Title
Comparative and developmental genomics in the moon jellyfish Aurelia species 1

Author
Gold, David

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Comparative and developmental genomics in the moon jellyfish *Aurelia species 1*

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in Biology

by

David Adler Gold

2014

ABSTRACT OF THE DISSERTATION

Comparative and developmental genomics in the moon jellyfish *Aurelia species 1*

by

David Adler Gold

Doctor of Philosophy in Biology

University of California, Los Angeles, 2014

Professor David K. Jacobs, Chair

This dissertation focuses on the transcriptome of the moon jellyfish (*Aurelia sp.1*). As the chapters progress, larger sets of genes are analyzed, and the work becomes decreasingly comparative in nature, and increasingly focused on *Aurelia*.

In the first chapter, I analyze the POU-class genes. I begin by using comparative genomics and "gene fishing" to resolve the topology of the POU gene tree. I then use ancestral state reconstruction to map the most likely changes in amino acid evolution for the conserved protein domains. Four of the six POU families evolved before the last common ancestor of living animals—doubling previous estimates—and was followed by extensive clade-specific gene loss. POU families best understood for their generic roles in cell-type regulation and stem cell pluripotency (*POU2*, *POU5*) show the largest number of nonsynonymous mutations, suggestive of functional evolution, while those better known for specifying subsets of neural and hormone-producing cell types (*POU1*, *POU3*) appear more similar to the ancestral protein.

In the second chapter, I annotate the homeodomain repertoire for *Aurelia sp.1*, and compare it to data from relatives that lack a medusa life stage (*Nematostella*, *Acropora*, and *Hydra*). Despite having simpler life cycles, the anthozoans *Nematostella* and *Acropora* have far more homeodomains than *Aurelia*, primarily because of clade-specific gene expansions. The one exception to this trend is the non-anterior Hox genes, where *Aurelia* has seven paralogs compared to *Nematostella* and *Acropora's* two. RNA-Seq analyses suggest that these non-anterior Hox genes are expressed dynamically through the *Aurelia* life cycle, and therefore represent candidate genes for future studies in medusozoan bodyplan evolution.

In the final chapter, I offer a broad analysis of the developmental transcriptome for *Aurelia sp.1*. Two major shifts in gene expression occur during the life cycle, correlating with formation of the two "adult" morphs (the transition from primary polyp to polyp, and from polyp to strobila). The morphologically complex medusa stage that distinguishes *Aurelia* from other model cnidarians is not enriched in novel genes, but is enriched in many conserved cell-signaling pathways, transcription factor domains, and neuroactive receptors.

The dissertation of David Adler Gold is approved.

Barnett Schlinger

Robert Wayne

Volker Hartenstein

David K. Jacobs, Committee Chair

University of California, Los Angeles

2014

To Kayleigh Perkov, who fought alongside me every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

LIST OF FIGURES

Chapter 1

Chapter 2

LIST OF FIGURES (CONTINUED)

Chapter 3

ACKNOWLEDGMENTS

I would like to begin by gratefully acknowledging the programs and institutions that funded my research: the NASA/MIT Foundations of Complex Life Astrobiology Team, the Department of Education (Graduate Assistance in Areas of National Need Grant), and the National Institutes of Health (Genomic Analysis and Interpretation Training Grant T32HG002536).

I want to thank my committee members for their guidance. I am also indebted to many professors who were not on my committee, but either directly helped me in my research, or played a formative role in my thinking. This includes Mike Alfaro, Doug Erwin, Marc LaFlamme, Barbara Natterson, Todd Oakley, Nipam Patel, Matteo Pellegrini, Kevin Peterson, Susannah Porter, Bruce Runnegar, Bill Schopf, Janet Sinsheimer, Rob Steele, Roger Summons, and Blaire Van Valkenburgh. I want to give special thanks to David Jacobs for believing in me, and providing me with the freedom to pursue graduate research in a multitude of directions.

