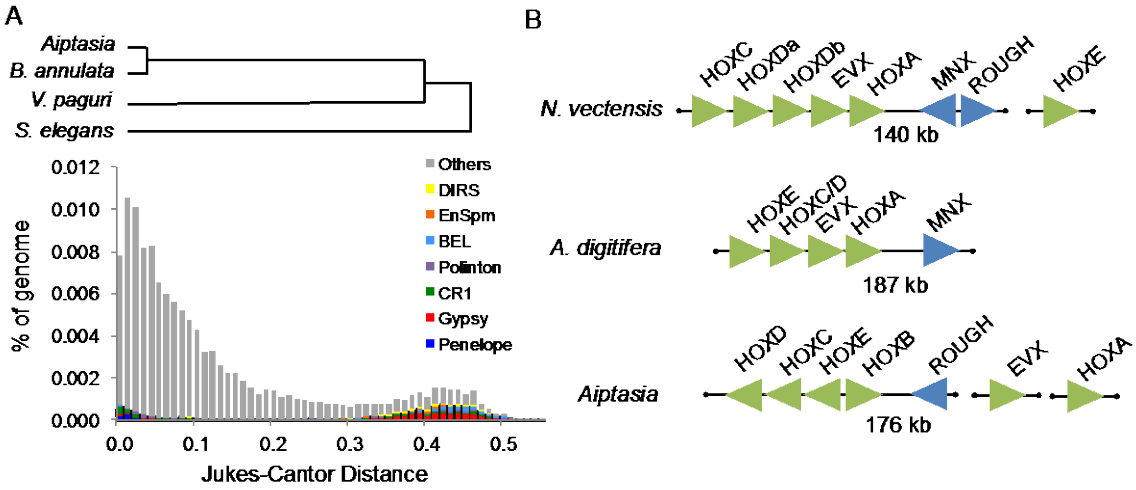

Fig 2.4. Genomic rearrangements in *Aiptasia*. (*A*) Correlation between a period of speciation and a period of high transposable-element (TE) activity. (Top) Approximate times of divergence of *Aiptasia* from the anemone species *Sagartia elegans, Verrilactis paguri,* and *Bartholomea annulata* based on Nei-Gojobori synonymous-substitution rates in the *COX3* gene (see Materials and Methods), which correspond to the Jukes-Cantor distances used to estimate the times of past TE activity. *Aiptasia* and *B. annulata* are symbiotic with dinoflagellates; *V. paguri* and *S. elegans* are not. (Bottom) The dynamics of seven distinct TE classes, plotted as the cumulative percentage of the genome covered by each class (ordinate) at given Jukes-Cantor substitution distances [abscissa; calculated based on the nucleotide differences between the individual genomic TEs and the consensus sequence for the corresponding TE family (see SI Materials and Methods)]. (*B)* Arrangements of *HOX* gene clusters in three anthozoans. *N. vectensis* and *A. digitifera* genes are named as described

previously (15); *Aiptasia* genes are named based on the sequence similarities as shown in Fig. 2.5C. Arrows indicate directions of transcription. Green: Hox and Hox-related genes, Blue: other homeobox genes.

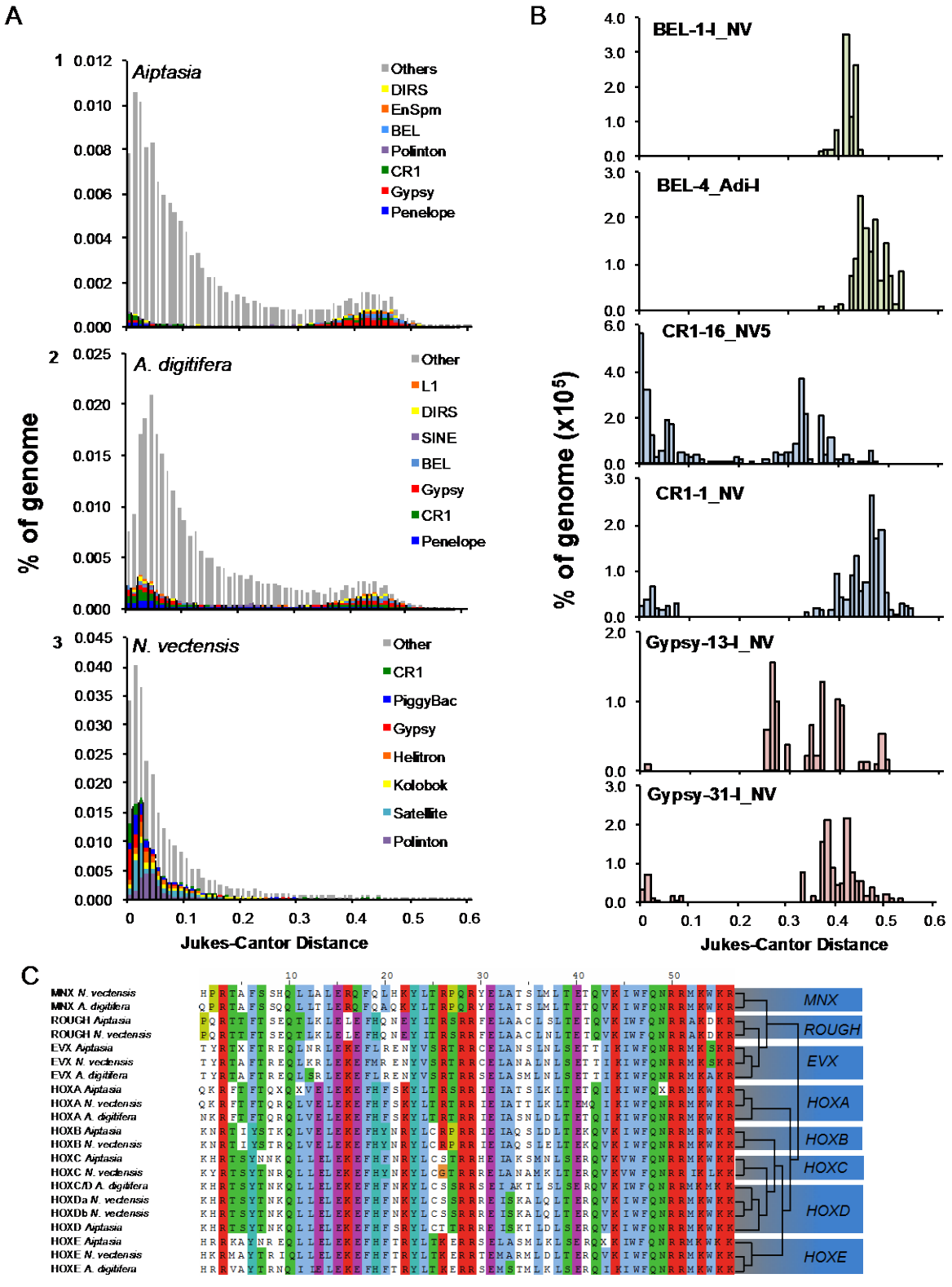
A
1
Aiptasia
Others
DIRS
EnSpm
BEL
Polinton
CR1
Gypsy
Penelope
2
A. digitifera
Other
L1
DIRS
SINE
BEL
Gypsy
CR1
Penelope
3
N. vectensis
Other
CR1
PiggyBac
Gypsy
Helitron
Kolobok
Satellite
Polinton
% of genome
Jukes-Cantor Distance
B
BEL-1-I_NV
BEL-4_Adi-I
CR1-16_NV5
CR1-1_NV
Gypsy-13-I_NV
Gypsy-31-I_NV
% of genome (x10^5)
Jukes-Cantor Distance
C
MNX N. vectensis
MNX A. digitifera
ROUGH Aiptasia
ROUGH N. vectensis
EVX Aiptasia
EVX N. vectensis
EVX A. digitifera
HOXA Aiptasia
HOXA N. vectensis
HOXA A. digitifera
HOXB Aiptasia
HOXB N. vectensis
HOXC Aiptasia
HOXC N. vectensis
HOXC/D A. digitifera
HOXDa N. vectensis
HOXDb N. vectensis
HOXD Aiptasia
HOXE Aiptasia
HOXE N. vectensis
HOXE A. digitifera
MNX
ROUGH
EVX
HOXA
HOXB
HOXC
HOXD
HOXE

Fig. 2.5. Genomic rearrangements and synteny. (*A*) Diversification of transposable elements (TEs) in *Aiptasia* (1), *A. digitifera* (2), and *N. vectensis* (3). For each organism, periods of TE diversification were identified by plotting the cumulative percentage of the genome covered by each of the seven most abundant TE classes in that organism (ordinate) against the Jukes-Cantor substitution distances [abscissa; calculated based on the nucleotide differences between the individual genomic TEs and the consensus sequence for the corresponding TE family (see SI Materials and Methods)]. (*B*) Plots as in *A* for the two most abundant families in the BEL (top), CR1 (middle), and Gypsy (bottom) TE classes in *Aiptasia*; the Gypsy and BEL families show predominantly an ancient divergence, whereas the CR1 families appear to have undergone at least two periods of divergence. (*C*) Alignment of the homeobox domains of HOX-like proteins of *Aiptasia* (this study)*, N. vectensis* (15), and *A. digitifera* (16) (see Materials and Methods, Dataset S2.2)*.*

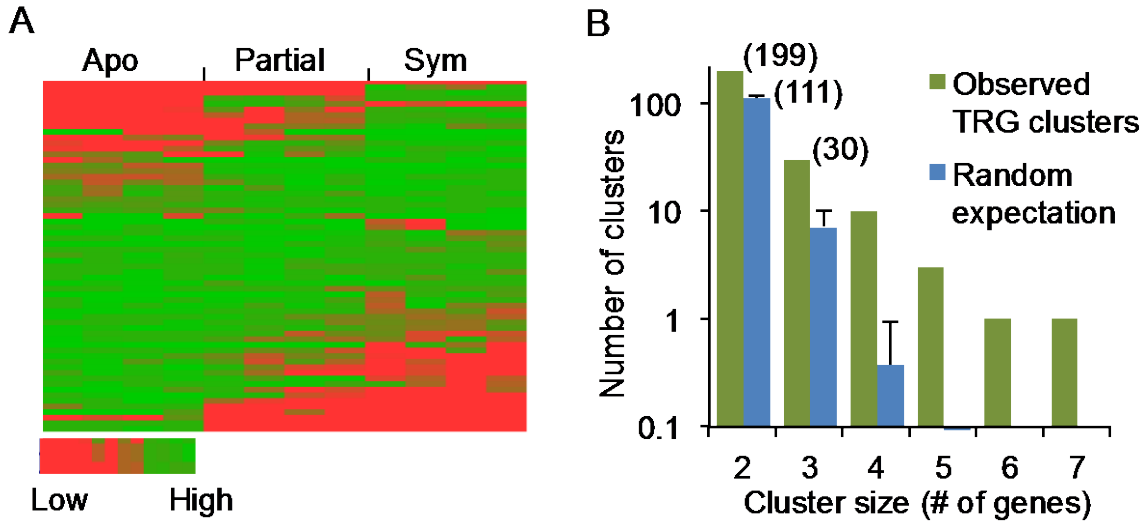

Fig. 2.6. Differential expression and non-random chromosomal clustering of taxonomically restricted genes (TRGs) in *Aiptasia*. (*A*) Heatmap of FPKM expression values (see SI Materials and Methods) for 63 putative TRGs with expression-level changes of ≥8-fold (up or down) between partially (Partial) or fully (Sym) infected anemones and anemones without dinoflagellates (Apo). (*B*) To evaluate clustering, the observed numbers of clusters of two or more putative TRGs without intervening genes were compared to the expectations based on a random distribution of such genes in the genome (see SI Materials and Methods). Error bars indicate standard deviations; numbers in parentheses indicate the actual numbers of clusters of those sizes.

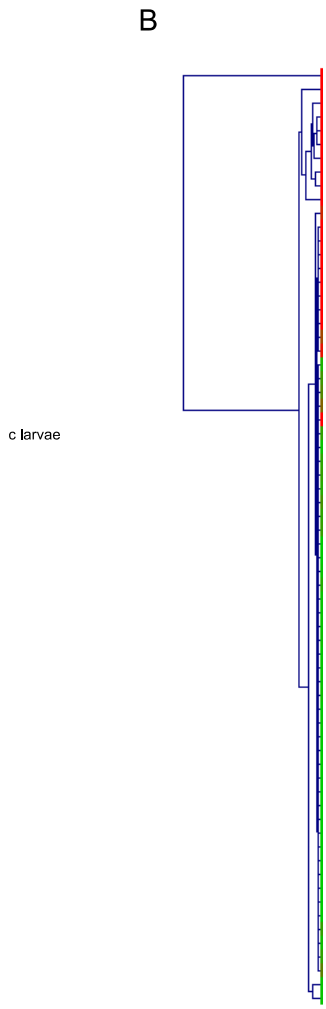

Fig. 2.7. Differential expression and lack of operon-like behavior of the apparently taxonomically restricted genes (TRGs). (*A*) MDS plot (see Materials and Methods) of expression levels of 1,946 putative TRGs among symbiotic and developmental states. Replicates (same color) of each developmental and symbiotic state cluster together and separate clearly along both the developmental (horizontal) and status-of-infection (vertical) axes, indicating overall changes in gene expression between the different states. (*B*) Expression levels (FPKM values) in different developmental and symbiotic states of putative TRGs present in clusters with ≥4 genes per cluster (see Materials and Methods).

Each sample (see panel *A*) is represented by one column in the heat map, and genes originating from the same cluster are indicated by the same color in the key on the right.

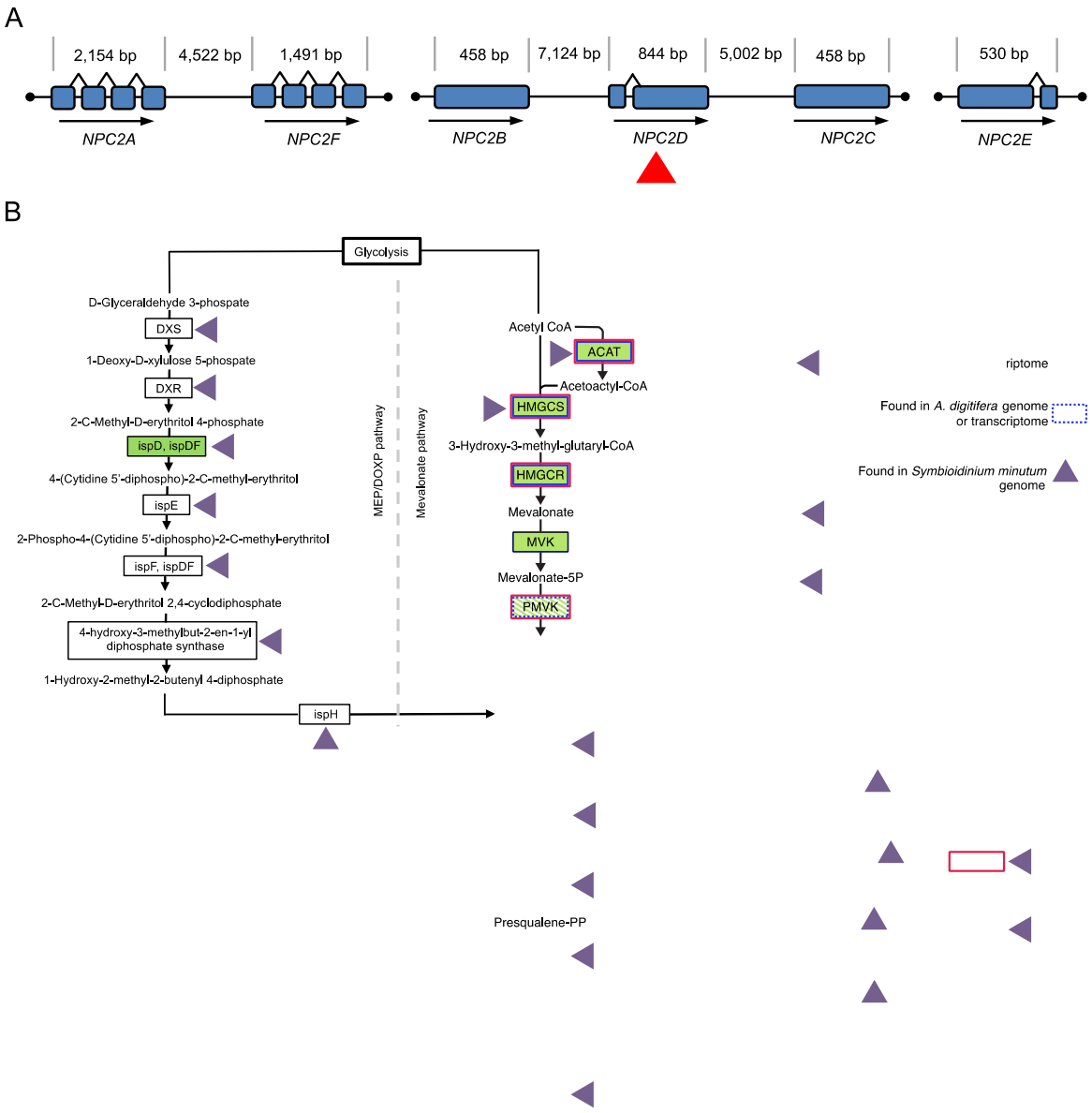

Fig. 2.8. Sterol transport and biosynthesis in cnidarian-dinoflagellate symbiosis. (*A*) The six *Aiptasia* genes encoding Npc2-family sterol-transport proteins are found at three genomic loci that appear to have different source sequences based on their distinct intron-exon structures. Arrows, gene orientations; blue boxes, exons; black carets, introns; numbers, distances between start and stop codons; arrowhead, gene that is upregulated in symbiotic relative to

aposymbiotic anemones (19, 21) and whose product appears to localize to the symbiosome (22). *NPC2A-E* were defined previously (19); *NPC2F* was newly identified from the genome sequence (Dataset S2.5). (*B*) Absence in cnidarians, but presence in *Symbiodinium*, of enzymes essential for the *de novo* synthesis of sterols (KEGG Accession No. ko00100). Results from KEGG-assisted searching (see SI Materials and Methods) of the predicted *Aiptasia, N. vectensis, A. digitifera,* and *S. minutum* protein sets are shown. As expected, cnidarians, like other animals, can apparently synthesize isoprenoids (which have a variety of biochemical functions) by the mevalonate pathway (KEGG Accession No. ko00900). In contrast, *Symbiodinium* can apparently use the MEP/DOXP pathway (known also from plants and apicomplexan parasites) to synthesize isoprenoids (KEGG Accession No. ko00900).

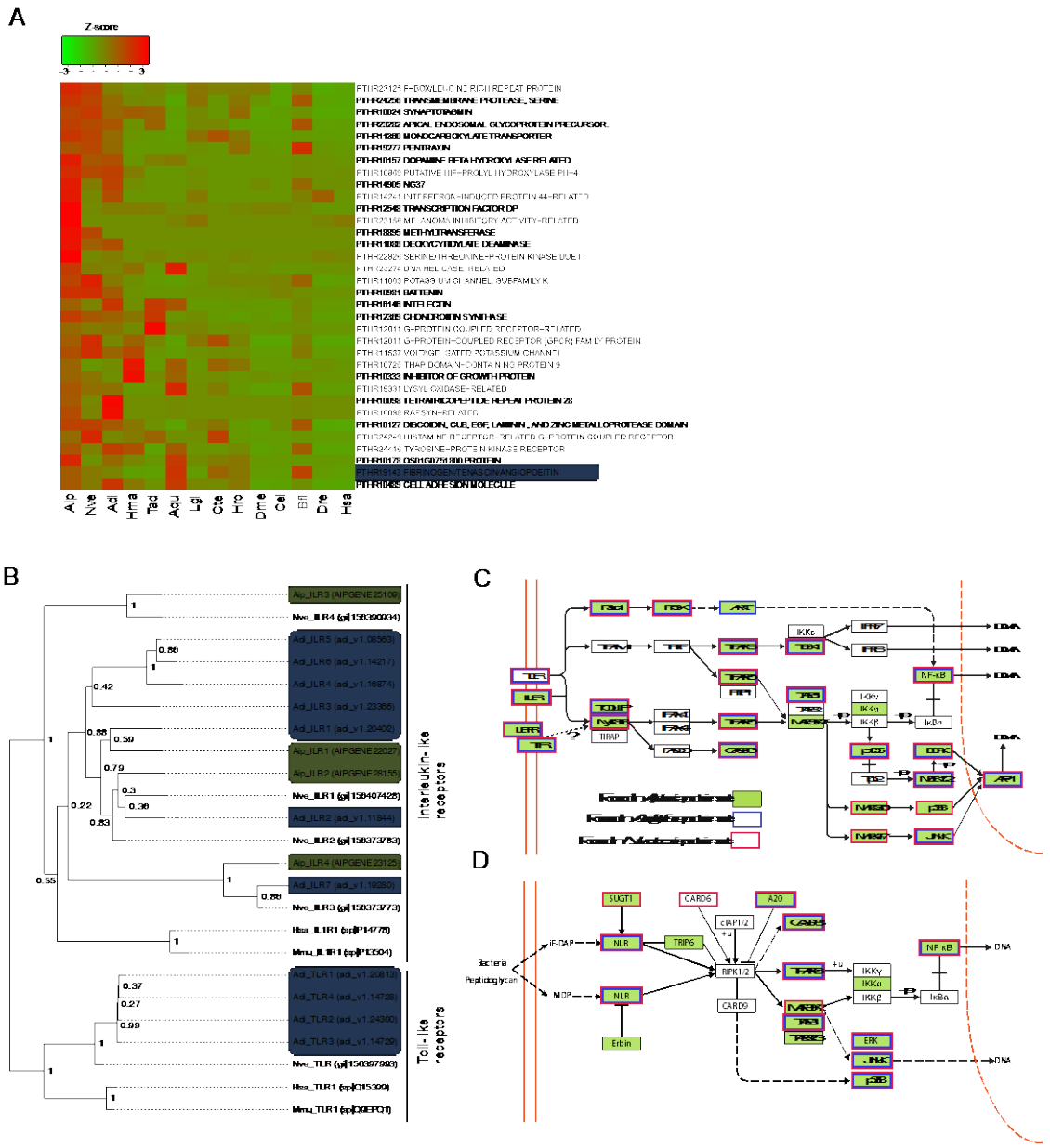

Fig. 2.9. Mechanisms potentially involved in the interaction of *Aiptasia* with *Symbiodinium* and other microbes. (*A*) Functional-domain enrichment in the *Aiptasia* genome relative to 13 other metazoan genomes (Table 2.13) based on PANTHER-domain annotation (see Materials and Methods). The fibrinogen domain is highlighted. (*B*) Maximum-likelihood phylogenetic tree (see SI Materials and Methods) of the Toll-like-receptor (TLR) and Interleukin-like-

receptor (ILR) proteins of three anthozoans (see also (39)). Species abbreviations as in panel A; gene identifiers in parentheses; numbers on nodes, bootstrap values based on 100 iterations; green and blue highlights, *Aiptasia* and *A. digitifera,* respectively. (*C,D*) Components of putative TLR and ILR (*C*) and "nucleotide-binding and oligomerization domain" (NOD)-like (*D*) pattern-recognition and signaling pathways identified in the *Aiptasia* and other anthozoan predicted gene sets by a combination of Pfam annotation and KEGG-pathway analysis (see Materials and Methods). Color code for organisms applies to both panels.

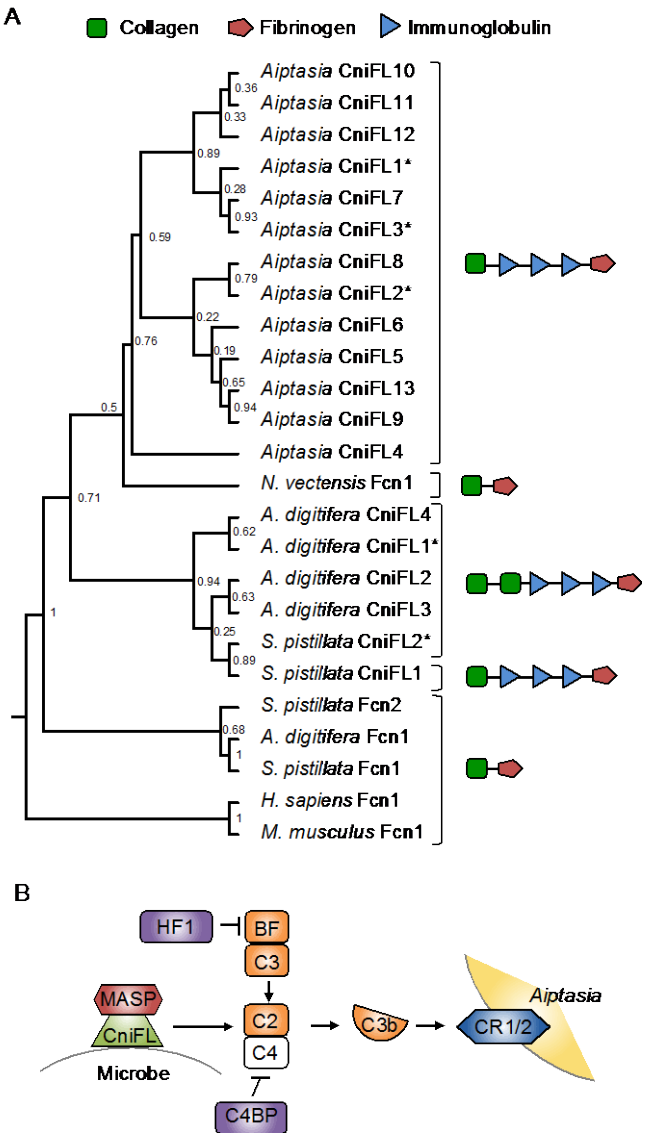

Fig. 2.10. Cnidarian ficolin-like proteins (CniFLs), a newly recognized family of putative pattern-recognition receptors. (*A*) Maximum-likelihood phylogenetic tree of all CniFLs (see text) identified in *Aiptasia* and stony corals (*A. digitifera* and *S. pistillata*) and of all canonical ficolins (Fcns) found in stony corals and the anemone *N. vectensis* (Dataset S2.7). A human and a mouse ficolin (one of three and two, respectively, in those species) were also included as an out-

group. The tree is based on an alignment of a 113-amino-acid sequence (with gaps removed) that spans portions of the collagen and fibrinogen domains. Most of the identified CniFLs contain three central immunoglobulin domains, but five (indicated by *) contain only two. Numbers on nodes denote bootstrap values. (*B*) Components of the lectin-complement pathway that are encoded in the *Aiptasia* genome (based on KEGG analysis – see SI Materials and Methods) and may be involved in signaling downstream of the CniFLs. MASP, mannose-binding-lectin-associated serine protease; C2, C3, and C3b, complement components 2 and 3 and the cleavage product of C3; HF1, complement factor H; BF, complement factor B; C4BP, complement component 4 binding protein; CR, complement receptor. Complement component C4 was not unequivocally identified in the currently available sequence, but is included here for the sake of completeness (Dataset S2.8).

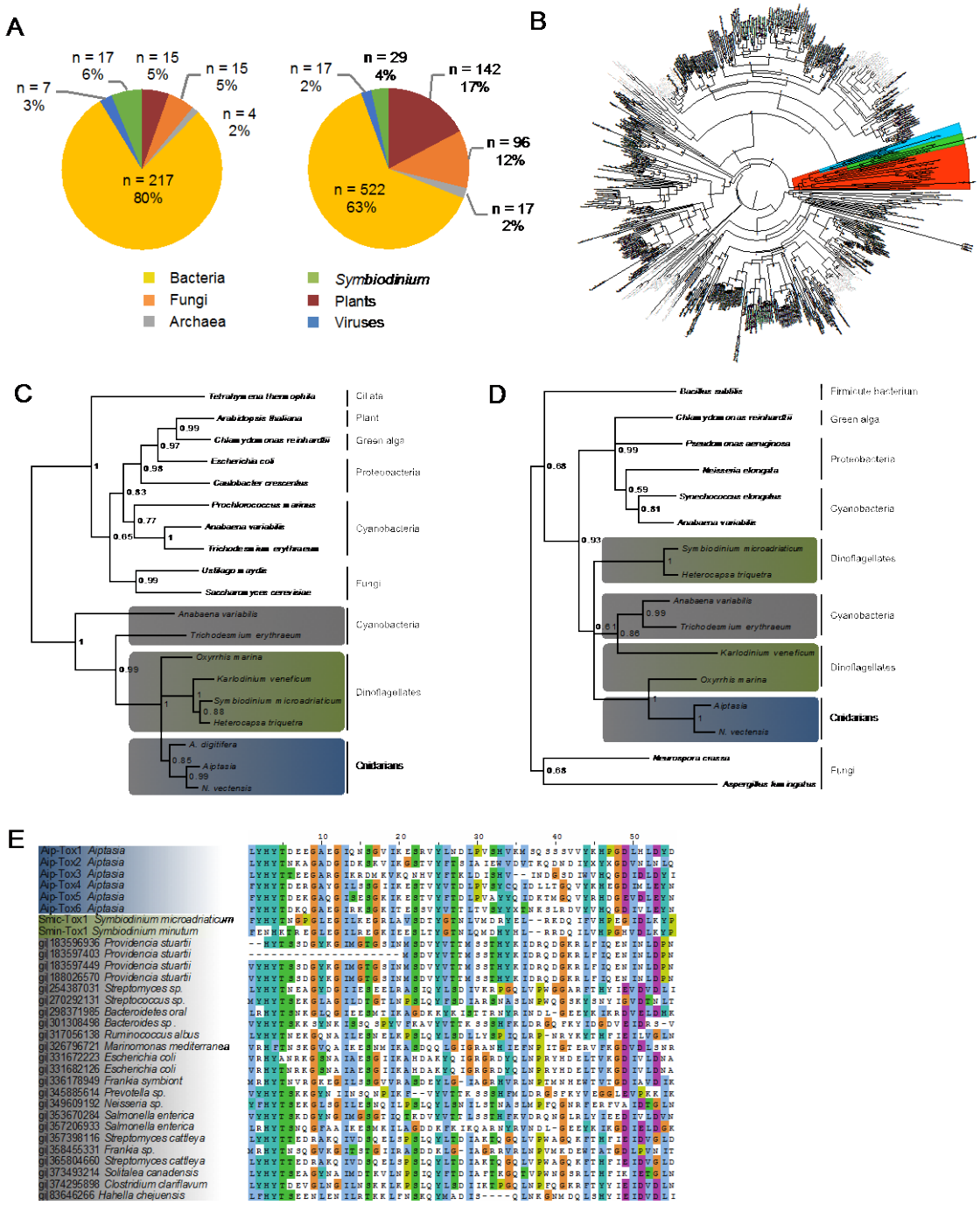

Fig. 2.11. Evidence for horizontal transfer of genes into *Aiptasia* and other cnidarians from prokaryotes and *Symbiodinium*. (*A*) Probable taxa of origin (based on protein-alignment data; see SI Materials and Methods) of HGT

candidates specific to *Aiptasia* (left) and to cnidaria (right). Note that the 823 proteins shown in the right-hand graph include the 275 proteins shown in the left-hand graph. Of the 29 cnidarian proteins whose best alignment is to a *Symbiodinium* protein (right), 17 are *Aiptasia*-specific (left); of the other 12, 10 are present in the *N. vectensis* genome and all 12 in the *A. digitifera* genome [in contrast to a previous report (9)]. (*B*) Maximum-likelihood phylogenetic tree (see SI Materials and Methods) for one of the 12 cnidarian HGT candidates of putative dinoflagellate/*Symbiodinium* origin that is not specific to *Aiptasia* (see panel A). This protein is a predicted Ca^{2+}-activated chloride channel; similar results were obtained with each of the three other proteins examined in this way [i.e. DHQS-*O*-MT fusion protein (AIPGENE19170), 3-hydroxybutyrate dehydrogenase (AIPGENE6203), and heme oxygenase (AIPGENE14419)]. The distinct, shaded clade includes proteins from cnidarians (blue; one each from *Aiptasia, A. digitifera, S. pistillata,* and *H. magnipapillata*), *Symbiodinium* (green; two each from *S. minutum* and *S. microadriaticum*), and a variety of mostly unicellular eukaryotes [red; in order from the top: one from a brown alga, three from three different species of diatoms, one from a parasitic dinoflagellate, one from a foraminiferan, five from five different species of the protozoan genus *Leishmania*, two from two species of flatworms (platyhelminthes), one from the bacterial genus *Clostridium*, and two from the ciliate *Tetrahymena*]. (*C,D*) Phylogenetic trees of 3-DHQS (*C*) and *O*-MT (*D*) domain sequences (see SI Materials and Methods). Despite the report of a 3-DHQS-*O*-MT fusion protein in *A. digitifera* (9), we were unable to identify any such gene in the data available from

www.marinegenomics.oist.jp, so no *A. digitifera O*-MT domain is included in panel (*D*). *Aiptasia, A. digitifera,* and *N. vectensis* do all contain recognizable homologues of the cyanobacterial enzymes "(ATP-grasp" and "NRPS-like") proposed to catalyze the next two steps in the synthesis of the mycosporine amino acid shinorine. Numbers on nodes, bootstrap values based on 1,000 iterations; grey, green, and blue shading, cyanobacteria, dinoflagellates, and cnidarians, respectively. (*E*) Alignment of *Aiptasia* (blue shading), *Symbiodinium* (green shading), and reference bacterial (grey shading) Tox-Art-HYD1 domains after gap removal. *Aiptasia TOX1, 4,* and *6* are the three genes identified initially as HGT candidates; *TOX2, 3,* and *5* were identified by their Pfam annotations (see text). The closely related *TOX4-6* are also closely linked in the genome (see text). Colored shading highlights some of the conserved amino-acid positions.

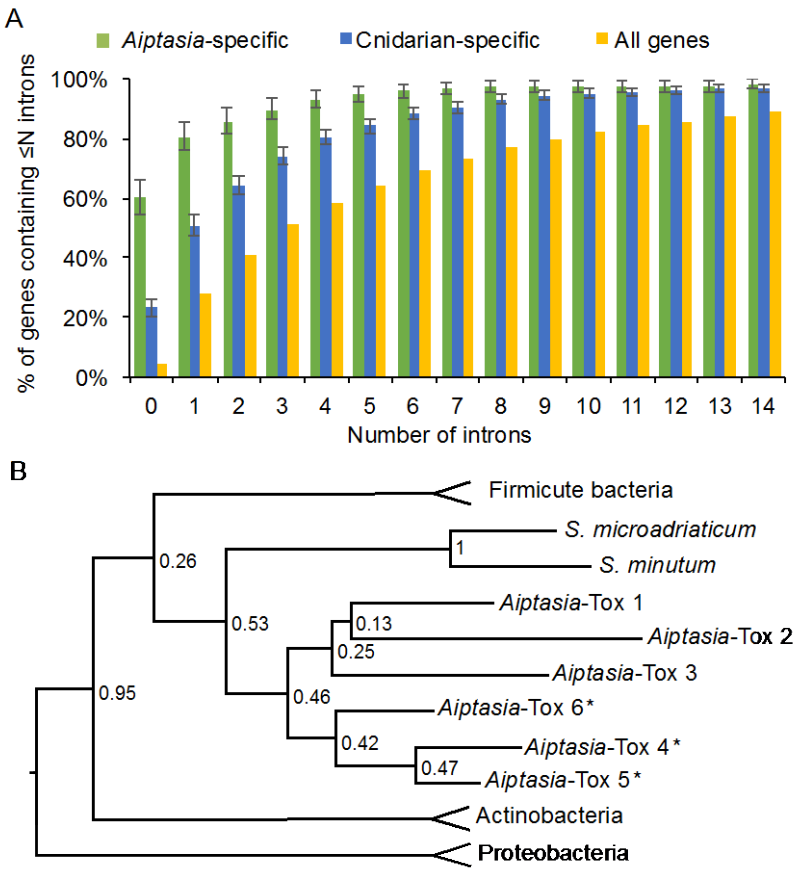

Fig. 2.12. Evidence for horizontal gene transfer in *Aiptasia.* (*A*) Cumulative-distribution plot of intron numbers in *Aiptasia*-specific (green) and cnidarian-specific (blue) HGT candidates (see text) and in all 29,269 *Aiptasia* gene models (yellow). Because of the smaller sample sizes of the first two categories (275 and 823, respectively), 95% binomial-proportion confidence intervals (see SI Materials and Methods) are shown for them. (*B*) Maximum-likelihood phylogenetic tree for the *Aiptasia* Tox-Art-HYD1 domains apparently derived by intraspecific duplications after an HGT event. For simplicity, the numerous and varied bacterial species whose Tox-Art-HYD1 domains form the outer branches of the tree are not shown individually (see Fig. 2.11E). Numbers on nodes denote bootstrap values. *, the three proteins in one genomic region (see text).

2.7 Supplemental Information

Supplemental Information (Datasets S2.1 - S2.12) are archived and accessible at the KAUST library.