Finally, I want to thank the family that helped me through this long process: my wife Kayleigh Perkov, my parents Steven Gold and Susan Adler-Gold, my mother-in-law Lisa Ancich, my grandparents Pearl Gold, Richard Gold, Dorothy Adler, and Reuben Adler, my sister Lauryn Gold, and my cousin Jason Lobell, who is the strongest man I know.

BIOGRAPHICAL SKETCH

EDUCATION AND NOTABLE COURSEWORK

B.S. Ecology and Evolutionary Biology, University of California, Irvine. 2007.

Embryology: Concepts & Techniques in Modern Developmental Biology. Summer 2011. Marine Biological Laboratory. Woods Hole, Massachusetts.

International School of Astrobiology: Origins of the Building Blocks of Life. Summer 2012. Universidad Internacional Menéndez Pelayo. Santander, Spain

PUBLICATIONS

1) Ghisalberti M., Gold D.A., Laflamme M., Clapham M.E., Narbonne G., Summons R.E., Johnston D.T., and Jacobs D.K. (2014) Canopy flow models identify the advantage of size in the oldest communities of multicellular eukaryotes. *Current Biology* 24(3) 305–309.

2) Gold D.A., Robinson J., Farrell A.B., Harris J.M., Thalmann O., and Jacobs D.K. (2014) Attempted DNA extraction from a Columbian mammoth (Mammuthus columbi): Prospects for ancient DNA from asphalt deposits. *Ecology and Evolution* 4(4) 329–336.

3) Gold, D.A., and Jacobs, D.K. (2013) Stem cell dynamics in Cnidaria: are there unifying principles? *Development Genes and Evolution* 223(1-2) 53-66.

4) Takashima, S., Gold, D.A., and Hartenstein, V. (2013) Stem cells and lineages of the intestine: a developmental and evolutionary perspective. *Development Genes and Evolution* 223(1-2) 85-102.

5) Jacobs, D.K., Gold, D.A., Nakanishi, N., Yuan, D., Camara, A., Nichols, S.A., and Hartenstein, V. (2010) Basal Metazoan Sensory Evolution. pp. 175-193 in Key Transitions in Animal Evolution, B. Schieirwater and R. DeSalle eds. CRC Press.

SELECT AWARDS AND HONORS

Agouron Institute Geobiology Post-Doctoral Fellowship (2014-2016)

NIH Genomic Analysis Training Program (2011-2013)

NASA Astrobiology Institute Scholarship (2012)

Development Cover Contest Winner (2012; Volume 139, issue 12)

Collegium of University Teaching Fellows (2011)

The Company of Biologists Ltd Scholarship (2011)

Lorus & Margery Milne Scholarship (2011)

GAANN Graduate Support Fellow (2009-2011)

Chapter 1: Early Expansion and Evolutionary Dynamics of POU Class Genes

ABSTRACT

The POU genes represent a diverse class of animal-specific transcription factors that play important roles in neurogenesis, pluripotency, and cell-type specification. While previous attempts have been made to reconstruct the evolution of the POU class, these studies have been limited by a small number of representative taxa, and a lack of sequences from basally branching organisms. In this study, we performed comparative analyses on available genomes and sequences recovered via "gene fishing" to better resolve the topology of the POU gene tree. We then used ancestral state reconstruction to map the most likely changes in amino acid evolution for the conserved domains. Our work suggests that four of the six POU families evolved before the last common ancestor of living animals—doubling previous estimates—and was followed by extensive clade-specific gene loss. Amino acid changes are distributed unequally across the gene tree, suggesting that the order in which paralogs diverged is not indicative of their similarity to the ancestral sequence. POU families best understood for their generic roles in cell-type regulation and stem cell pluripotency (*POU2*, *POU5*) show the largest number of nonsynonymous mutations, suggestive of functional evolution, while those better known for specifying subsets of neural and hormone-producing cell types (*POU1*, *POU3*) appear more similar to the ancestral protein. Overall, the distribution of paralogs across the animal tree suggests that many POU genes could have unresolved roles in development, or that they specified ancestral cell types in early metazoan evolution, and subsequently diverged as cell type complexity increased.