2.8 References

1. Dubinsky Z, Stambler N (2010) *Coral reefs: an ecosystem in transition* (Springer Science & Business Media).
2. Davy SK, Allemand D, Weis VM (2012) Cell biology of cnidarian-dinoflagellate symbiosis. *Microbiol Mol Biol Rev* 76(2):229–261.
3. Hoegh-Guldberg O, et al. (2007) Coral reefs under rapid climate change and ocean acidification. *Science (80-)* 318(5857):1737–1742.
4. Weis VM, Davy SK, Hoegh-Guldberg O, Rodriguez-Lanetty M, Pringle JR (2008) Cell biology in model systems as the key to understanding corals. *Trends Ecol Evol* 23(7):369–376.
5. Thornhill DJ, Xiang Y, Pettay DT, Zhong M, Santos SR (2013) Population genetic data of a model symbiotic cnidarian system reveal remarkable symbiotic specificity and vectored introductions across ocean basins. *Mol Ecol* 22(17):4499–4515.
6. Schoenberg DA, Trench RK (1980) Genetic Variation in *Symbiodinium* (=*Gymnodinium*) *microadriaticum* Freudenthal, and Specificity in its Symbiosis with Marine Invertebrates. III. Specificity and Infectivity of *Symbiodinium microadriaticum*. *Proc R Soc London Ser B Biol Sci* 207(1169):445–460.
7. Hambleton EA, Guse A, Pringle JR (2014) Similar specificities of symbiont uptake by adults and larvae in an anemone model system for coral biology. *J Exp Biol* 217(Pt 9):1613–1619.
8. Putnam NH, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science (80-)* 317(5834):86–94.
9. Shinzato C, et al. (2011) Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* 476(7360):320–323.
10. Gregory TC (2015) Animal Genome Size Database. 2015. Available at: http://genomesize.com/result_species.php?id=1309.

11. Simakov O, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493(7433):526–531.
12. Putnam NH, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453(7198):1064–1071.
13. Lynch M, Walsh B (2007) *The origins of genome architecture* (Sinauer Associates Sunderland).
14. Chapman JA, et al. (2010) The dynamic genome of Hydra. *Nature* 464(7288):592–596.
15. Chourrout D, et al. (2006) Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* 442(7103):684–687.
16. DuBuc TQ, Ryan JF, Shinzato C, Satoh N, Martindale MQ (2012) Coral comparative genomics reveal expanded Hox cluster in the cnidarian-bilaterian ancestor. *Integr Comp Biol* 52(6):835–841.
17. Tautz D, Domazet-Loso T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* 12(10):692–702.
18. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 25(9):404–413.
19. Lehnert EM, et al. (2014) Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3* 4(2):277–295.
20. Wang JT, Douglas AE (1999) Essential amino acid synthesis and nitrogen recycling in an alga–invertebrate symbiosis. *Mar Biol* 135(2):219–222.
21. Ganot P, et al. (2011) Adaptations to endosymbiosis in a cnidarian-dinoflagellate association: differential gene expression and specific gene duplications. *PLoS Genet* 7(7):e1002187.
22. Dani V, Ganot P, Priouzeau F, Furla P, Sabourault C (2014) Are Niemann-Pick type C proteins key players in cnidarian-dinoflagellate endosymbioses? *Mol Ecol* 23(18):4527–4540.
23. Giner JL (1993) Biosynthesis of marine sterol side chains. *Chem Rev* 93(5):1735–1752.

24. Withers NW, Kokke WC, Fenical W, Djerassi C (1982) Sterol patterns of cultured zooxanthellae isolated from marine invertebrates: Synthesis of gorgosterol and 23-desmethylgorgosterol by aposymbiotic algae. *Proc Natl Acad Sci U S A* 79(12):3764–3768.
25. Kokke WCMC, Fenical W, Bohlin L, Djerassi C (1981) Sterol synthesis by cultured zooxanthellae; implications concerning sterol metabolism in the host-symbiont association in caribbean gorgonians. *Comp Biochem Physiol Part B Comp Biochem* 68(2):281–287.
26. Yamashiro H, Oku H, Higa H, Chinen I, Sakai K (1999) Composition of lipids, fatty acids and sterols in Okinawan corals. *Comp Biochem Physiol Part B Biochem Mol Biol* 122(4):397–407.
27. Bosch TC (2013) Cnidarian-microbe interactions and the origin of innate immunity in metazoans. *Annu Rev Microbiol* 67:499–518.
28. Fraune S, Bosch TC (2007) Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. *Proc Natl Acad Sci U S A* 104(32):13146–13151.
29. Gochfeld DJ, Aeby GS (2008) Antibacterial chemical defenses in Hawaiian corals provide possible protection from disease. *Mar Ecol Prog Ser* 362:119–128.
30. Bayer T, et al. (2013) The microbiome of the Red Sea coral *Stylophora pistillata* is dominated by tissue-associated Endozoicomonas bacteria. *Appl Env Microbiol* 79(15):4759–4762.
31. Vidal-Dupiol J, et al. (2011) Innate immune responses of a scleractinian coral to vibriosis. *J Biol Chem* 286(25):22688–22698.
32. Fusetani N, Toyoda T, Asai N, Matsunaga S, Maruyama T (1996) Montiporic acids A and B, cytotoxic and antimicrobial polyacetylene carboxylic acids from eggs of the scleractinian coral *Montipora digitata*. *J Nat Prod* 59(8):796–797.
33. Kvennefors EC, Leggat W, Hoegh-Guldberg O, Degnan BM, Barnes AC (2008) An ancient and variable mannose-binding lectin from the coral *Acropora millepora* binds both pathogens and symbionts. *Dev Comp*

Immunol 32(12):1582–1592.

34. Wood-Charlson EM, Hollingsworth LL, Krupp DA, Weis VM (2006) Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. *Cell Microbiol* 8(12):1985–1993.
35. Fujita T (2002) Evolution of the lectin-complement pathway and its role in innate immunity. *Nat Rev Immunol* 2(5):346–353.
36. Zhang SM, Adema CM, Kepler TB, Loker ES (2004) Diversification of Ig superfamily genes in an invertebrate. *Science (80-)* 305(5681):251–254.
37. Kvennefors EC, et al. (2010) Analysis of evolutionarily conserved innate immune components in coral links immunity and symbiosis. *Dev Comp Immunol* 34(11):1219–1229.
38. Kimura A, Sakaguchi E, Nonaka M (2009) Multi-component complement system of Cnidaria: C3, Bf, and MASP genes expressed in the endodermal tissues of a sea anemone, Nematostella vectensis. *Immunobiology* 214(3):165–178.
39. Poole AZ, Weis VM (2014) TIR-domain-containing protein repertoire of nine anthozoan species reveals coral-specific expansions and uncharacterized proteins. *Dev Comp Immunol* 46(2):480–488.
40. Miller DJ, et al. (2007) The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. *Genome Biol* 8(4):R59.
41. Franzenburg S, et al. (2012) MyD88-deficient *Hydra* reveal an ancient function of TLR signaling in sensing bacterial colonizers. *Proc Natl Acad Sci U S A* 109(47):19374–19379.
42. Lange C, et al. (2011) Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol* 28(5):1687–1702.
43. Hamada M, et al. (2013) The complex NOD-like receptor repertoire of the coral *Acropora digitifera* includes novel domain combinations. *Mol Biol Evol* 30(1):167–176.
44. Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. *Trends Genet* 27(4):157–163.
45. Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene

transfer in bdelloid rotifers. *Science (80-)* 320(5880):1210–1213.

46. Artamonova II, Mushegian AR (2013) Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts. *Appl Env Microbiol* 79(22):6868–6873.
47. Starcevic A, et al. (2008) Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, *Nematostella vectensis*, have microbial origins. *Proc Natl Acad Sci U S A* 105(7):2533–2537.
48. Balskus EP, Walsh CT (2010) The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. *Science (80-)* 329(5999):1653–1656.
49. Waller RF, Slamovits CH, Keeling PJ (2006) Lateral gene transfer of a multigene region from cyanobacteria to dinoflagellates resulting in a novel plastid-targeted fusion protein. *Mol Biol Evol* 23(7):1437–1443.
50. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L (2012) Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct* 7:18.
51. Deng Q, Barbieri JT (2008) Molecular mechanisms of the cytotoxicity of ADP-ribosylating toxins. *Annu Rev Microbiol* 62:271–288.
52. Franzenburg S, et al. (2013) Distinct antimicrobial peptide expression determines host species-specific bacterial associations. *Proc Natl Acad Sci U S A* 110(39):E3730–8.
53. Sunagawa S, et al. (2009) Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its dinoflagellate endosymbiont. *BMC Genomics* 10:258.
54. Xiang T, Hambleton EA, DeNofrio JC, Pringle JR, Grossman AR (2013) Isolation of clonal axenic strains of the symbiotic dinoflagellate Symbiodinium and their growth and host specificity. *J Phycol* 49(3):447–458.
55. Pettay DT, LaJeunesse TC (2007) Microsatellites from clade B *Symbiodinium spp.* specialized for Caribbean corals in the genus

Madracis. *Mol Ecol Notes* 7(6):1271–1274.

56. LaJeunesse TC, Parkinson JE, Reimer JD (2012) A genetics-based description of *Symbiodinium minutum* sp. nov. and *S. psygmophilum* sp. nov. (Dinophyceae), two dinoflagellates symbiotic with cnidaria. *J Phycol* 48(6):1380–1391.
57. Altshuler D, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407(6803):513–516.
58. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
59. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108(4):1513–1518.
60. Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. *Genome Biol* 13(6):R56.
61. Boetzer M, Henkel C V, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.
62. Shoguchi E, et al. (2013) Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol* 23(15):1399–1408.
63. Bradnam KR, et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2(1):10.
64. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1:i351–8.
65. Smit AFA, Hubley R, Green PJ RepeatMasker Open-3.0. Available at: http://www.repeatmasker.org.
66. Rodríguez E, et al. (2014) Hidden among Sea Anemones: The First Comprehensive Phylogenetic Reconstruction of the Order Actiniaria (Cnidaria, Anthozoa, Hexacorallia) Reveals a Novel Group of Hexacorals. *PLoS One* 9(5):e96998.
67. Gusmao LC, Daly M (2010) Evolution of sea anemones (Cnidaria:

Actiniaria: Hormathiidae) symbiotic with hermit crabs. *Mol Phylogenet Evol* 56(3):868–877.

68. Rodríguez E, Barbeitos M, Daly M, Gusmão LC, Häussermann V (2012) Toward a natural classification: phylogeny of acontiate sea anemones (Cnidaria, Anthozoa, Actiniaria). *Cladistics* 28(4):375–392.
69. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948.
70. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server issue):W609–12.
71. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13(5):555–556.
72. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
73. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
74. Consortium TU (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42(Database issue):D191–8.
75. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
76. Dimmer EC, et al. (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* 40(Database issue):D565–70.
77. Haas BJ, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654–5666.
78. Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32(Web Server issue):W309–12.
79. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12(4):656–664.
80. Liew YJ, et al. (2014) Identification of microRNAs in the coral *Stylophora*

pistillata. *PLoS One* 9(3).

81. Thomas PD, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13(9):2129–2141.
82. Finn RD, et al. (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222–30.
83. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.
84. R Development Core Team (2012) R: A language and environment for statistical computing. Available at: http://www.r-project.org/.
85. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28(1):27–30.
86. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
87. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 10(1):71–73.
88. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
89. Howe EA, Sinha R, Schlauch D, Quackenbush J (2011) RNA-Seq analysis in MeV. *Bioinformatics* 27(22):3209–3210.
90. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
91. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.
92. Ronquist F, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539–542.
93. Tavare S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci* 17:57–86.

94. Fisher RA (1922) On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. *J R Stat Soc* 85(1):87–94.
95. Dunn OJ (1961) Multiple Comparisons among Means. *J Am Stat Assoc* 56(293):52–64.
96. Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641–1650.
97. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23(10):1282–1288.
98. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
99. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164–1165.
100. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30(12):2725–2729.
101. Kearse M, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–1649.
102. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.
103. Boschetti C, et al. (2012) Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. *PLoS Genet* 8(11):e1003035.
104. Moran Y, et al. (2008) Concerted evolution of sea anemone neurotoxin genes is revealed through analysis of the Nematostella vectensis genome. *Mol Biol Evol* 25(4):737–747.
105. Zhang G, et al. (2012) The oyster genome reveals stress adaptation and

complexity of shell formation. *Nature* 490(7418):49–54.

106. Takeuchi T, et al. (2012) Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. *DNA Res* 19(2):117–130.

107. Flot JF, et al. (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. *Nature* 500(7463):453–457.

Chapter 3: Insights into miRNA-mediated host transcriptome modulations in the cnidarian-dinoflagellate endosymbiosis of *Aiptasia*

Sebastian Baumgarten[a], Maha J. Cziesielski[a], Ludivine Thomas[a], Craig T. Michell[a], Lisl Y. Esherick[b], Manuel Aranda[a], John R. Pringle[a], Christian R. Voolstra[a]

[a] Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[b] Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

3.1 Abstract

Cnidarians living in symbiotic relationships with dinoflagellates are known to undergo transcriptomic changes during the infection with *Symbiodinium* and maintenance of a stable endosymbiont population. However, the precise regulatory mechanisms modulating the host transcriptome are not known. Here we report on the analysis of post-transcriptional gene regulation by miRNAs in the small sea anemone *Aiptasia*, a model system for cnidarian-dinoflagellate host-microbe interactions. We find that *Aiptasia* encodes mainly species-specific miRNAs and provide evidence for recent differentiation of miRNAs within the *Aiptasia* genome that are commonly conserved among anthozoans. Our analysis of miRNA expression shows that both conserved and species-specific host miRNAs are differentially expressed in response to endosymbiont infection. Using cross-linking immunoprecipitations of Argonaute, the central protein of the miRNA-induced silencing complex, we identified miRNA binding sites on a transcriptome wide scale. We found that symbiotically regulated miRNAs target genes that were previously recognized to be involved in biological processes related to *Symbiodinium* infection. Taken together, our findings provide evidence for an effective role of miRNAs in modulating the host transcriptome during the onset and maintenance of the cnidarian-dinoflagellate endosymbiosis.

3.2 Introduction

The trophical and structural basis of one of the most important marine ecosystems, coral reefs, relies on the functional endosymbiosis of a cnidarian coral host and their photosynthetic dinoflagellate symbionts in the genus

Symbiodinium that reside within vesicles (symbiosomes) of the host's gastrodermal cells. In this mutualistic endosymbiosis, the cnidarian (e.g. coral) host offers a sheltered environment to the symbiont and provides the inorganic nutrients needed for photosynthesis, whereas the symbiont transfers the majority of its photosynthates to the host (1). This tight nutrient coupling allows the animal-alga pair to grow in otherwise extremely oligothrophic waters. Most symbiotic cnidarians acquire their endosymbionts with high specificity from the environment ("horizontally") (2), either after sexual reproduction during the host's larval phase or during the recovery of a stress-induced symbiosis breakdown (i.e. "bleaching"). There is growing evidence that the acquisition of the dinoflagellate symbiont as well as the establishment and maintenance of the endosymbiosis is accompanied by specific changes of host gene expression, including genes involved in pattern recognition and innate immunity, endocytosis and trans-membrane nutrient transport (3–8). These seemingly specific transcriptional responses also suggest an effective role of higher-level regulatory mechanisms that help in the modulation and orchestration of gene expression that are still elusive.

MicroRNAs (miRNAs) are small (~22 nt) noncoding RNAs that post-transcriptionally down-regulate gene expression through complementary binding of the target mRNAs. In this process, the mature miRNA is first bound to an Argonaute protein and guides the latter to binding sites within the target transcript that are partially complementary to the miRNA, resulting in translational inhibition or target degradation (9). One miRNA can have a multitude of mRNA targets and

the action of individual miRNAs often includes widespread downstream effects, making miRNAs regulatory hubs for the fine-tuning of gene expression and the orchestration of broad transcriptomic changes (10). Recent studies have shown that miRNAs are also pervasively involved in the regulation of and response to eukaryotic endoparasites, e.g. *Leishmania donovani* and putatively also *Plasmodium falciparium* (11, 12), making miRNA expression and regulation of target genes highly attractive candidates for transcriptome modulation in the cnidarian-dinoflagellate endosymbiotic relationship.

Using the cnidarian model system *Aiptasia* and its recently sequenced genome (4), we sought to investigate the role of host miRNA regulation by integrating small RNA and mRNA sequencing over a range of different states of endosymbiont infection. Together with cross-linking immunoprecipitations (CLIP) of the *Aiptasia* Ago protein, we identified high confidence ternary interactions between the Ago effector protein, the regulatory miRNA and its target mRNA transcript on a transcriptome wide scale. This approach allowed us to reveal miRNA-mediated modulations of genes and pathways implicated in biological processes that are regarded contributory for a functional symbiosis of cnidarians.

3.3 Results

3.3.1 Annotation of conserved and species-specific miRNAs in the *Aiptasia* genome

We annotated the *Aiptasia* miRNA repertoire using small RNA libraries sampled from *Aiptasia* polyps at different states of infection with their natural

endosymbiont *Symbiodinium minutum* (strain SSB01). Four replicates of *Aiptasia* polyps were sampled each in an aposymbiotic state (without endosymbionts), at an intermediate state (12 days post infection [p.i.]), and a fully symbiotic state (30 days p.i.) (Materials and Methods) (Figs. 3.1A,B, Table 3.1).

The initial inspection of the *Aiptasia* small RNA repertoire revealed a distinct length distribution with two small RNA populations being seemingly overrepresented, one with nucleotide lengths of 19-21 nt and a second population of larger small RNAs with 27 - 31 nt (Fig. 3.2A). The former comprised putative miRNAs as well as small-interfering RNAs (siRNAs), whereas the latter represented putative PIWI-interacting RNAs (piRNAs) that constitute the majority of the sequenced small RNAs. The length distribution as well as a strong bias towards uridine in the 5' identity of both small RNA populations represents a distinct pattern that is consistent with previous small RNA characterization efforts in the cnidarian phylum (13–17).

From the pooled small RNA samples (see above), we annotated 46 high-confidence miRNAs that are transcribed from a total of 60 distinct genomic loci (Dataset S3.1). These mature *Aiptasia* miRNAs were stringently filtered for known characteristics of metazoan miRNAs that result from the evolutionary conserved biogenesis pathway of miRNAs in bilaterians and cnidarians (Materials and Methods) (15). These characteristics include 1) a distinct length of ~ 21 nt, 2) folding of the precursor miRNA transcript (pre-miRNA) into a hairpin structure of ~ 60 nt; 3) homogenous 5' ends of the mature miRNA reads resulting from the precise cleavage of the pre-miRNA by the Dicer endonuclease, and 4)

uneven frequencies of small RNA reads mapping to either side of the hairpin pre-miRNA stem, which is attributed to the strand selection process in which the mature miRNA is bound to the Ago effector protein, whereas the complementary small RNA sequence of the pre-miRNA stem is degraded (Fig. 3.2B). Eleven of the 46 *bona fide* miRNA were previously annotated in other anthozoan species, with the largest overlap (n = 10) being with the closest relative of *Aiptasia,* the starlet sea anemone *Nematostella vectensis.* (Fig. 3.2C) (13–17).