INTRODUCTION

The POU genes represent a large class of DNA-binding transcription factors known for their roles in cell-type specification and broad developmental regulation (Ryan and Rosenfeld 1997; Phillips and Luisi 2000). The POU homolog *Oct-4* has been extensively studied, as it is the most critical of the four "Yamanaka factors" used to induce pluripotent stem cells in mammals (Niwa et al. 2000; Takahashi and Yamanaka 2006; Ng and Surani 2011). The POU name is an acronym derived from the mammalian genes *Pit-1*, *Oct-1*, and *Oct-2*, as well as the *Caenorhabditis elegans* gene *unc-86*, which all share a 150 amino acid region of high sequence similarity (Herr et al. 1988). Although POU genes have been identified in animals as diverse as sponges and humans, there exists strong conservation within the major domains (Figure 1). POU genes feature a modular, tripartite structure, consisting of an N-terminal POU-specific domain (POU_S), a C-terminal homeodomain (POU_{HD}), and a linker region of varying length connecting the two. The secondary structure of both POU_S and POU_{HD} domains consists of a series of α-helices, which make multiple contacts with DNA through hydrogen bonding with the phosphate backbone or directly to nucleotides (Jacobson et al. 1997; Reményi et al. 2001; Jauch et al. 2010; Esch et al. 2013). In both domains, the third helix serves as the recognition helix, binding to the major groove of DNA and making the majority of direct contacts with nucleotides (Assa-Munt et al. 1993; Dekker et al. 1993; Jacobson et al. 1997). As Figure 1 suggests, these contact regions are often, though not always, the most invariant sites within the POU class.

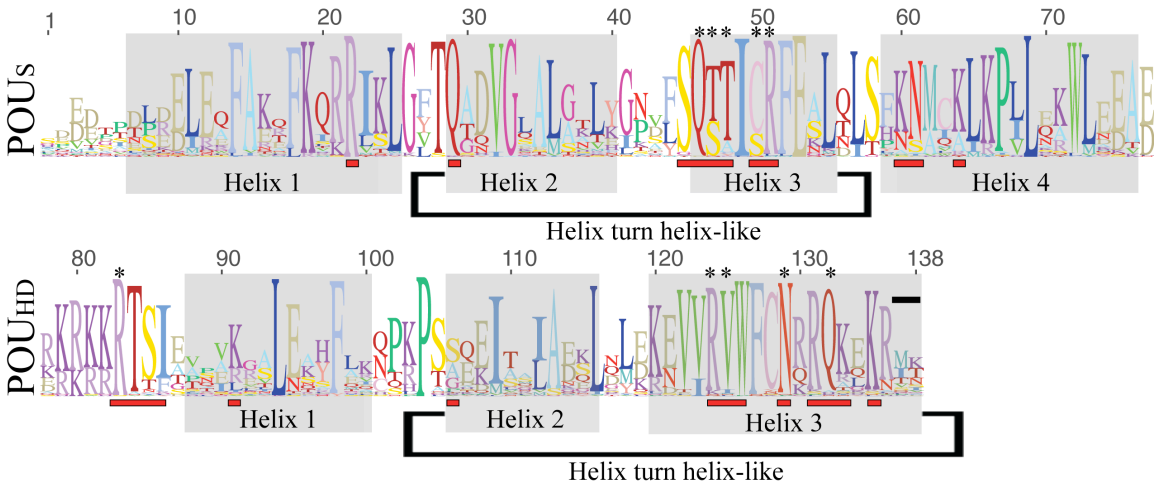

Figure 1: Structure and variation within the POU_S and POU_{HD} domains.

Probable α-helix sub-domains are shaded in grey, while amino acids known to make contact with DNA in at least one POU class are marked with a red bar, and the subset which make direct contacts with nucleotides are marked with an asterisk, based on (based on Klemm et al. 1994; Jacobson et al. 1997; Reményi et al. 2001; Reményi et al. 2003; Jauch et al. 2010; Esch et al. 2013). The combined height of the amino acids at each position indicates the degree of sequence conservation, while the height of each individual amino acid indicates its relative frequency. This figure is based on the alignment we used for our phylogenetic analyses (see Methods and File S1), and was created using the Sequence Logo function in Geneious.