3.3.2 Identification of *in vivo* mRNA-miRNA protein interactions

In order to further confirm the identity of *Aiptasia* miRNAs as well as to identify functional molecular interactions between miRNAs and their cognate target mRNAs, we conducted cross-linking immunoprecipitation (CLIP) experiments of *Aiptasia* Ago, the core protein of the miRNA induced silencing complex (miRISC). *Aiptasia* encodes for two Ago proteins that differ in the number of amino acid residues, similar to their two homologs in *N. vectensis* (18). The canonical *Aiptasia* Ago protein (AipAgo1) is 853 aa long, similar to the Ago proteins found in bilaterians, whereas the second protein is considerably longer (1,083 aa). For our further experiments, we focused on the canonical AipAgo1 and designed polyclonal antibodies (Ab) against a peptide antigen in the protein C-terminus Fig. 3.3A). First, we confirmed the specificity of our Ab by Western Blot and retrieved AipAgo1 at the expected molecular weight (~96 kDa, Fig. 3.3B). Using whole mount immunohistochemical staining of adult *Aiptasia* polyps, we also find that AipAgo1 is present in the whole organism (Fig. 3.3C) and on a subcellular

level, AipAgo1 seems to be distributed throughout the cytoplasm, but tends to cluster locally at sites putatively corresponding to Processing (P-) bodies, a pattern that has been observed previously (Fig. 3.3D) (19).

We followed a CLIP protocol that was previously applied successfully for tissue samples and cell cultures to identify RNA-protein interactions, including the annotation of ternary miRNA-mRNA-Ago interactions (20, 21) (Fig. 3.4A, Materials and Methods). Following *in vivo* UV-cross-linking of miRNAs to their target mRNAs and the Ago protein, a central step in this protocol includes RNase digestion of free mRNA tails, both up- and downstream of the miRNA-Argonaute binding sites. The protection from nuclease digestion by the Ago protein ensures that only miRNAs and mRNA sequences (i.e., CLIP tags) corresponding to the putative binding site of the Argonaute protein are further processed. Exclusive sequencing of CLIP tags enables the subsequent detection of Ago binding sites with high confidence by searching for specific enrichments (i.e., peaks) of CLIP tags at sites of putative miRNA-mRNA interaction (Figs. 3.4A and 3.5A).

We performed CLIP experiments on four biological samples with two replicates of aposymbiotic and fully symbiotic anemones each. In a first step, we used the immunoprecipitated RNAs to confirm the bioinformatically annotated miRNAs and could verify that all of the 46 predicted pre-miRNA were mapped by miRNAs that were co-immunoprecipitated with AipAgo1. In addition, we found that the numbers of CLIP tags mapping to the annotated pre-miRNAs strongly correlate between the replicates (Fig. 3.5B) as well as the two symbiotic states (Fig. 3.5C).

Applying stringent criteria of processing and filtering (Materials and Methods, Fig. 3.5A), we next aligned the CLIP tags to the *Aiptasia* genome. We obtained a total of 112,903 and 195,897 distinct regions of CLIP tag alignments from the two aposymbiotic and symbiotic samples, respectively (Table 3.2). Inspection of clusters of overlapping CLIP tags at distinct genomic loci showed that the height of the CLIP peaks (i.e., the number of overlapping CLIP tags per cluster) also correlated well between the two replicates and also between the two symbiotic states (Fig. 3.6A). The correlation of replicates observed in the Ago-RNA interactions suggests a high technical reproducibility and robustness of the CLIP protocol, similar to observations in previous CLIP experiments (20, 22). In addition, the correlations between the aposymbiotic and symbiotic samples in both the miRNA-Ago (Fig. 3.5B) as well as the mRNA-Ago interactions (Fig. 3.6A) might indicate that detectable, biological differences of these interactions between the two symbiotic states are few, suggesting a high specificity of the binding sites independent of the symbiotic state. Following this coherence of Ago-RNA interactions, we pooled the CLIP tags of all four biological samples for subsequent analyses, although we acknowledge the possibility of infection state specific differences in miRNA-mRNA-Ago interactions.

We expected *bona fide* RNA-protein interactions to be captured multiple times independently over CLIP experiments, which produces clusters of overlapping CLIP tags. For this reason, we determined significant CLIP peaks within exonic regions using the modified False Discovery Rate algorithm proposed by Yeo *et al.* (23) implemented in pyicoclip (24) (Materials and Methods). This produced a

set of 2,269 CLIP peaks (FDR ≤ 0.01) covering genomic regions with an average length of 60 nt and representing putative sites of mRNA-Ago binding. The majority (63%) of the significant RNA tag clusters were present within the coding sequences of transcripts, with an additional 27% and 10% falling within the 5' and 3' UTRs of transcripts, respectively (Fig. 3.6B, Materials and Methods).

To assign interactions between the identified mRNA-Ago binding sites and the sites of functional miRNA-mRNA binding, we next mapped the mature miRNAs to the identified CLIP peak sequences, which resulted in 3,377 interactions between any of the miRNAs and the mRNA-Ago sites that feature a minimum free binding energy (MFE) of ≤ -15 kcal/mol (Dataset S3.2). Of those, 619 CLIP peak sequences feature more than one equally good miRNA-mRNA interaction. Further inspection of the latter showed that 226 CLIP peak sequences were targeted by the two almost identical miRNAs, namely miR-2022 and miR13, while another 30 CLIP peak sequences were targeted by the same miRNA multiple times equally well (i.e., with the same number of binding miRNA nucleotides, see Materials and Methods).

In higher metazoans, it has been shown that binding of as few as 7 nucleotides in the 5' seed region of a mature miRNA is efficient for effective miRNA-mRNA binding (25). In contrast to that, we found that many of our interactions seemed to include a much higher number of binding nucleotides from the miRNA, with a third of the interactions (n = 1,061) featuring a seemingly high complementarity to the mRNA target with ≥ 17 miRNA nucleotides involved in the binding interaction (Fig. 3.6C). This higher complementarity between miRNAs and their cognate

mRNA targets was previously described for cnidarians (15), and seems to be frequently related to the regulation of the target mRNA by slicing, a mode of action similar to plant miRNAs and presumably an ancient state of miRNA function (15). One example of such regulation through high complementarity is potentially represented by the binding of spi-miR-7 to a protein of the *Aiptasia* cnidarian Ficolin-like protein (CniFL) family, a class of putative pattern recognition receptors that was only recently described in the cnidarian phylum and appears to be only present in symbiotic anthozoans (4) (Fig. 3.6D).

3.3.3 *Aiptasia* miRNA are regulated during the onset and maintenance of the endosymbiosis

To further assess the role of miRNA post-transcriptional regulation in the onset and maintenance of the cnidarian-dinoflagellate endosymbiosis, we analyzed changes in miRNA expression among three different time points of infection with *Symbiodinium* (Fig. 3.1, Materials and Methods). Overall, *Aiptasia* miRNAs seem to be strongly symbiotically regulated (Fig. 3.7A), with 12 of the 46 annotated miRNAs altering their expression level significantly (FDR ≤ 0.01) between aposymbiotic *Aiptasia* and fully symbiotic anemones (Fig. 3.7B). Among the significantly differentially expressed miRNAs, the majority is upregulated between 1.98 to 11.98 fold in symbiotic anemones, with the species-specific aip-new-miR2 (Fold Change = 3.35) and aip-new-miR-24 (Fold Change = 11.98) featuring the overall highest and lowest average expression levels, respectively (Fig. 3.7C). Comparison of the fold changes of miRNA expression between different time

points of *Symbiodinium* infection revealed that the majority of miRNAs already show a similar intensity of upregulation at an intermediate state of infection compared to fully symbiotic *Aiptasia.* Only few miRNAs showed additional increases in expression in fully symbiotic anemones (Fig. 3.7D), which might indicate priming or front-loading of miRNA expression and regulation independent of the total count of intracellular *Symbiodinium.*

Interestingly, two of the highest upregulated miRNAs, miR2022 and aip-new-miR-13, are positioned in one genomic region spanning 275 kb that includes two loci encoding for a precursor transcript for aip-new-miR13 and one for miR-2022 (Fig. 3.8A, left). miR-2022 is strongly conserved within anthozoans (Fig. 3.2C) and a single locus encoding this miRNA is found in the genomes of *Stylophora pistillata* and *Nematostella vectensis* (15, 17). The mature miRNA sequences of miR-2022 and aip-new-miR13 are highly similar and differ in only one nucleotide position in the miRNA 3' region (Fig. 3.8A, right), which, together with the close genomic arrangement of the three loci, suggests a recent duplication of the miR-2022 locus in the *Aiptasia* lineage. To assess a putative co-control of these three miRNAs, we searched for conserved sequence motifs in the upstream region of the miRNA loci that could hint towards similar regulation. Interestingly, we found that the two loci of miR13, separated by more than 275 kb, share several long and highly conserved motifs in an 1kb upstream region (Fig. 3.8B), whereas the upstream regions of miR-2022 and any of the miR-13 loci are much less conserved. In contrast, the 1kb downstream regions of the three miRNA loci don't share any sequence conservation (Fig. 3.8C). Since the exact transcription start

site of the primary (pri-) miRNA transcript for any of the three miRNA loci is unknown, the conserved motifs between the two miR13 loci are potentially lying within the pri-miRNA transcript, making it possible that the high level of upstream conservation is implicated in either co-control of miRNA transcription or post-transcriptional processing of the pri-miRNA transcript.

3.3.4 *Aiptasia* miRNAs modulate genes involved in symbiont acquisition, signaling, and transmembrane transport

Post-transcriptional control of gene expression by miRNAs usually results in the downregulation of cognate target genes, either through endonucleatic cleavage or inhibition of mRNA translation (26). To assess any such changes in mRNA transcript abundance, we assessed gene expression from the same total RNA samples of the *Symbiodinium* infection experiment that we also used for small RNA sequencing and subsequent miRNA analyses (4). Combining simultaneous changes in gene and miRNA expression with miRNA-mRNA interactions of the CLIP experiment enabled us to further assess the putative role of miRNA in the endosymbiosis. We find a total of 885 interactions identified by CLIP between significantly differentially expressed miRNAs and their putative mRNA targets, among which we identify several target genes that were previously implicated in the onset and maintenance of cnidarian-dinoflagellate endosymbioses. These include a membrane receptor of the *transforming growth factor β receptor* family (*TGFβR*) (27), as well as the downstream messenger of *TNF-receptors, TRAF* (28). In addition, we identified a signalling cascade including the *Fibroblast-*

growth factor receptor (*FGFR*), its extracellular ligand *FGF*, and one of its downstream intracellular messengers, *GRB2,* to be under putative post-transcriptional miRNA control (Fig. 3.9A). Although we didn't find significant changes in expression of *TGFβR*, we found the expression levels of the *FGFR, FGF*, and the *TRAF* genes to decrease considerably in the symbiotic state (Fig. 3.9B). Interestingly, we also identified genes to interact with miRNAs that encode for proteins involved in the maintenance of the *Symbiodinium*-containing vesicle (i.e., symbiosome), including a homolog of *LAMP1*, a protein related to lysosome - and potentially symbiosome - trafficking and maturation as well as homologs of the sterol transporter *NPC1* and the peptide transporter *ABCB9*. In the case of *LAMP1* and *NPC1*, these genes are also down regulated in a fully symbiotic state (Fig. 3.9B), which might be partially attributable to the binding of their regulatory miRNAs.

3.4 Discussion

The establishment and maintenance of the cnidarian-dinoflagellate endosymbiosis is generally accompanied by distinct changes in host gene expression (3-8). In this study, we set out to annotate the miRNA repertoire of *Aiptasia,* analyze their expression patterns, and identify mRNA targets in order to elucidate their effective role as regulatory elements of the host transcriptome in the endosymbiosis with *Symbiodinium.*

Our annotation of *Aiptasia* miRNAs revealed an unusually low conservation of miRNA families within the cnidarian phylum compared to bilaterians, a feature

that seems to be inherent to all cnidarian species analyzed to date (15). Additionally, even within the miRNAs that are conserved among anthozoan anemones, we identified species-specific differences, namely the duplications of the miR-2022 locus, indicating lineage-specific functional divergence, as suggested by the evolving mature miRNA itself and the differentiating upstream regions of the miRNA loci. In particular, the conservation of an individual nucleotide substitution in the 3' region of miR-13 implies a functional importance of nucleotide binding beyond the 5' 'seed' binding. This finding is consistent with miRNA target studies in *N. vectensis* (15) and is also corroborated by the overall high number of miRNA nucleotides involved in mRNA target binding that we identified in the miRNA-mRNA interactions.

Although it is likely that the host transcriptome is regulated by more than one mechanism in response to *Symbiodinium* infection, the consistent changes of miRNA expression levels shown in this study provides evidence of their effective role as regulators of this endosymbiosis. The interaction of these miRNAs with *FGFR* and especially with *TGFβR* and the *TNFR/TRAF* pathways (Fig. 3.9), which were previously recognized to play key roles in the priming of immune and stress responses of symbiotic cnidarians, corroborates the hypothesis of a co-option of these signalling cascades as well as their regulatory elements to set up a stable endosymbiosis (27, 28).

In addition, especially proteins that are located at the symbiosome membrane are known to be highly important for the maintenance of the endosymbiosis, including proteins involved in symbiosome maturation (29, 30) as well as cross-

membrane nutrient exchange (3, 31). This might also hold true for the *Aiptasia* homologs of *LAMP1* and *NPC1* whose putatively critical fine-tuning of expression levels and function at the host-symbiont interface might be achieved in parts through their post-transcriptional modulation by miRNAs.

Our data suggest that miRNA are not only involved in defining the susceptibility of host organisms to endoparasites (11), but that post-transcriptional modulations of gene expression are also effective in mutualistic relationships such as the dinoflagellate endosymbiosis in anemones and corals. Such fine-tuning of transcription might add an additional layer of regulation for a wide range of biological processes and physiological responses including stress responses or host-symbiont specificities that should especially taken into account in future experiments that are focusing on gene expression on a transcriptional level only.

In summary, by performing a comprehensive analysis of miRNA expression and identification of *in vivo* ternary miRNA-mRNA-Ago interactions, we describe miRNA-mediated modulations of the cnidarian host transcriptome in response to *Symbiodinium* infection and provide novel insights into cnidarian gene regulatory mechanisms in general.

3.5 Materials and Methods

3.5.1 *Aiptasia* culture

Aiptasia anemones (strain CC7) were used for all experiments and were cultured as described previously (4). Briefly, anemones were kept in autoclaved seawater (ASW) at 25°C on a 12h:12h light-dark cycle and fed once a week with freshly hatched *Artemia nauplii* (brine shrimp) larvae. Water changes were performed on the days following feeding.

To generate dinoflagellate-free (aposymbiotic) *Aiptasia*, bleaching (i.e. the expulsion of the dinoflagellate endosymbionts) was initiated by subjecting the anemones to consecutive cold shocks in pre-chilled ASW for 4 h at 4°C. The bleached *Aiptasia* were subsequently kept in the dark for ≥ 6 month at 25°C with usual feeding and water change. Prior to experimental use, aposymbiotic *Aiptasia* were acclimatized to a 12 h:12 h light-dark cycle for at least two weeks to acclimatize the aposymbiotic anemones to experimental conditions for further experiments (i.e., *Symbiodinium* infection and CLIP, see below)

3.5.2 *Symbiodinium* infection experiment and RNA sequencing

A detailed protocol of the infection experiment is provided in the Supplemental Information at the end of this Chapter.

To obtain RNA samples from *Aiptasia* anemones during different states of *Symbiodinium* infection, adult aposymbiotic anemones were infected with

Symbiodinium minutum (strain SSB01) from a homologous, exponentially growing culture. The anemones were sampled at three time points over a total period of 30 days as described previously (4). Briefly, adult anemones were kept in 15 ml 6-well plates (Falcon # 353046) (~3-5 anemones per well) with autoclaved and sterile-filtered seawater (ASFW) and acclimatized for 16 days with regular feeding and water change (see above). The infection time course was laid out as follows: Day 1, addition of *Symbiodinium* cells to a final concentration of ~10^5 cells/ml; day 2, feeding with brine shrimp larvae without change of ASFW or adding additional *Symbiodinium*; day 3, change of ASFW and addition of algae at ~10^5 cells/ml; day 11, end of incubation with *Symbiodinium* cells by placing the anemones into fresh 6-well plates without further addition of *Symbiodinium* and regular feeding and water changes until final sampling after 30 days. For each time point (i.e., infection state), the anemones from one 6-well plate were pooled into one biological replicate and four of those replicates were sampled at day 1 (aposymbiotic anemones), day 12 (intermediate infected anemones), and day 30 (symbiotic anemones). Total RNA was extracted using TRIzol (Thermo Fisher # 15596-026) and small RNAs were fractionated using the mirVana small RNA enrichment kit (Thermo Fisher # AM1560) following the manufacturers' instructions.

All small RNA sequencing libraries were prepared using the Illumina TruSeq Small RNA Sample Library Kit (Illumina # RS-200-0012) and sequenced as single end reads on one lane of the Illumina HiSeq 2000 platform. The small RNA sequencing libraries are deposited at the NCBI Sequence Read Archive

(SRA) (accession no.: aposymbiotic: SRX1351928; intermediate: SRX1351929; symbiotic: SRX1351930, four replicates each, see Table 3.2). The processing of mRNA samples as well as the analysis of differentially expressed genes from the infection experiment is described in (4), mRNA sequencing libraries can be accessed at the NCBI SRA (accession no.: aposymbiotic: SRX757525; intermediate: SRX757526; symbiotic: SRX757528, four replicates each). To estimate the relative progression of *Symbiodinium* infection, the mRNA sequences were mapped to a *S. minutum* reference transcriptome (NCBI BioProject Accession No. PRJNA274852) using bowtie2 (32). The percentages of read pairs that aligned to the symbiont transcriptome without mismatches are shown in Fig. 3.1.

3.5.3 miRNA annotation and motif analysis

The small RNA reads were first trimmed by removing adapter sequences and filtered for low quality reads using Trimmomatic (33). Only small RNAs between 15 - 32 nucleotides in length were retained and unique sequences were collapsed using the collapse_reads.pl script from the miRDeep2 package (34). Next, we annotated regions of the *Aiptasia* genome encoding for non-coding RNAs (ncRNA), including snoRNAs, snRNAs, tRNAs, and eukaryotic rRNAs. Reference ncRNA sequences were retrieved from the RNAcentral database (35) and aligned to the *Aiptasia* genome using BLASTN (evalue cutoff ≤ 1e-5) (36). The *Aiptasia* ncRNA regions confirmed by this alignment were then used as a database for mapping of the small RNA libraries using bowtie2 (32). Small RNA

reads mapping to the *Aiptasia* ncRNA regions were subsequently removed from the further analysis.

miRNAs were annotated from the pre-processed and pooled small RNA libraries (Fig. 3.2A, Table 3.1) using the miRDeep2 package at default settings (34). To identify putatively conserved miRNAs previously annotated in other cnidarian genomes, we added a reference library of mature miRNA sequences from *N. vectensis* (13, 15), *H. magnipapillata* (16), and *S. pistillata* (17). An initial set of 145 miRNAs was predicted, and further inspected for characteristics required for *bona fide* miRNA annotation. These include, 1) folding of the precursor miRNA transcript (pre-miRNA) into an unbranched hairpin structure, 2) homogenous 5' ends of small RNA reads mapping to the pre-miRNA hairpin, 3) differing frequencies of mature miRNA and star miRNA reads mapping to the pre-miRNA stem, as well as 4) 2 nt overhangs of the mature miRNA and the star miRNA sequence in the miRNA duplex (sensu 37, 38). Applying these additional stringent criteria, we identified 46 miRNA mapping to 68 genomic loci. We then used the built-in scoring mechanism of miRDeep2 to assign putative probabilities of the 46 miRNA to be true-positives, performing 100 rounds of permutation (34, 39). For 45 of the miRNA, the true-positive probabilities were ≥ 84%, except for the evolutionary conserved miR-9454. Reviewing the prediction of miR-9454 shows that no star sequences of this miRNA were recovered in the *Aiptasia* small RNA libraries, decreasing its probability to be identified by the algorithm. However, because of the perfect conservation of the mature miRNA sequence, we retained miR-9454 for subsequent analyses.