Despite this significant conservation, POU proteins are capable of generating high levels of conformational diversity through complex interactions with DNA and other transcription factors. POU genes form a variety of heterodimers and homodimers that can bind to non-contiguous DNA strands (Voss et al. 1991; Jacobson et al. 1997; Scully et al. 2000; Reményi et al. 2001; Rodda et al. 2005). It is common for POU paralogs to share partially overlapping functions (Erkman et al. 1996; Tichy et al. 2008), and certain POU knockouts can be rescued by a paralog that is not normally expressed in the region (Friedrich et al. 2005). Some POU genes take on multiple isoforms, which oppose each

other in regulation, or work together to bind multiple trans factors (Konzak and Moore 1992; Lee and Salvaterra 2002; Theodorou et al. 2009). Even changes in the spacing between the two DNA binding domains can allow the same transcript to act as an activator in one scenario and a repressor in another (Scully et al. 2000). The last two amino acids of the homeodomain may be particularly important in driving dimerization (Reményi et al. 2001), and it has been hypothesized that the ability of POU genes to form heterodimers and homodimers could be as important as the DNA binding interface for the recognition of cis-regulatory modules (Jauch et al. 2010). Interestingly, this final dipeptide appears to be one of the most variable positions (Figure 1).

Since their initial discovery, more than one thousand POU sequences have been recovered from across the *Metazoa*. These are generally organized into six families (*POU1-POU6*). Multiple POU families have been described in every annotated animal genome, and many lineages, particularly vertebrates, have multiple paralogs in multiple families. This has resulted in an extensive nomenclature, summarized in Table 1. No POU genes have been recovered from a non-metazoan, which suggests that the POU_S domain represents an animal novelty that evolved from the more ancient homeodomain during the early evolution of animals (Degnan et al. 2009). However, the presence of multiple, and often non-overlapping, POU families in early-branching animal lineages makes rooting the POU gene tree difficult, and has led to conflicting topologies in gene tree reconstruction (Kamm and Schierwater 2007; Larroux et al. 2008; Ryan et al. 2010).

Table 1: Division of POU homologs into the six major classes, including common names.

	Mammalian Homologs	***Drosophila* Homologs**	***Caenorhabditis* Homologs**
POU1	*POU1F1 (Pit-1)*	None	None
POU2	*POU2F1 (Oct-1)*, *POU2F2 (Oct-2)*, *POU2F3 (Oct-11)*	*pdm-1* (*nubbin; dPOU-19; twain; dOct1*), *pdm-2* (*miti-mere; dOct-2*)	*Ceh-18*
POU3	*POU3F1 (Oct-6; SCIP), POU3F2 (Oct-7; Brn-2), POU3F3 (Oct-8; Brn-1), POU3F4 (Oct-9; Brn-4; DFN3)*	*vvl (cf1a; drifter)*	*Ceh-6*
POU4	*POU4F1 (Brn-3a; RDC-1; Oct-T1), POU4F2 (Brn-3b; Brn-3.2), POU4F3 (Brn-3c; Brn-3.1; DFNA15)*	*acj6 (Ipou)*	*Unc-86*
POU5	*POU5F1(Oct-3; Oct-4)*, *POU5F2 (SPRM-1), Pou2/V*	None	None
POU6	*POU6f1 (Brn-5; mPOU)*, *POU6f2 (Emb; RPF-1)*	*pdm-3*	None

To better understand the diversity and evolution of POU genes, we adopted a comparative genomics approach to reconstruct the class topology. The results of this study were corroborated with gene fishing, using degenerate PCR primers to capture novel POU homologs from a variety of understudied animal clades. Finally, we used ancestral state reconstruction to track the most likely trajectory of POU sequence evolution. Taken together, our results suggest that four out of the six major families of POU genes (*POU6*, *POU1*, *POU3*, and *POU4*) were present before the last common ancestor of all living animals, which is double the previous estimate. The POU families appear to have evolved primarily through gene duplication followed by neofunctionalization (Lynch and Conery 2000; Innan and Kondrashov 2010), where one paralog retains the ancestral amino acid sequence (and presumably aspects of the ancestral function), while the other duplicate takes on significantly more nonsynonymous mutations.

RESULTS

A Comparative Genomics Approach Resolves Many Aspects of the POU Gene Tree Topology

We began by surveying available animal genomes for POU-domain sequences. For phylogenetic analyses, we ultimately chose a subsample of taxa that included model laboratory animals as well as representatives of major clades from across the animal tree (discussed in detail in the Methods section). We employed both maximum likelihood and Bayesian approaches to tree building. To account for the low phylogenetic support of sequences from the sponges *Amphimedon queenslandica* and *Oscarella carmela*, as well as the ctenophore *Mnemiopsis leidyi,* we also attempted tree reconstruction excluding these taxa. The results of these analyses are summarized in Figure 2 (see Figures S1-S6 in the Supplementary Materials for full trees).

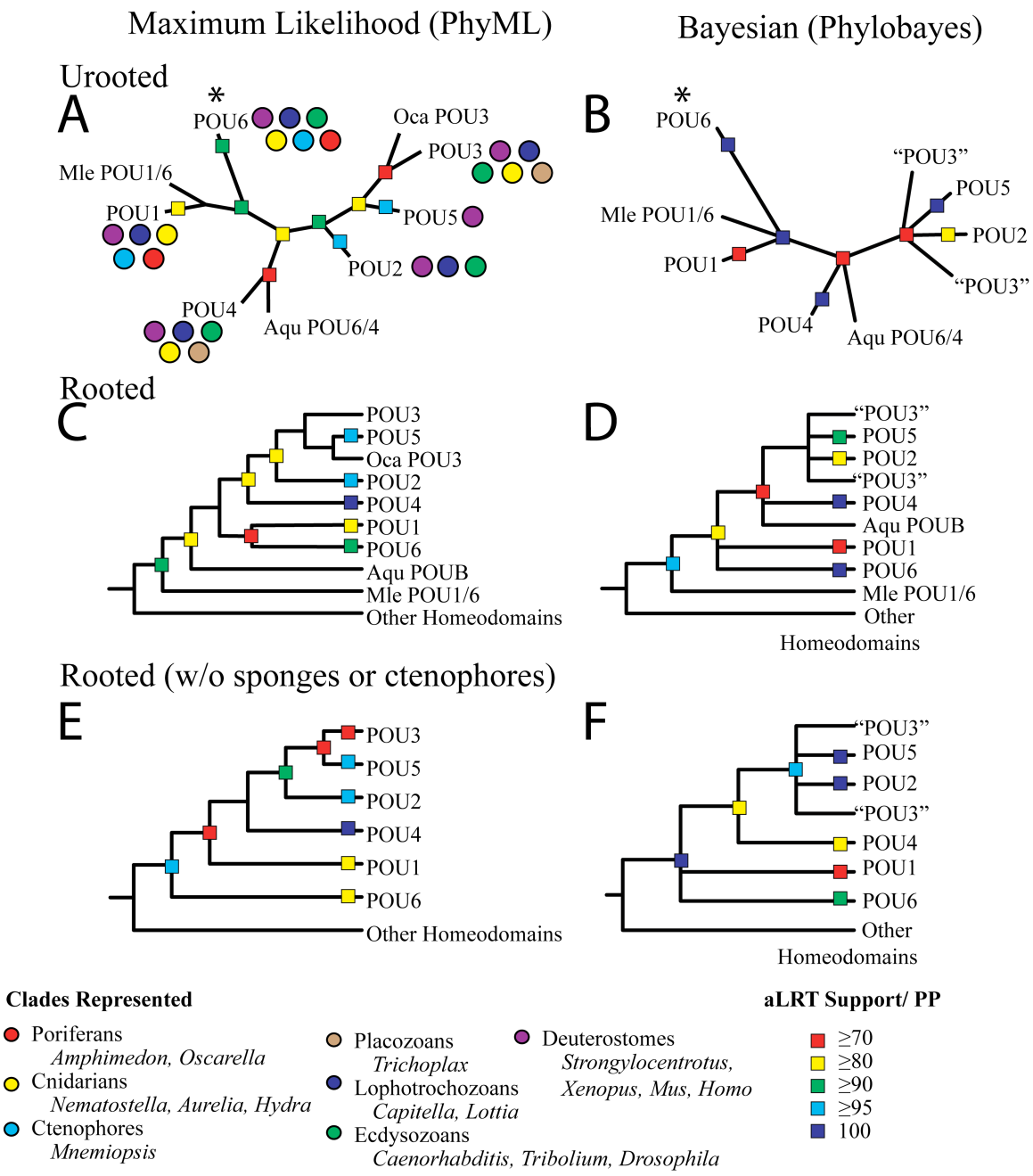

Figure 2. Summary of maximum likelihood and Bayesian reconstructions of our POU dataset.

See the Methods section and Supplementary Materials for more complete information on taxon sampling and support values for all nodes. Genes with uncertain phylogenetic position from *Amphimedon* (Aqu), *Oscarella* (Oca) and *Mnemiopsis* (Mle) are singled out. (A-B) Unrooted topologies. The location of the midpoint root is marked with an asterisk. (C-D) Topologies that have been rooted by the inclusion of

additional homeodomains. (E-F) Rooted topologies with all sequences from *Amphimedon*, *Oscarella* and *Mnemiopsis* removed.

While we were unable to generate a single topology across all analyses, we were able to resolve some areas of uncertainty regarding the relationships between POU families. Previous studies rooting the POU class with homeodomains have recovered *POU6* as the original POU family. However, there has been disagreement whether the next family to diverge was *POU4* (e.g. Ryan et al. 2010) or *POU1* (e.g. Larroux et al. 2008; Millane et al. 2011); different rooting methods have also produced alternate topologies for the same dataset (Kamm and Schierwater 2007). In contrast, all of our