To identify motifs putatively involved in the transcriptional control of the miR-2022 loci, we extracted the sequence 1 kb upstream of the pre-miRNAs start position of each locus. The motif analysis was performed using the MEME web suite (40), searching for up to 10 motifs that occur zero or only one time per region (Fig. 3.8B). For comparison, the 1 kb downstream sequence of each locus was also used for motif identification using the same settings (Fig. 3.8C).

3.5.4 AipAgo1 antibody design and Western Blot

Aiptasia Argonaute (AipAgo) proteins were initially identified by BLASTP alignments of the *N. vectensis* Ago protein sequences (UniProt accession no. U3MIG1, U3MH35) against the *Aiptasia* genomic protein set (http://aiptasia.reefgenomics.org) (E-value ≤ 1e-5).

An antigen of 25 amino acids in length corresponding to residues 811 - 835 of the AipAgo1 C-terminus (Fig. 3.3A) was synthesized by Acris Antibodies (Herford, Germany) and used for the generation of polyclonal antibodies (Ab). The Ab was raised in rabbits and affinity-purified using the synthesized antigen.

For Western Blots of AipAgo1*, Aiptasia* protein lysates (see below) were resolved by size on an SDS polyacrylamide gel for ~1.5 h at 120 V using a 2-gel Tetra and Blotting Module (Bio-Rad # 1660828EDU). The proteins were transferred to a PVDF membrane for 2 h at 100 V using the same 2-gel module as for size selection (Bio-Rad # 1660828EDU) and blocked overnight at 4°C using StartingBlock T20 (PBS) (Thermo Fisher # 37538). For AipAgo1 detection, the membrane was incubated for 2 h at 37°C with 20 µg Ab diluted 1:1,000 in TRIS

buffered saline supplemented with 0.05% (v/v) Tween20 (TBST), followed by 3 h incubation at room temperature with a secondary goat-anti rabbit DyLight Fluor 488 antibody (Thermo Fisher # A-11034) diluted 1:10,000 in TBST. Protein bands were visualized using a Typhoon Scanner 9410 (GE Healthcare).

3.5.5 Immunohistochemical staining

Aposymbiotic *Aiptasia* anemones were macerated according to David (41). Briefly, whole animals were placed in a maceration solution containing glycerine, glacial acetic acid, and water (ratio 1:1:13) at room temperature. The tissue was left soaking for 30 minutes with occasional gentle shaking until the animals were visibly dissociated. The resulting cell suspension was fixed by addition of 0.1 volumes of 20% formaldehyde and incubated for 15 minutes at room temperature. 50-100 µl of the fixed cell suspension were subsequently spread on superfrost microscope slides, left to dry at room temperature and incubated twice for 15 minutes with PBST (phosphate-buffered saline with 0.1% triton). Slides were then incubated for two hours in PBSTB blocking solution consisting of PBST, 1% bovine serum albumin (BSA), and 10% goat serum.

Staining of AipAgo1 was performed using anti-AipAgo1 (see above) and a conjugated goat-anti-rabbit DyLight Fluor 650 (Thermo Fisher # 84546) as secondary antibody.

First, slides were either incubated with primary antibody (dilution: 1:250 in PBSTB) or in PBSTB alone (negative control) in a humidified chamber at 4°C over night. Next, the secondary antibody incubation (dilution 1:400 in PBSTB)

was conducted for one hour in a humidified chamber at room temperature. Cells were subsequently also stained for actin using Dylight 488 Phalloidin (Thermo Fisher # 21833) and DNA using Hoechst 33342 (Thermo Fisher # H1399) for 30 minutes and 15 minutes, respectively. Between every staining incubation step, the slides were washed thrice for 15 minutes in PBST.

3.5.6 Crosslinking immunoprecipitation (CLIP) and RNA tag sequencing

A detailed protocol of the crosslinking immunoprecipitation of AipAgo1 is provided in the Supplemental Information at the end of this Chapter.

3.5.6.1 UV crosslinking

For crosslinking immunoprecipitation (CLIP) of AipAgo1, we used aposymbiotic adult anemones as well as adult anemones harboring exclusively *S. minutum* strain SSB01 symbionts with two replicates for each symbiotic state (see above). An overview of the CLIP experiment is shown in Fig. 3.4A. For each replicate, ~15 adult anemones (5-10 mm disc diameter) were sampled in ice-cold PBS and triturated with plastic mortar and pestle. The tissue samples were placed in 10 cm petri dishes on ice and irradiated three times at 254 nm with 400 mJ cm^2 using a Spectrolinker XL-1000 UV cross-linker (Spectronics) and subsequently collected by centrifugation. NP40 cell lysis buffer (Thermo Fisher # FNN0021) supplemented with 1 mM PMSF and protease inhibitor (Sigma Aldrich # P2714) was used for cell lysis following the manufacturers' instructions. Briefly, packed

tissue samples were re-suspended in three volumes of cell lysis buffer, disrupted using a Wheaton glass homogenizer and incubated for 30 min on ice. The homogenate was subsequently centrifuged at 13,000 rpm for 10 min at 4°C and the supernatant protein lysates were stored at -80°C until further processing. Final protein concentrations were ~3 µg/µl protein lysate as determined by Bradford assay.

3.5.6.2 Immunoprecipitation (IP)

The Dynabeads Protein G Immunoprecipitation Kit (Thermo Fisher #100007D) was used for the IP of crosslinked RNA-AipAgo1 complexes from the protein lysates (see above). 20 µg of rabbit anti-AipAgo1 Ab were bound to 100 µl Protein G coated Dynabeads for each sample according to the manufacturer's protocol. 700 µl of protein lysate were incubated with the Ab-coupled Dynabeads for 1h at 4°C by end over end rotation followed by two washes of the beads using 400 µl of washing buffer provided with the kit for each wash. mRNA bound to AipAgo1 complexes were trimmed on the beads for 6 min using 0.4 U RNase A/T1 (Thermo Fisher # EN0551) diluted in 400 µl washing buffer, immediately followed by one additional wash of the beads using 400 µl washing buffer to remove residual RNase A/T1. The RNA-AipAgo1 complexes were eluted from the Dynabeads using the elution provided with the kit and supplemented with 4x NuPAGE LDS sample buffer (Thermo Fisher # NP0008) and 10x NuPAGE reducing agent (Thermo Fisher NP0004). Beads diluted in pre-mixed elution

solution were heated for 10 min at 70°C and the supernatant was collected for SDS-PAGE size selection (see below).

As a first positive control of successful RNA co-immunoprecipitation, RNA tags were directly isolated from the beads after the last wash. The RNA tags were isolated from the beads using TRIzol LS (Thermo Fisher # 10296010) and the mirVana small RNA enrichment kit (Thermo Fisher # AM1560) and visualized on a Bioanalyzer 2100 (Agilent Technologies, RNA Pico Chip) (Fig. 3.4C).

As negative control, the IP was also performed using an unspecific IgG antibody and RNA tags that were potentially co-purified with the unspecific IgG antibody were also directly isolated from the beads using TRIzol LS (Thermo Fisher # 10296010), the mirVana small RNA enrichment kit (Thermo Fisher # AM1560), and visualized on a Bioanalyzer 2100 (Agilent Technologies, RNA Pico Chip) (Fig. 3.4D).

3.5.6.3 SDS-PAGE and nitrocellulose transfer

To remove potential co-purified contaminants that survived the immunoprecipitation and washes, the immunoprecipitate was subjected to SDS-PAGE gel electrophoresis. 20 µl of immunoprecipitate from the previous step were loaded per lane on a Novex NuPAGE Bis-Tris 4-12% mini gel (Thermo Fischer # NP0321BOX) and the proteins were resolved by size for 1 h at 150 V using NuPAGE MOPS running buffer Thermo Fisher NP0001) in a X-Cell SureLock Mini-Cell (Thermo Fisher # EI0002).

To additionally ensure the removal of free RNA that was potentially running at the same size as the RNA-AipAgo1 complexes, the proteins were transferred for 2 h at 100 V to a nitrocellulose membrane (Sigma Aldrich # GE10600002) using 10% (v/v) methanol NuPAGE transfer buffer (Thermo Fisher # NP0006) in a 2-gel Transfer and Blotting Module (Bio-Rad # 1660828EDU).

3.5.6.4 RNA tag isolation and protein digestion

RNA tags were isolated from a region of the nitrocellulose membrane corresponding to the predicted molecular weight of the AipAgo1 protein (~96 kDa) as well as a region ~50 kDa above to include larger RNA-AipAgo1 complexes running at higher molecular weights (Fig. 3.4B). The membrane regions were excised and cut into 1- to 2-mm squares and collected in RNase-free microcentrifuge tubes. The AipAgo1 protein was digested by mixing the cut membrane pieces with 400 µl 1X TE buffer supplemented with 0.5% SDS and 150 µg proteinase K (Quiagen # 19131) for 2 h at 55°C in a Thermoblock at 1000 rpm. TRIzol LS and the mirVana kit were used for RNA isolation from the supernatant following the manufacturers' instructions. The RNA tags were subsequently visualized using a Bioanalyzer 2100 (Agilent Technologies, RNA Pico Chip) which also allowed to assess the trimming of the protein-bound RNA to an anticipated RNA tag length of ~50-100nt from the earlier RNase treatment.

3.5.6.5 RNA tag sequencing

The miRNA and mRNA tags were first dephosphorylated using thermosensitive alkaline phosphatase (NEB # M0371) and subsequently treated with T4 polynucleotide kinase to specifically phosphorylate the 5’ ends of the RNA as preparation for specific RNA sequencing adapter ligation. The sequencing libraries were then prepared using the TruSeq small RNA library prep kit (Illumina # RS-200-0012) following the manufacturer’s instruction with 24 cycles of final library PCR amplification. The resulting cDNA library was sequenced single-end (1x 101 bp) on one lane of the Illumina HiSeq 2000 platform. The sequencing libraries are deposited at the NCBI SRA (accession no. SRX1351926)

3.5.7 CLIP tag analysis

The RNA reads were first trimmed for adapter sequences and filtered for low-quality sequences using Trimmomatic (33). For the detection of miRNA in the CLIP RNA libraries, RNA tags were aligned to the pre-miRNA sequences using NovoAlign (http://www.novocraft.com), requiring a minimum alignment of 15 nucleotides and allowing soft-clipping of the RNA read ends. The number of CLIP tags mapping to each pre-miRNA were then correlated between the replicates of each biological sample (Figs. 3.5B,C). For this correlation of miRNA counts, the raw counts in each sample were normalized by the total number of CLIP tags in the respective library, resulting in equal and comparable numbers of total CLIP tags of each sample.

For the CLIP analysis, the RNA tags were aligned to the whole genome using novoalign (http://www.novocraft.com), requiring a minimum alignment of 20 nucleotides, allowing a read to match no more than 3 times in the genome and enabling soft-clipping of the RNA read ends. The subsequent CLIP tag filtering and analysis were performed following the scripts and pipeline described in (21) (http://zhanglab.c2b2.columbia.edu/index.php/CIMS). Briefly, the CLIP tag alignments were first converted into the BED file format for convenient handling of all subsequent analysis steps and putative PCR duplicates where then collapsed based on the alignment coordinates to create a set of unique CLIP tags in each of the four samples (Table 3.2). Overlapping CLIP tags (i.e., peaks) and their respective height (i.e., number of reads per peak) were subsequently identified, requiring a minimum overlap of CLIP tags of 1 nucleotide within a peak. The bedtools 'intersect' tool (42) was then used to identify genomic regions that feature CLIP peaks across all samples. CLIP tag counts were normalized similar to the CLIP miRNA counts (see above) and the normalized height of CLIP peaks containing ≥ 3 CLIP tags at the same genomic region in each sample were correlated (Fig. 3.6A).

Pyicoclip (24) was used to identify significant CLIP peaks within exonic regions of the *Aiptasia* genome, using the pooled unique CLIP tags of all four samples. Pyicoclip implements the modified FDR algorithm proposed by (23) and was run with 1,000 rounds of permutation. Briefly, pyicoclip calculates the background frequency of detected CLIP peaks by randomly placing the identical number of CLIP tags into exonic regions and calculating the probability of obtaining the

same CLIP peak height as from the real data. Only exonic CLIP peaks with a FDR value ≤ 0.01 were retained for further analyses, resulting in a total of 2,269 CLIP peak regions. The number of CLIP peaks falling in each exonic region (i.e., 5' UTR, CDS, 3' UTR) was obtained from the genome feature file (gff) of the *Aiptasia* genome assembly. The CLIP peak sequences plus regions of 15 nt upstream and downstream were then extracted from the *Aiptasia* genome using bedtools 'getfasta' (42). Next, RNAhybrid (43) was used to map the 46 *bona fide* miRNAs (see above) to the CLIP peak sequences requiring 1) no *a priori* helix constraints, 2) a minimum free binding energy (MFE) ≤ -15 kcal/mol, 3) allowing GU wobbles in the miRNA-target alignment, and 4) allowing a maximum of 1 nt bulges in either side of the alignment. If a CLIP peak region was targeted by multiple miRNAs, we only kept the interaction that featured the highest number of binding nucleotides between the miRNA and the CLIP peak region. This resulted in a total of 3,377 miRNA-mRNA interactions, with 619 CLIP peak sequences featuring more than one equally good miRNA-mRNA interaction. A summarizing workflow of the CLIP tag and miRNA-mRNA interaction analysis is shown in Fig. 3.5A.

3.5.8 Differential expression analysis of miRNAs

To assess expression levels of miRNAs, we mapped each pre-processed small RNA library independently to the identified miRNA precursor sequences using the script quantifier.pl from the miRDeep2 package (34). The miRNA read counts were then normalized to the total number of small RNA reads mapping to any of

the 46 *bona fide* pre-miRNA sequences in each individual sample for cross-sample comparison. The differential expression analysis was performed with three replicates of each time point (i.e. infection state) using the edgeR package (44) implemented in R (45). For visualization of miRNA expression levels, Z-scores of significantly differentially expressed miRNAs (FDR ≤ 0.01) were calculated over all replicates/infection states and visualized in MeV (46). Multidimensional scaling (MDS) plots were created in edgeR using the "plotMDS" function. Distances between pairs of RNA samples correspond to the leading log2-fold-changes [i.e. average (root-mean-square) of the largest absolute log2-fold-change] (44, 45).

All data reported in this chapter are accessible at the NCBI BioProject database (accession no. PRJNA297831).

3.7 Tables and Figures

Table 3.1. Overview of small RNA sequencing libraries

Sample[+]	NCBI SRA accession	Library Depth	Cleaned smRNA counts (15-32 nt)*	miRNA counts
A1	SRR2716067	11,597,080	7,571,784	1,897,191
A2	SRR2716068	11,624,496	8,858,912	1,285,097
A3	SRR2716069	10,616,536	7,400,484	929,482
A4	SRR2716070	13,608,809	8,811,193	771,261
I1	SRR2716071	13,072,948	10,282,811	629,133
I2	SRR2716072	11,373,666	8,415,424	281,079
I3	SRR2716073	13,215,508	9,618,468	474,734
I4	SRR2716074	12,827,426	9,376,063	492,852
S1	SRR2716063	13,083,889	10,794,704	1,121,781
S2	SRR2716064	13,921,704	10,813,917	680,290
S3	SRR2716065	10,170,186	6,870,577	498,689
S4	SRR2716066	13,391,941	10,619,254	543,512

[+] A: library from aposymbiotic anemones; I: Intermediate infected; S: symbiotic.

* The sum of these reads (n = 109,433,591) is comprised of 13,422,387 unique small RNA sequences and was used as input for the miRNA annotation by miRDeep2.

Table 3.2. Overview of the CLIP tag mapping and filtering

Sample	A1	A2	S1	S2
NCBI SRA accession	SRR2716058	SRR2716059	SRR2716060	SRR27160601
Number of CLIP tag alignments (≥ 20 nt)	158,125	195,927	261,767	343,111
Number of unique CLIP tags (≥ 20 nt)	52,747	60,156	88,264	107,633
Genomic position with ≥ 1 CLIP tag present in both replicates	14,979		25,199	
Genomic position with ≥ 1 CLIP tags present in all samples	10,061			

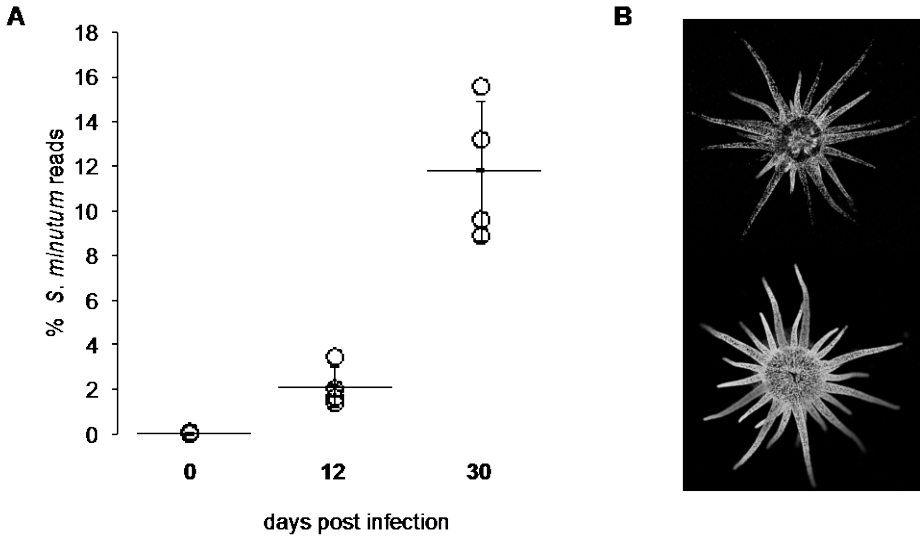

Fig. 3.1. Infection of *Aiptasia* anemones with *S. minutum* (strain SSB01) dinoflagellates. (*A*) Increase of the intracellular *S. minutum* population in *Aiptasia* hosts shown as the percentage of mRNA reads mapping to a reference transcriptome of *S. minutum* (see Materials and Methods). The white circles designate the biological replicates (n = 4) for each time point (i.e., infection state) at 0 days, 12 days p.i. and 30 days p.i. Black lines represent the mean. (*B*) The increasing density of endosymbionts is visualized by the chlorophyll fluorescence of *Symbiodinium* residing within *Aiptasia* host cells at the intermediate [12 days p.i., top] and fully symbiotic state (30 days p.i., bottom).

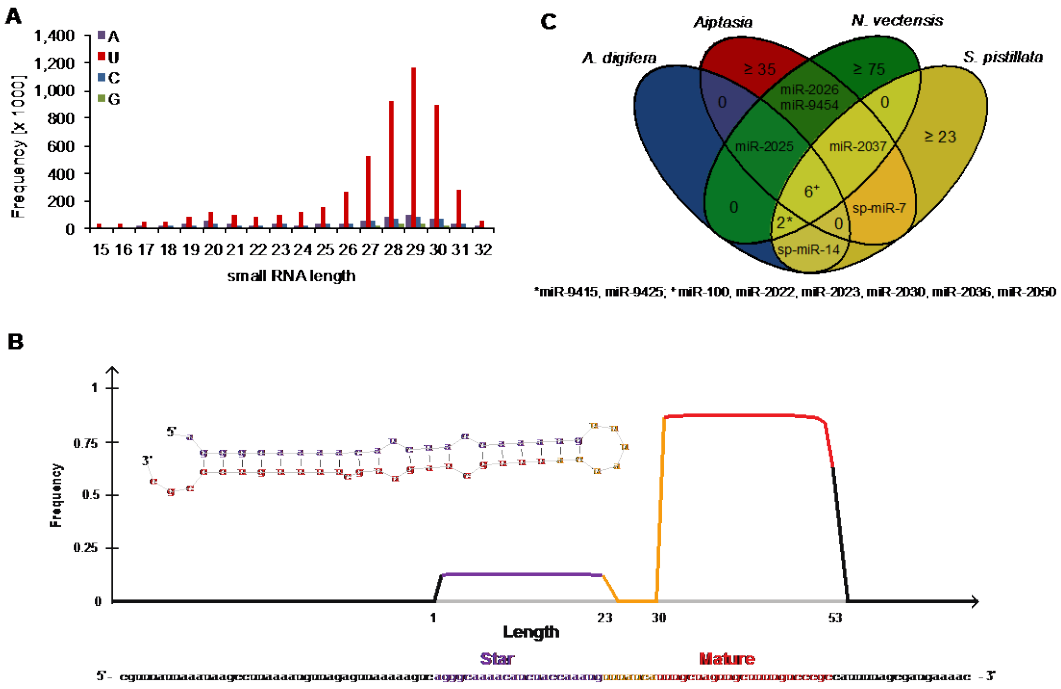

Fig. 3.2. Annotation of the *Aiptasia* miRNA repertoire. (*A*) Two populations of unique small RNA reads are overrepresented in the *Aiptasia* genome at lengths of 19-21 nt and 27 - 31 nt, representing putative siRNA/miRNA and piRNA, respectively. The small RNA reads are sorted by their 5' identity. (*B*) Overlap of evolutionary conserved mature miRNAs. Only six miRNAs are conserved among all anthozoan species and the largest conserved miRNA repertoire is found between the two actinarian *Aiptasia* and *N. vectensis.* More species-specific miRNAs might be annotated with the increasing completeness of genome assemblies and small RNA sequencing. (*C*) Sequence (bottom) and secondary structure of the miRNA-2022 precursor. The frequencies of small RNAs that map to either side of the hairpin (ordinate) is shown along the primary sequence (abscissa). Small RNAs matching the mature miRNA sequence (red) feature

homogenous 5’ ends and considerably higher read counts than the star-miRNA (purple).

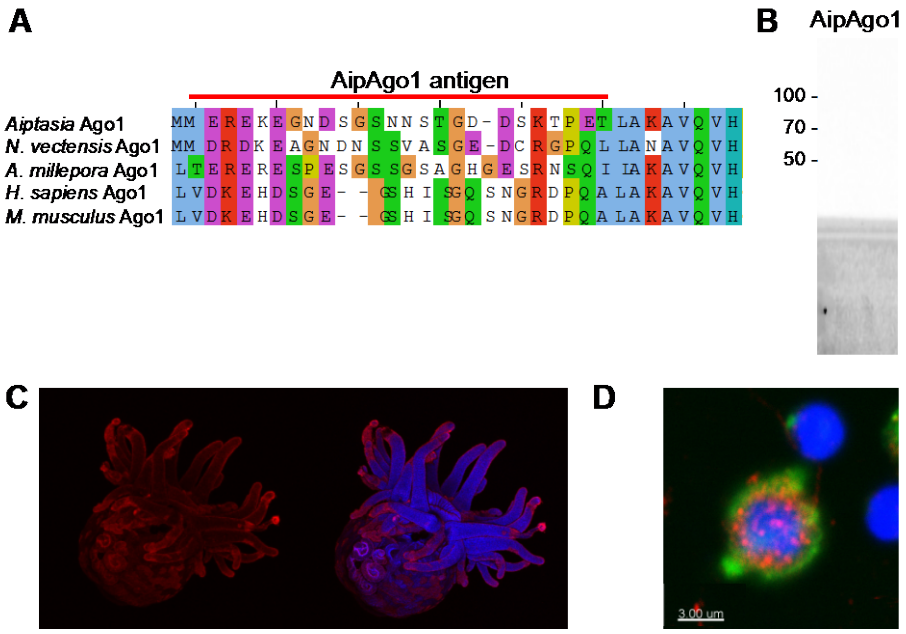

Fig. 3.3. Characterization of AipAgo1. (*A*) Alignment of the C-terminal ends of the canonical *Aiptasia, N. vectensis, A. millepora* Argonaute proteins as well as human and murine Argonaute 1. The red bar above designates the *Aiptasia* peptide sequence used as antigen for antibody generation. (*B*) Western Blot of AipAgo1 at the expected molecular weight of ~96 kDa. (*C*) Left: AipAgo1 is present throughout *Aiptasia* polyps as assessed by whole mount immunohistochemical stainings of AipAgo1. Right: Overlay of AipAgo1 (red) and Hoechst nuclei staining (blue). (*D*) Subcellular localization of AipAgo1 (red) in the cytoplasm showing a tendency of the protein to accumulate in specific locations that presumably reflect P-bodies. DNA is stained with Hoechst nuclei stain (blue) and the actin cytoskeleton with phalloidin (green, see Materials and Methods).

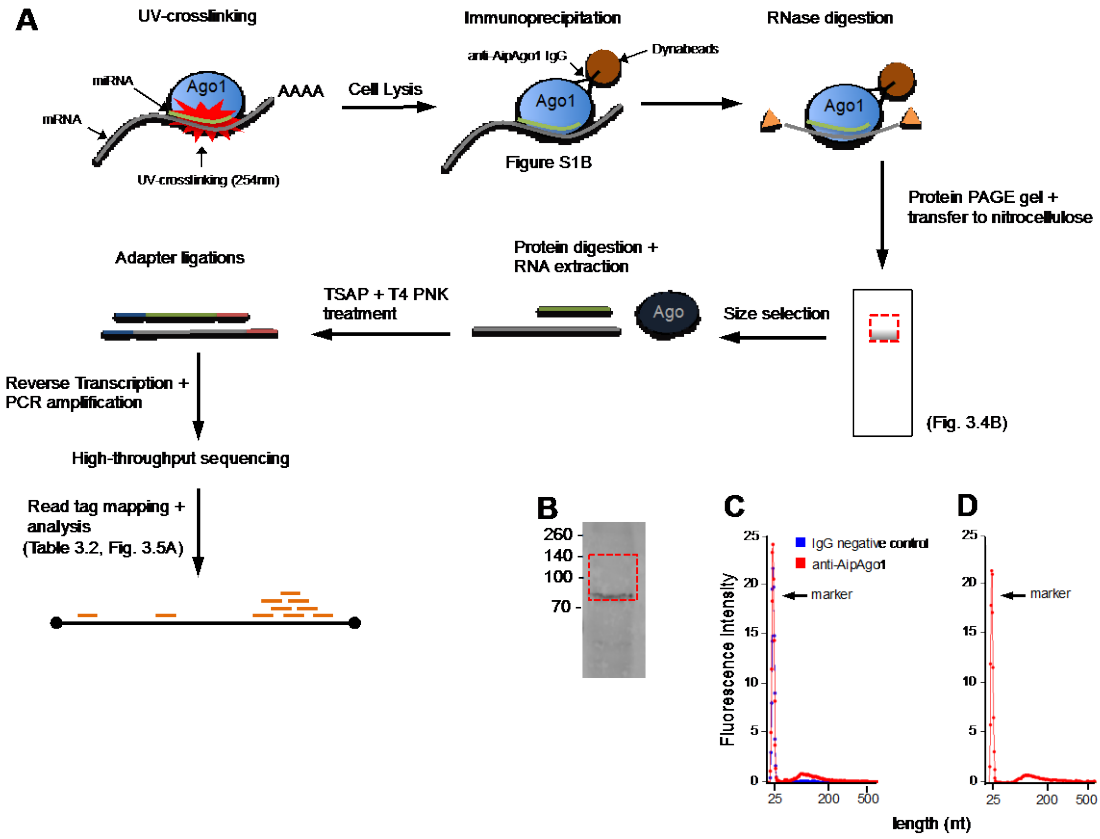

Fig. 3.4. Workflow of the AipAgo1 CLIP experiment. (*A*) The tripartite miRNA-mRNA-Ago complex is crosslinked with UV light *in vivo* following cell lysis and immunoprecipitation of AipAgo1 and the bound RNA. Next, mRNA is digested to using RNase A/T1 preserving only tags protected by AipAgo1 around the miRNA binding site. The protein-RNA complex is then resolved by molecular size and plotted on nitrocellulose. Membrane regions corresponding to the molecular weight of the AipAgo1 protein (~96 kDa) were excised followed by the digestion of AipAgo1, RNA extraction, sequencing and analysis (*B*) Western Blot of AipAgo1. The dashed red-box designates the membrane area that was excised

for protein and RNA extraction during the CLIP experiment. (*C*) Electropherogram of RNA tags extracted after immunoprecipitation. The red line shows the length distribution of RNA tags that were extracted from the immunoprecipitation using anti-AipAgo1 antibodies. No RNA tags are detected when a control IgG antibody is used for the immunoprecipitation. (*D*) Same as (*C*), but showing the length distribution of RNA tags that are extracted from the nitrocellulose membrane after the complete CLIP experiment. The RNA tags where then used as input for RNA library generation.

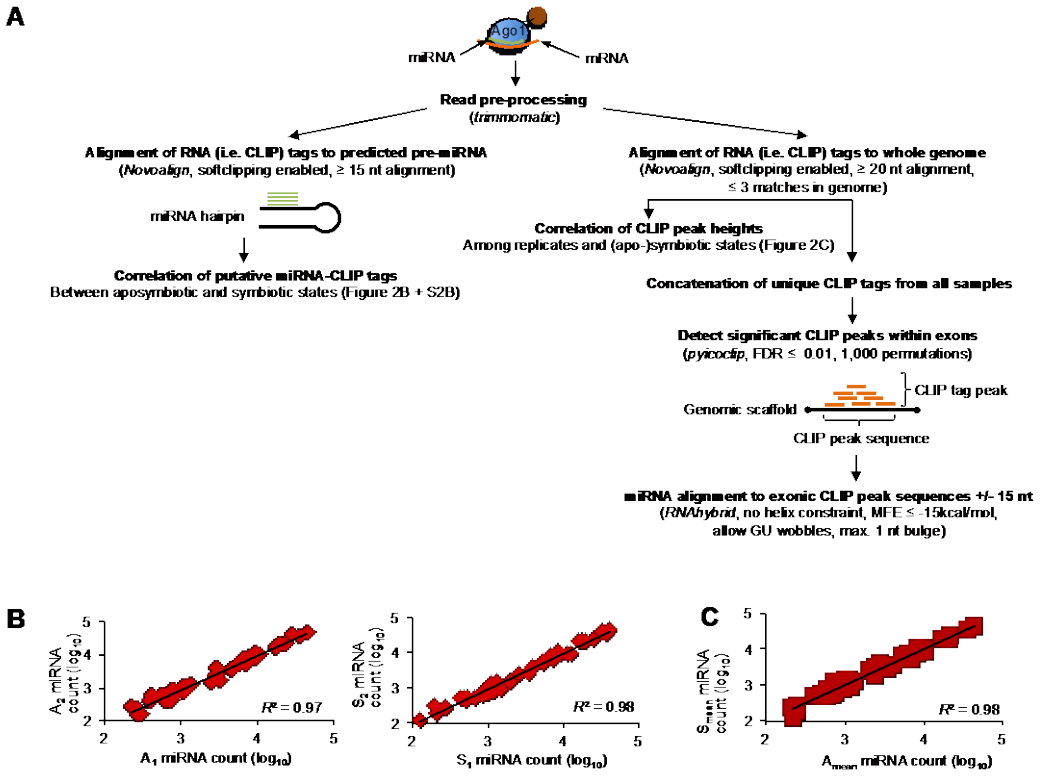

Fig. 3.5. CLIP tag processing workflow. (*A*) Workflow of the CLIP tag analysis (see Materials and Methods). (*B*) Correlation of miRNA read counts that were recovered from the independent CLIP experiments of the two aposymbiotic (left) and symbiotic (right) replicates. (*C*) Correlation of the average miRNA counts recovered from the two aposymbiotic and symbiotic samples used in the CLIP experiment.

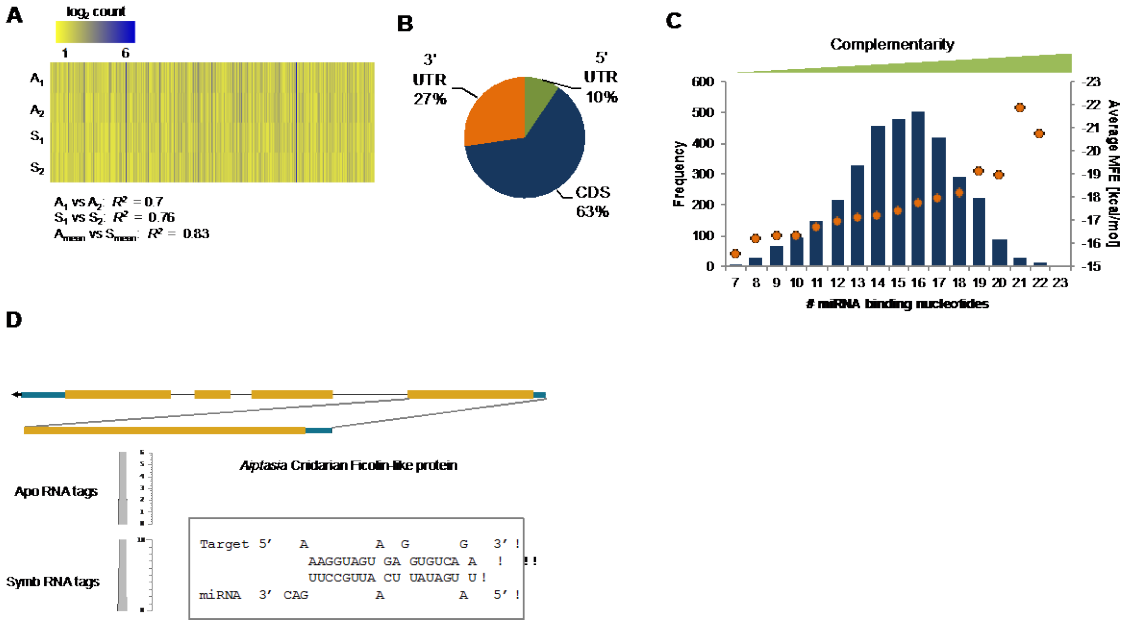

Fig. 3.6. Analysis of ternary miRNA-mRNA-Ago interactions. (*A*) Correlation of CLIP peak heights (≥ 3 CLIP tags) that are recovered in all four independent replicates of the CLIP experiment at the same genomic positions. The correlation coefficients are shown below. (*B*) Percentages of miRNA-mRNA interactions identified within exonic regions of the *Aiptasia* genome. UTR: untranslated regions; CDS: coding sequence. (*C*) Frequencies of miRNA-mRNA interactions featuring a specific number of miRNA nucleotides involved in the mRNA target binding (blue bars, primary axis) and the average minimum free energy for each of the miRNA-mRNA interaction bins (orange diamonds, secondary axis). (*D*) Genome browser view of an *Aiptasia* Cnidarian Ficolin-like protein (CniFL) gene with CLIP tags aligning in the coding sequence of exon 1 (grey bar: CLIP tag coverage). The CLIP tags of the pooled aposymbiotic and symbiotic samples

map to the same genetic region that putatively represents a target site of spi-miR-7 (box). Blue: untranslated region; yellow: coding sequence; arrow: direction of transcription.