analyses preferred *POU1* as the closest paralog to *POU6*, although the nature of that relationship varied, with *POU1* and *POU6* occasionally forming sister clades or a polytomy (Figure 2C, 2D, 2F). Still, there are several reasons to prefer *POU6* as an outgroup to the other extant POU families. Topologies illustrated in Figure 2C and 2D include highly divergent and poorly-supported sequences from *Mnemiopsis* and *Amphimedon*, which increases the probability of long-branch attraction artifacts. When these sequences are removed, maximum likelihood strongly supports *POU6* as the outgroup (Figure 2E), while the Bayesian phylogeny generates a polytomy between *POU6* and *POU1* (Figure 2F). Given the small number of phylogenetically informative sites in our alignment, it is likely that the Bayesian approach lacks sufficient information to produce a topology with strong posterior support. Midpoint rooting on the unrooted topologies places the root within the *POU6* family (Figure 2A, 2B), and an additional rooting process used during ancestral state reconstruction (discussed in detail in the Methods section) also supports *POU6* as the outgroup.

Following *POU6* and *POU1*, all of our analyses support *POU4* as the next paralog to diverge. One gene from the sponge *Amphimedon*, which has previously been described as *POUB* (Larroux et al. 2008), has an affinity with *POU4* in some of our analyses, and *POU6* in others. This was followed by either a split between *POU2* and *POU3/5* (maximum likelihood analyses; Figure 2A, 2C, 2E) or a polytomous *POU3* "bush" which includes monophyletic *POU2* and *POU5* classes (Bayesian analyses; Figure 2B, 2D, 2F). As Figure 2A illustrates, *POU3* includes representatives from a number of basally branching animal taxa, including cnidarians, the placozoan *Trichoplax adherans*, and possibly the sponge *Oscarella*. *POU2* was only recovered from bilaterian animals, while *POU5* appears restricted to vertebrates, which supports the hypothesis that these families represent more recent, clade-specific duplications.