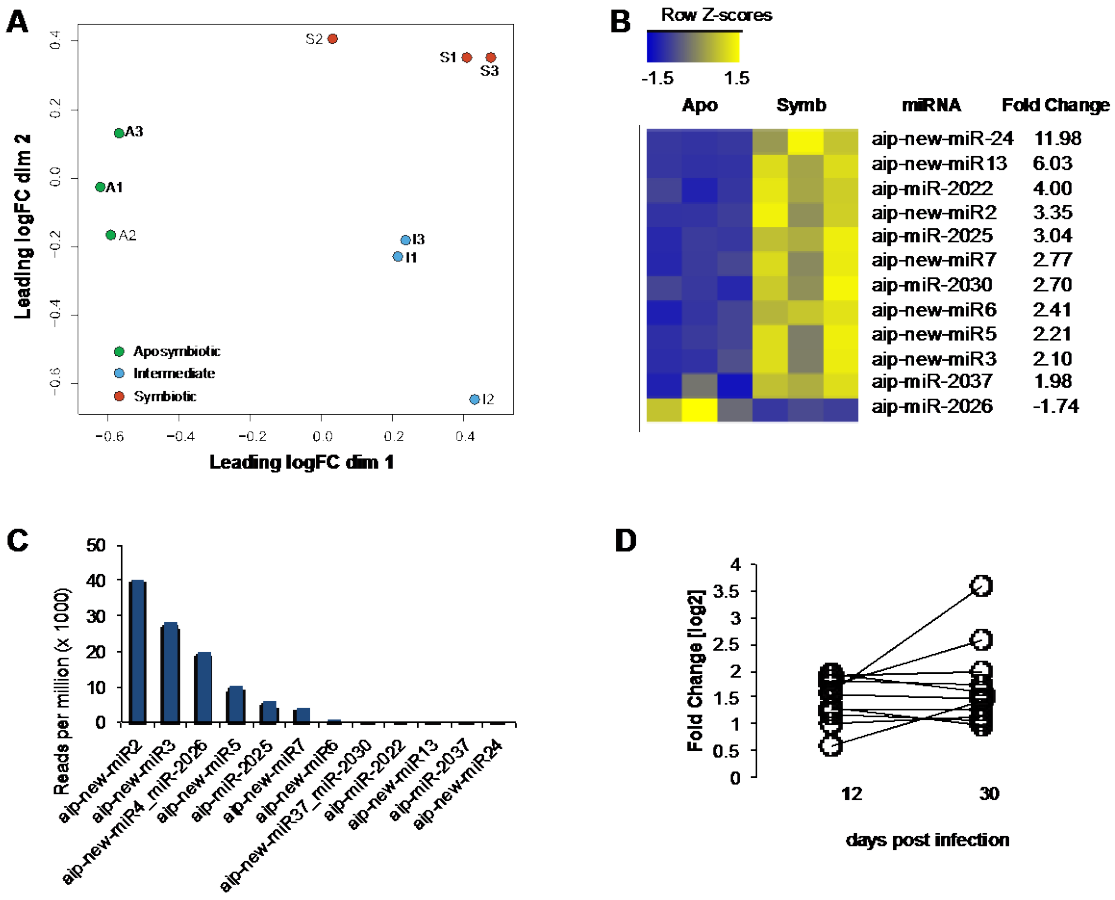

Fig. 3.7. Differential expression analysis of *Aiptasia* miRNA. (*A*) MDS plot of miRNA expression at different infection states showing the replicates (n = 3, indicated by same color) of each infection state cluster together and separate clearly along the symbiotic (horizontal) axis, indicating overall changes in expression levels between the different states. (*B*) The heatmap of significantly differentially expressed miRNAs (FDR ≤ 0.01) shows the majority of miRNAs are upregulated in the symbiotic state compared to *Aiptasia* without endosymbionts. (*C*) Average counts (reads per million, RPM) of the 12 differentially expressed miRNA calculated over all replicates. (*D*) Fold changes (log2) of up-regulated

miRNAs are at similar intensities at the intermediate infected state (12 days p.i.) and the fully symbiotic state (30 days p.i.) when compared to aposymbiotic *Aiptasia* anemones.

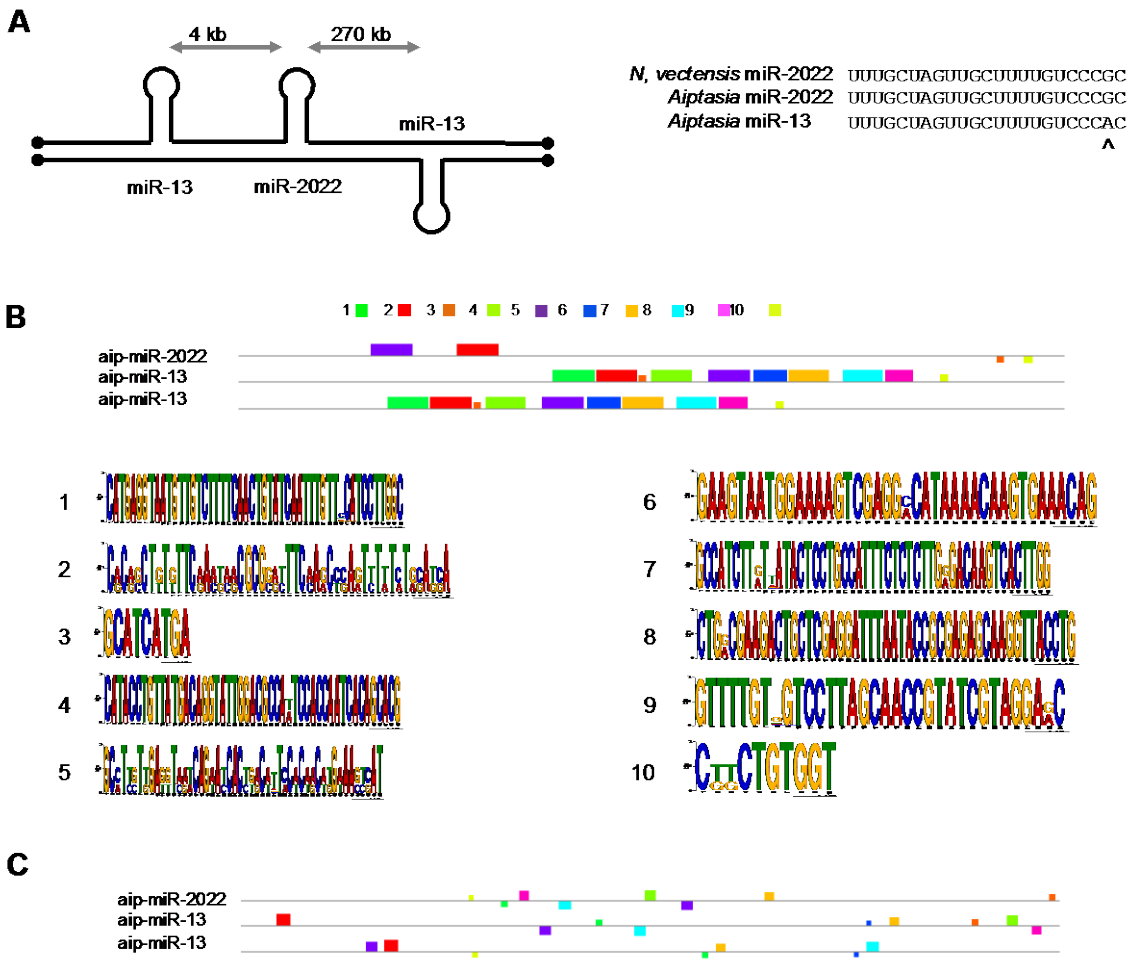

Fig. 3.8. Characterization of the miR-2022/13 locus. (*A*) Genomic position of the miR-2022 and miR-13 loci in the *Aiptasia* genome (left). The arrows indicate the distance in kilobases (kb) between the predicted hairpin precursors. miR-2022 is also found in other anthozoan genomes, e.g. *N. vectensis* (right), whereas miR-13 is *Aiptasia* specific and seems to have duplicated from the miR-2022 locus and differs in one nucleotide position in the 3' region (indicated by ^). (*B*) Location of conserved sequence motifs detected in a region 1 kb upstream of the miR-2022 and miR-13 loci (top). The two miR-13 loci are highly conserved, but show no conserved sequence similarity with the miR-2022 upstream region. (*C)*

Motifs in the 1 kb downstream region of the three loci are considerably shorter than in the upstream region and are not conserved in their genomic position.

A

miRNA	miRNA FC	Gene	Gene FC
aip-miR5			
aip-miR3			
aip-miR-2025			
aip-miR-5			

Fig. 3.9. Symbiotically regulated miRNAs target genes presumably involved in the endosymbiosis onset and maintenance. (*A*) Model of genes and pathways that are putatively regulated by differentially expressed miRNAs and which are functioning in biological processes previously recognized to be involved in the endosymbiosis onset and maintenance. (*B*) Overview of miRNA expression levels and the expression of their cognate target genes that were identified in the CLIP experiment. FC: Fold Change (symbiotic relative to the aposymbiotic state).

3.7 Supplemental Information

Supplemental Information (Dataset S3.1 and S3.2) are archived and accessible at the KAUST library.

Protocol for infection of aposymbiotic *Aiptasia* polyps with *Symbiodinium minutum* strain SSB01

Water: Autoclaved, natural seawater; filtered through a 0.22 µm glass-fiber filter

Incubator settings: Temperature 25°C; light-dark cycle: 12h:12h

Experimental setup: 5-10 aposymbiotic anemones per well of a 6-well plate in a total of ~10 ml water. During sampling, the anemones of every 6-well plate are pooled as one replicate.

Day -16: Put *Aiptasia* from the dark into ambient light conditions (12h:12h light: dark cycle)

Day -12: Setup new *Symbiodinium minutum* SSB01 culture in 150 ml F2-Si medium

Day -7: last feeding with *Arteimia* nauplii larvae

Day -6: water change

Day -3: water change

Day 0: Sampling t_0, add *S. minutum* SSB01 at a final concentration of ~ 10^5 cells/ml

Day 1: feeding

Day 2: water change, reinfect with *S. minutum* SSB01 at a final concentration of ~ 10^5 cells/ml

Day 10: Stop *S. minutum* SSB01 exposure. Transfer anemones in new 6-well plates

Day 12: Sampling t_{12}, feeding

Day 13: water change

Day 17: feeding

Day 18: water change

Day 23: feeding

Day 24: water change

Day 30: Sampling t_{30}

Protocol for Crosslinking immunoprecipitation and high-throughput sequencing (HITS-CLIP) of the *Aiptasia* Argonaute protein

This protocol was adapted from the HITS-CLIP protocol developed by Moore *et al.* (21). It also includes information from the CLASH protocol from Helwak *et al.* (47) and several manufacturer protocols as outlined below.

Equipment

- UV cross-linker (254 nm; e.g. Stratagene Stratalinker 2400)
- XCell SureLock Mini-Cell (Life Technologies)
- Mini-Trans-Blot Transfer cell (Bio-Rad)
- Dynal Magnetic Rack (Life Technologies)
- 1.5 ml Wheaton glass homogenizer
- 1.5 ml RNase free microcentrifuge tubes
- 0.2 ml PCR tubes
- 15 ml and 50 ml Falcon tubes
- Tissue culture dish (10 and 15 cm)
- refrigerated microcentrifuge
- Thermoblock
- SpeedVac

Reagents

- PBS

- Nuclease-free water
- 0.3 M PMSF dissolved in DMSO
- EDTA
- Methanol
- 1 M Tris stock solution
- 1xTE buffer + 0.5 vol% SDS
- NP40 Cell Lysis buffer (Life Technologies # FNN0021)
- Protease inhibitor cocktail (e.g. Sigma # P2714)
- 10X PhosSTOP phosphatase inhibitor (e.g. Sigma-Aldrich # 04906845001)
- Dynabeads Immunoprecipitation Kit (Life Technologies # 10007D)
- anti-AipAgo antibody (custom, e.g. from Acris Antibodies)
- RNase A/T1 (Life Technologies # EN0551)
- 4x NuPAGE LDS sample buffer (Life Technologies # NP0007/8)
- 10x NuPAGE reducing agent (Life Technologies # NP0009)
- 4-12% Novex NuPAGE Bis-Tris mini gel (Life Technologies # NP0321BOX)
- 20x NuPAGE MOPS running buffer (Life Technologies # NP0001)
- 20x NuPAGE transfer buffer (Life Technologies # NP0006)
- Whatman Protran BA85 nitrocellulose (Sigma # Z670898)
- Proteinase K (Qiagen # 19131)
- TRIzol LS reagent + chloroform (Life Technologies # 10296-028)

- thermosensitive Alkalinephosphatase (NEB # M0371, includes cutSMART buffer)
- T4 Polynucleotide Kinase (Life Technologies # 18004-010)
- 10 mM ATP (e.g. Life Technologies # PV3227)
- RNase Inhibitor (e.g. Life Technologies # N8080119)
- mirVana small RNA enrichment kit (Life Technologies # AM1560)
- Illumina small RNA Library Prep Kit (Illumina # RS-200-0012)
- Ultrapure agarose (e.g. Life Technologies # 16500-100)
- 100 bp DNA ladder (e.g. NEB # N3231S)
- 6X Gel loading dye (e.g. NEB # B7024S)
- Broad Range Protein Ladder (Life Technologies # 26634)

1. Sampling and cross-linking (see (21)

1| Use a plastic mortar and pestle to triturate ~15 anemones with ~0.5 - 1 cm disk diameter to create a gross tissue suspension in ice-cold PBS. Transfer the tissue suspension to a 10-cm tissue culture dish and place it on ice.

2| Irradiate the triturated tissue three times at 400 mJ cm^2 in the UV cross-linker, by swirling between each irradiation to keep it cold and to maximize exposed surfaces for cross-linking. The Stratalinker or Spectrolinker cross-linkers have UV detectors that monitor the actual dose delivered. The units are labeled such that 1 = 0.1 J per m^2; hence, a setting of 4,000 on the machine is 400 mJ cm^2

3| Collect the cells into a 15 or 50 ml conical tube and pellet them by centrifugation at 200g for 5 min at 4°C. Remove the supernatant, resuspend the cell pellet in 1 ml of cold PBS and transfer it to a microcentrifuge tube. Re-pellet the cells at ~1,000g for 5 min at 4 °C in the microcentrifuge (each tube should have a maximum of 200 - 300 µl of packed cells or tissue).

2. Cell lysis

1| prepare a 1M Tris stock solution

30 ml 1M Tris 3,6342 g (adjust pH with HCL to 7.4)

2| Prepare NP40 stock solution (100 ml)

250mM NaCl 1.461 g

5mM EDTA 0.1461 g

50 mM Tris 5 ml of 1M Tris stock solution

95 ml H_2O

1 ml NP40

Prepare lysis buffer

1| 2.5 ml of cell lysis buffer must be supplemented prior to use with:

- 1 mM PMSF (i.e., 8.5 µl 0.3 M stock per 2.5 ml cell lysis buffer).

- 0.3 M stock: add 0.0523 g PMSF in 1ml DMSO

- 25 µl of proteinase inhibitor cocktail
- 25 µl EDTA
- 250 µl 10X PhosSTOP solution

2| For cross- linked tissue, suspend cell pellets in a volume of lysis buffer roughly three times the volume of packed tissue. The tissue might be resistant to lysis, so gentle mechanical disruption with a Wheaton glass homogenizer can be applied.

3| Lyse the cell pellet in cell lysis buffer for 30 minutes, on ice, with flicking at 10 minute intervals.

4| Transfer the extract to microcentrifuge tubes and centrifuge at 13,000 rpm for 10 minutes at 4°C.

5| Aliquot the clear lysate to clean microfuge tubes. These samples are ready for assay. Lysates can be stored at -80°C. Avoid multiple freeze/thaws.

3. Immunoprecipitation

See also the protocol for Dynabeads Immunoprecipitation Kit Protein G for steps 3.1 to 3.4.

3.1. Bead preparation

NOTE: Loading capacity of Ab and Dynabeads: 50 µl beads can hold up 10 µg Ab. For the final experiment, use 700 µl lysate (~ 3 µg/µl protein concentration) per sample and divide it into 2x 350 µl lysate. The samples can be run as pairs with 50 µl beads and 10 µg Ab per 350 µl lysate according to the manufacturer's protocol. The pairs are then pooled again after the last wash before the protein elution. Moore et al. recommend 50 µl of beads as minimum input to avoid loss during washes.

1| Resuspend Dynabeads® in the vial (vortex >30 sec or tilt and rotate 5 min).

2| Transfer 50 µl (1.5 mg) Dynabeads to a 1.5 ml microcentrifuge tube.

3| Place the tube on the magnet to separate the beads from the solution, and remove the supernatant.

4| Remove the tube from the magnet.

5| Add your antibody (Ab) (typically 1 - 10 µg) diluted in 200 µl of Binding-Washing Buffer to the tube from step 4 above. The optimal amount of Ab needed depends upon the individual Ab used (I used 10ug).

6| Incubate with rotation for 30 min at room temperature.

7| Place the tube on the magnet and remove the supernatant.

8| Remove the tube from the magnet and resuspend the beads-Ab complex in 200 Ab Binding - Washing Buffer. Wash by gentle pipetting.

For storage of Ab-conjugated Dynabeads, use PBS (pH 7.4) with 0.01–0.1% Tween-20 to prevent aggregation.

3.2. Immunoprecipitation

1| Place the tube containing Dynabeads-Ab complex on the magnet and remove the supernatant.

2| Add 350 µl of your sample containing the antigen (Ag) (typically 100 - 1,000 µl) and gently pipette to resuspend the Dynabeads-Ab complex (-> 350 µl)

Reminder: One sample = 700 µl, but run as pairs of 2x 350 µl.

3| Rotate the beads/lysate mix end over end for 1 h at 4°C.

4| Place the tube on the magnet.

5| Wash the Dynabeads-Ab-antigen complex 2 times using 200 µl washing buffer for each wash. Separate beads and supernatant on the magnet between each wash, remove supernatant, and resuspend by gentle pipetting.

3.3. On bead RNase digestion (see also Helwak *et al.* (47))

1| Dilute RNase A1/T 1:50 in water (-> 0.2U/µl). Can be stored at 4 °C for at least 2 years.

2| Start RNase digestion by adding 1 µl (0.2U) of diluted RNase A/T1 per 200 µl IP washing buffer (same concentration as in Helwak paper [0.5U into 500 µl buffer]).

3| Incubate for 6 min at 20°C. During the digestion, tap the samples gently every 30 seconds to prevent Dynabeads from settling.

4| Remove the RNase digestion buffer completely and discard it.

5| Wash the Dynabeads-Ab-antigen complex 1 time using 200 µl washing buffer to remove RNase completely. Separate on the magnet and remove supernatant.

6| Resuspend the Dynabeads-Ab-antigen complex in 100 µl washing buffer and transfer the bead suspension to a clean tube. This is recommended to avoid co-elution of proteins bound to the tube wall.

NOTE: Pool the pair of samples again into the same tube (results in 100 µg of beads in 200 µl of Washing Buffer per sample)

3.4. Elute Target Antigen

NOTE: The Sample Buffer is stored at 4°C and becomes very viscous. Bring it to room temperature for easier pipetting several minutes before use.

-> Denaturing elution

1| Place the tube containing Dynabeads-Ab-Ag complex on the magnet and remove the supernatant.

2| Add 39 µl Elution Buffer and 21 µl premixed NuPAGE LDS Sample Buffer/NuPAGE Sample Reducing Agent mix.

3| For 4 samples (each with 50 µl, i.e. two replicates with 100 µg beads each), prepare:

Sample (Volume Elution Buffer)	39 µl
NuPAGE LDS Sample Buffer (4x)	15 µl
NuPAGE Reducing Agent (10x)	6 µl
Total	60 µl

4| Gently pipette to resuspend the Dynabeads-Ab-Ag complex.

5| Heat for 10 min at 70°C on a Thermoblock with 1000 rpm.

6| Place the tube on the magnet and load the supernatant/sample onto a gel. 1 lane can hold 20 µl, so divide the 60 µl sample in 3 separate lanes.

4. Size Selection on PAGE Gel

See also Moore *et al.* (21)

1| Prepare 1x Running Buffer

20X NuPAGE MOPS	40 ml
Water	760 ml
Total	800 ml

2| Load the supernatants on a Novex NuPAGE Bis-Tris 8-12% mini gel, by dividing samples across two or more lanes if necessary (*sensu* Moore et al.: not more than 20 µg IgG per lane)

2| Run the gel at constant voltage (150 V) for ~1h (till the sample buffer reaches bottom of the gel)

5. Wet transfer on nitrocellulose membrane

1| Preparation of NuPAGE Transfer Buffer:

NuPAGE Transfer Buffer (20x)	50 ml
Methanol	100 ml
Water	850 ml
Total	1000 ml

2| Prepare the sandwich parts as follows:

Nitrocellulose - Place the membrane directly into a shallow dish containing 50 ml of 1X NuPAGE Transfer Buffer for several minutes.

.**Filter paper** - Soak the filter paper briefly in 1X NuPAGE Transfer Buffer immediately prior to use.

.**Gel** - Use the gel immediately following the run. **Do not soak the gel in transfer buffer.**

3| Transfer the gel to membrane using the BioRad Mini-Trans-Blot Transfer cell for 2 h at 100 V in 1X NuPAGE transfer buffer with 10% (vol/vol) methanol.

6. RNA tag isolation

See also Moore *et al.* (47) and Helwak *et al.* (21)

1| Identify the signal ~50 kDa above the identified weight of the protein of interest, using molecular weight markers as a guide. Use two scalpels to carefully dice each excised band (i.e. 1 lane) into 1- to 2-mm squares, and then transfer these to an RNase-free, 1.5-ml microentrifuge tube.

2| Repeat these steps for each sample (i.e. lane) to be processed, by changing scalpels in between.

Pause Point: The nitrocellulose membrane can be frozen at −20 °C for at least a week.

7. Protein digestion and RNA extraction

1| Prepare 0.5% SDS-TE buffer for 10 samples (4 ml):

100X TE buffer	40 µl
20% SDS	100 µl
Nuclease-free water	3,860 µl
Total	4,000 µl

2| Dilute proteinase K:

Proteinase K (20 µg/µl)	150 µg (7.5 µl)
1X 0.5% SDS-TE buffer	400 µl

3| Incubate the membrane slices for 2 h at 55 °C with 1000 rpm in the Thermoblock.

4| Discard the membrane and perform the RNA extractions using TriZol LS from the supernatant. Follow the manufacturers protocol by starting with 400 µl sample and add 1.2 ml of TriZol LS. Continue with 320 µl of chloroform for the separation of the aqueous phase.

5| Transfer the aqueous phase into fresh tubes and continue with the mirVana protocol for small RNA extraction and enrichment following the manufacturer's instructions.

8. Dephosphorylation of 5'/3' RNA ends

See also the protocol for Illumina directional mRNA library preparation

1| Dry down RNA from 100 µl to 16 µl in a SpeedVac (30°C)

2| Prepare and add :

CutSmart Buffer	2 µl
RNAse Inhibitor	1 µl
1U rSAP	1 µl
Total	20 µl

3| Incubate for 30 min at 37°C in the Thermoblock

4| Stop reaction at 65°C for 5 min in Thermoblock

9. PNK treatment

RNA	20 µl
5X forward PNK buffer	10 µl
ATP (10mM)	5 µl
T4-PNK	2 µl
RNase Inhibitor	1 µl
Water	12 µl

Total 50 µl

Incubate 60 min at 37°C, hold on 4°C

10. RNA cleanup

1| Follow the cleanup procedure of the mirVana small RNA extraction and enrichment protocol (*IV. Additional Procedures/ A. Isolation of Small RNAs from Total RNA Samples*)

11. RNA sequencing cDNA library gel size selection

1| Dry down the RNA from 100 µl to 5 µl in a SpeedVac. Proceed with cDNA library preparation using the TruSeq small RNA library prep kit following the manufacturer's protocol.

2| Cast a 2% agarose gel using µltrapure agarose and 1X TAE buffer with Sybr Safe.

3| Take 10 µl of the library and mix with 2 µl 6X Gel loading dye

4| Use 100 bp DNA ladder for size reference

5| Run the gel for ~1 h at 80V

6| Identify the signal under a UV transilluminator and cut a band in the size range 100 - 300 bp from the gel.

7| Use the Quiagen MinElute Kit for RNA gel extraction and cleanup following the manufacturer’s protocol.

3.8 References

1. Muscatine L, Porter JW (1977) Reef Corals: Mutualistic Symbioses Adapted to Nutrient-Poor Environments. *Bioscience* 27(7):454–460.
2. Trench RK (1987) Dinoflagellates in non-parasitic symbioses. *Biol Dinoflag*:530–570.
3. Lehnert EM, et al. (2014) Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3* 4(2):277–295.
4. Baumgarten S, et al. (2015) The genome of *Aiptasia*, a sea anemone model for coral symbiosis. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1513318112.
5. Ganot P, et al. (2011) Adaptations to endosymbiosis in a cnidarian-dinoflagellate association: differential gene expression and specific gene duplications. *PLoS Genet* 7(7):e1002187.
6. Voolstra CR, et al. (2009) The host transcriptome remains unaltered during the establishment of coral-algal symbioses. *Mol Ecol* 18(9):1823–1833.
7. Rodriguez-Lanetty M, Phillips WS, Weis VM (2006) Transcriptome analysis of a cnidarian-dinoflagellate mutualism reveals complex modulation of host gene expression. *BMC Genomics* 7:23.
8. Schnitzler CE, Weis VM (2010) Coral larvae exhibit few measurable transcriptional changes during the onset of coral-dinoflagellate endosymbiosis. *Mar Genomics* 3(2):107–116.
9. Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell* 136(4):642–655.
10. Esau C, et al. (2006) miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab* 3(2):87–98.
11. Ghosh J, Bose M, Roy S, Bhattacharyya SN (2013) *Leishmania donovani* targets Dicer1 to downregulate miR-122, lower serum cholesterol, and facilitate murine liver infection. *Cell Host Microbe* 13(3):277–288.

12. LaMonte G, et al. (2012) Translocation of sickle cell erythrocyte microRNAs into *Plasmodium falciparum* inhibits parasite translation and contributes to malaria resistance. *Cell Host Microbe* 12(2):187–199.
13. Grimson A, et al. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455(7217):1193–1197.
14. Juliano CE, et al. (2014) PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. *Proc Natl Acad Sci U S A* 111(1):337–342.
15. Moran Y, et al. (2014) Cnidarian microRNAs frequently regulate targets by cleavage. *Genome Res*. doi:10.1101/gr.162503.113.
16. Krishna S, et al. (2013) Deep sequencing reveals unique small RNA repertoire that is regulated during head regeneration in *Hydra magnipapillata. Nucleic Acids Res* 41(1):599–616.
17. Liew YJ, et al. (2014) Identification of microRNAs in the coral *Stylophora pistillata*. *PLoS One* 9(3).
18. Moran Y, Praher D, Fredman D, Technau U (2013) The Evolution of MicroRNA Pathway Protein Components in Cnidaria. *Mol Biol Evol* 30(12):2541–2552.
19. Leung AKL, Sharp PA (2013) Quantifying Argonaute proteins in and out of GW/P-bodies: implications in microRNA activities. *Adv Exp Med Biol* 768:165–82.
20. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460(7254):479–486.
21. Moore MJ, et al. (2014) Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* 9(2):263–293.
22. Weyn-Vanhentenryck SM, et al. (2014) HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep* 6(6):1139–52.

23. Yeo GW, et al. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* 16(2):130–137.
24. Althammer S, Gonzalez-Vallinas J, Ballare C, Beato M, Eyras E (2011) Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics* 27(24):3333–3340.
25. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136(2):215–233.
26. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10(2):94–108.
27. Detournay O, Schnitzler CE, Poole A, Weis VM (2012) Regulation of cnidarian-dinoflagellate mutualisms: Evidence that activation of a host TGFbeta innate immune pathway promotes tolerance of the symbiont. *Dev Comp Immunol* 38(4):525–537.
28. Barshis DJ, et al. (2013) Genomic basis for coral resilience to climate change. *Proc Natl Acad Sci U S A* 110(4):1387–1392.
29. Chen MC, Cheng YM, Hong MC, Fang LS (2004) Molecular cloning of Rab5 (ApRab5) in *Aiptasia pulchella* and its retention in phagosomes harboring live zooxanthellae. *Biochem Biophys Res Commun* 324(3):1024–1033.
30. Chen M-C, Cheng Y-M, Sung P-J, Kuo C-E, Fang L-S (2003) Molecular identification of Rab7 (ApRab7) in *Aiptasia pulchella* and its exclusion from phagosomes harboring zooxanthellae. *Biochem Biophys Res Commun* 308(3):586–595.
31. Dani V, Ganot P, Priouzeau F, Furla P, Sabourault C (2014) Are Niemann-Pick type C proteins key players in cnidarian-dinoflagellate endosymbioses? *Mol Ecol* 23(18):4527–4540.
32. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
33. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

34. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 40(1):37–52.
35. Consortium TR (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res* 43(Database issue):D123–9.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
37. Baumgarten S, et al. (2013) Integrating microRNA and mRNA expression profiling in Symbiodinium microadriaticum, a dinoflagellate symbiont of reef-building corals. *BMC Genomics* 14(1):704.
38. Tarver JE, Donoghue PC, Peterson KJ (2012) Do miRNAs have a deep evolutionary history? *Bioessays* 34(10):857–866.
39. Friedlander MR, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26(4):407–415.
40. Bailey TL, et al. (2009) MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res* 37:202–208.
41. David CN (1973) A quantitative method for maceration of hydra tissue. *Wilhelm Roux' Arch für Entwicklungsmechanik der Organisme* 171(4):259–268.
42. Quinlan AR (2014) *BEDTools: The Swiss-Army Tool for Genome Feature Analysis.* doi:10.1002/0471250953.bi1112s47.
43. Krüger J, Rehmsmeier M (2006) RNAhybrid: MicroRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 34:451–454.
44. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
45. R Development Core Team (2012) R: A language and environment for statistical computing. Available at: http://www.r-project.org/.
46. Howe EA, Sinha R, Schlauch D, Quackenbush J (2011) RNA-Seq analysis in MeV. *Bioinformatics* 27(22):3209–3210.

47. Helwak A, Tollervey D (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc* 9(3):711–28.

Chapter 4: Conclusions

Studying the molecular basis of the cnidarian-dinoflagellate endosymbiosis is key to understand the susceptibility and resilience of coral reef ecosystems whose structural and trophical basis is formed by the tripartite metaorganism consisting of the coral animal host, their dinoflagellate endosymbionts as well as the associated prokaryotic microbiome. This thesis focused on laying the foundation for the continuous study of the molecular basis of host-microbe interactions in the cnidarian holobiont by presenting the assembly and a comprehensive analysis of the *Aiptasia* genome, the laboratory model system for coral symbiosis. This resource on hand, the thesis then sought to enhance the understanding of host transcriptional responses in the context of *Symbiodinium* infection by providing novel insights into miRNA-mediated post-transcriptional gene modulations and cnidarian gene regulatory mechanisms in general.

In Chapter 2, a traditional Illumina short-read sequencing approach was taken to assemble a technically high-quality genome of *Aiptasia,* which exceeds previously available cnidarian genome assemblies in its completeness and contiguity. While this can be attributed to a large part to the smaller genome size of *Aiptasia* and the resulting higher coverage of sequencing information, the presumably high rates of polymorphisms within the clonal population used for sequencing still prevented an even better result. In the nearest future, this will soon be overcome by including additional genomic information into the assembly processes that are drawn from long-molecule sequencing (e.g. PacBio SMRT) or

the revelation of large-scale chromatin structures using for example Hi-C sequencing. Only recently, the implementation of these techniques improved the assembly of the *Hydra* genome by one order of magnitude in regard of its contiguity, making analysis on a near-chromosome scale feasible.

The availability of a full genome assembly for the first time overcomes the lack of baseline information on the gene content in *Aiptasia* and in a comparative analysis it revealed several features that might be crucial for its endosymbiotic lifestyle as well as those of corals. While there is an ongoing debate about the evolutionary origin of the endosymbiosis in corals and anemones, we could show that the underlying biochemical mechanisms of their symbiotic relationships with dinoflagellates are likely to be similar for several reasons as outlined in this thesis. First, genes and pathways that are considered contributory in the endosymbiosis seem to be highly conserved among corals and anemones and the finding presented in this work of a phylum-specific and novel family of putative pattern-recognition receptors found exclusively in symbiotic cnidarians suggests common selective pressures. Furthermore, the shared specificity of an individual symbiont genotype (i.e. *S. minutum* SSB01) that is found in both corals and anemones implies that the underlying processes enabling the onset and maintenance of a functional endosymbiosis might be similar in both anthozoan lineages. Taken together, these findings as well as previous comparative studies, such as on the nature of photosynthate translocation and coral bleaching in the absence of reactive oxygen species (1, 2), highlight the usefulness of the model-system approach to study the coral symbiosis under laboratory conditions.

The actual power of the sequenced *Aiptasia* genome will also lie in facilitating the development of novel technical approaches for better-designed experiments and by providing a resource for studies investigating (epi-) genetic mechanisms outside the protein-coding genome. To smoothen this transition into the post-genomic era of *Aiptasia*, care was taken to ensure the full accessibility of its genome on all levels of information content. As a central database, all sequencing libraries, the whole genome assembly including the sequence, the annotation of genome features as well as gene and protein models are accessible at the NCBI database (NCBI BioProject accession no: PRJNA261862) and searchable via the central BLAST webpage. In addition, a second database (http://aiptasia.reefgenomics.org) was put online as a focal point for *Aiptasia* related genome research, which includes a genome browser and a BLAST database and which will also include updated versions of the genome assembly in the future.

While a suite of studies investigated changes in the protein-coding genome during the onset and maintenance of the cnidarian-dinoflagellate endosymbiosis in corals and anemones, the second part of this thesis investigated for the first time the role of gene regulatory elements in the endosymbiosis by focusing on miRNA as one of the major class of small regulatory RNAs. The results highlight the importance of miRNA-mediated transcriptome modulations by showing 1) that cnidarian miRNAs are differentially expressed during the onset and maintenance of the endosymbiosis 2) the suitability of Argonaute cross-linking immunoprecipitations to identify high confidence miRNA-mRNA binding sites on

a transcriptome wide scale and 3) that symbiotically regulated miRNAs target genes that are acting in biological processes which were previously identified to be involved in a functional endosymbiosis. These findings suggest that miRNA-mediated modulation of the host-transcriptome is not only crucial in the transcriptional response to endoparasites (3, 4), but might also be important in the mutualistic relationship of cnidarians and dinoflagellates. Incorporating data on post-transcriptional regulation will also provide a critical link for future experiments that focus either on gene expression on the transcriptional level or different protein contents.

While these findings should greatly improve our understanding on miRNA action in cnidarians, further research should also focus especially on the functionality of the seemingly high complementarity between cnidarian miRNAs and their mRNA targets. Building up on the CLIP protocol presented in this work, direct molecular interactions could be investigated using an additional inter-molecular ligation step between the miRNA and its target. Using such advancement in the CLIP protocol called cross-linking and sequencing of hybrids (CLASH) would make studies seeking to unravel the function of individual miRNA nucleotide substitutions, the transcriptional control of individual miRNA loci and their evolutionary origin feasible (5, 6).

The action of host miRNAs most likely is only one process in a plethora of regulatory elements with a putative effective role during *Symbiodinium* infection and maintenance of a stable endosymbiont population. Among others, especially methylation states on the DNA and RNA level as well as chromatin restructuring

by histone modification should be considered in future research. These fundamental processes are likely to contribute to the symbiosis functionality and also to its breakdown and, with the genome on hand, are now realistic to be investigated in great detail.

Future directions in *Aiptasia* research will also greatly depend on the further establishment of common methods for their laboratory handling and genetic modification where especially the closing of the *Aiptasia* life cycle is in urgent need. While spawning and sexual reproduction can be introduced on a regular basis in cultural setups, only the induction of metamorphosis from planula larvae into full adult polyps will allow the generation of transgenic organisms by e.g. CRISPR-Cas9 genome editing techniques. Thus, a major focus of *Aiptasia* research in the nearest future should lie on these ‘best practices’ to complete the transition from the ‘emerging’ to the ‘established’ cnidarian-dinoflagellate model system *Aiptasia.*

4.1 References

1. Burriesci MS, Raab TK, Pringle JR (2012) Evidence that glucose is the major transferred metabolite in dinoflagellate-cnidarian symbiosis. *J Exp Biol* 215(19):3467–77.
2. Tolleter D, et al. (2013) Coral bleaching independent of photosynthetic activity. *Curr Biol* 23(18):1782–1786.
3. LaMonte G, et al. (2012) Translocation of sickle cell erythrocyte microRNAs into *Plasmodium falciparum* inhibits parasite translation and contributes to malaria resistance. *Cell Host Microbe* 12(2):187–199.
4. Ghosh J, Bose M, Roy S, Bhattacharyya SN (2013) *Leishmania donovani*

targets Dicer1 to downregulate miR-122, lower serum cholesterol, and facilitate murine liver infection. *Cell Host Microbe* 13(3):277–288.

5. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153(3):654–665.
6. Grosswendt S, et al. (2014) Unambiguous Identification of miRNA: Target site interactions by different types of ligation reactions. *Mol Cell* 54(6):1042–1054.

Appendix

Permission letter

Permission letter from the office of the executive editor of the "*Proceedings of the National Academy of Science of the USA*" for reprint of my work.

Sebastian Baumgarten, MS
Red Sea Research Center
4700 King Abdullah University of Science and Technology, KAUST
Building 2, Level 2, Office 2246, PO #3714
23955-6900 Thuwal, Saudi Arabia
Phone: +966 (0) 544701256
email: sebastian.baumgarten@kaust.edu.sa

Copyright Permission Letter

To whom it may concern,

I am completing my dissertation at the King Abdullah University of Science and Technology (KAUST) on 'The genome of *Aiptasia* and the role of microRNA in cnidarian-dinoflagellate endosymbiosis' and I kindly request your permission to reprint in my dissertation extracts from the following:

The genome of Aiptasia, a sea anemone model for coral symbiosis (2015) Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, Li Y, Hambleton EA, Guse A, Oates ME, Gough J, Weis VM, Aranda M, Pringle JR, Voolstra CR. *Proceedings of the National Academy of Science of the USA*
doi: 10.1073/pnas.1513318112 [published ahead of print online on August 31st, 2015]

I'd like to request permission for all parts of the article mentioned above, on which I'm the first author and which is openly accessible at the PNAS webpage. I would like to use the requested material to be included in my PhD dissertation entitled 'The genome of *Aiptasia* and the role of microRNA in cnidarian-dinoflagellate endosymbiosis' on which I'm the only author. The

permission is intended exclusively for non-profit use. The dissertation will be available online and open access at the KAUST library. The requested permission extends to any future revisions and editions of my dissertation, including non-exclusive world rights in all languages, and to the prospective publication of my dissertation. These rights will in no way restrict republication of the material in any other form by you or by others authorized by you. By signing this letter you (or your company) confirm that you own the copyright to the above-described material. If these arrangements meet with your approval, please sign this letter where indicated below, and return it to me by email attachment.
Thank you very much,

Sebastian Baumgarten

PERMISSION GRANTED FOR THE ABOVE REQUESTED USE:

Kay McLaughlin for Diane Sullenberger, PNAS Executive Editor

Date: 9/25/2015

Please cite the original PNAS article in full when re-using the material.