Gene fishing recovers putative POU3 and POU4 classes in sponges.

By using a diverse selection of taxa, we uncovered several unanticipated results regarding the distribution of POU genes across the animals. Firstly, our analyses provide good support for a *POU3* homolog in *Oscarella*, as well as moderate support for a *POU4* homolog in *Amphimedon*. This potentially doubles the number of POU families identified in the sponges, as previous analyses have only recognized *POU6* and *POU1* homologs in the Porifera (Larroux et al. 2008). A second surprise comes from the taxon distribution of the *POU1* family. *POU1* is present in early-branching animals, such as cnidarians, ctenophores, and sponges, as well as vertebrates and the chordate amphioxus (Jacobs and Gates 2003; Candiani et al. 2008). Our analyses suggest that *POU1* is also present in the annelid *Capitella teleta*, but absent from all other sampled protostomes. The

identification of *POU1* in annelids is not new, as it has previously been described in the polychaete worm *Platynereis dumerilii* (Raible et al. 2005), but the hypothesis that the annelids are the only protostomes to retain this homolog has not been formalized. Indeed, while we were also able to recover a candidate *POU1* from the genome of the leech *Helobdella robusta* (Figure S7), we found no other protostome *POU1* candidates in the NCBI database, or in any additional publically available protostome genomes.

To corroborate these results, we performed a gene fishing experiment, using degenerate PCR primers to amplify POU genes from a variety of understudied animal lineages (summarized in Table 2). Family designations for the recovered genes were determined using BLAST, alignments of the linker regions (Figure S7), and phylogenetic analysis (Figure S8). In our phylogenetic analyses, the linker was discarded, for although the region is often conserved within POU families, it is difficult to homologize between them. However, this also makes the linker a good candidate for supporting family affinity, as it reduces the probability that our phylogenetic results are caused by convergent evolution within the otherwise largely invariant POU_S and POU_{HD} domains.

Table 2: Results of gene fishing experiments.

Species Name	Phylum	Class	POU Genes Recovered
Acarnus erithacus	Porifera	Demospongiae	POU1, POU4 (2)
Tethya aurantia	Porifera	Demospongiae	POU4
Spongilla sp.	Porifera	Demospongiae	POU4
Haliclona sp.	Porifera	Demospongiae	POU1
Rhabdocalyptus dawsoni	Porifera	Hexactinellida	POU1 (2), POU3
Pleurobrachia sp.	Ctenophora	Tentaculata	POU1
Agaricia sp.	Cnidaria	Anthozoa	POU1, POU3
Anthopleura elegantissima	Cnidaria	Anthozoa	POU3
Fungia sp.	Cnidaria	Anthozoa	POU1
Pelagia colorata	Cnidaria	Scyphozoa	POU4
Convolutriloba sp.	Acoelomorpha	Acoela	POU3, POU4
Notoplana acticola	Platyhelminthes	Turbellaria	POU3, POU4
Stylochus tripartitus	Platyhelminthes	Turbellaria	POU3, POU4 (3)
Alitta virens (formally *Nereis virens*)	Annelida	Polychaeta	POU3, POU4
Phragmatopoma californica	Annelida	Polychaeta	POU4
Hydroides sp.	Annelida	Polychaeta	POU4
Acanthina sp.	Mollusca	Gastropoda	POU4
Kelletia kelletii	Mollusca	Gastropoda	POU3 (2), POU4
Transennella sp.	Mollusca	Bivalvia	POU4
Crassostrea gigas	Mollusca	Bivalvia	POU3

Our gene fishing results are consistent with comparative genomic inferences regarding clade-specific gene gain and loss of POU families. The recovery of a *POU3* homolog in the hexactinellid *Rhabdocalyptus dawsoni* and *POU4* homologs from the demosponges *Acarnus erithacus*, *Tethya aurantia*, and *Spongilla sp.* strongly support our interpretation

of the *Amphimedon* and *Oscarella* data, and collectively double the number of POU families known from the sponges. Because sponges commonly house a variety of symbiotic and commensal organisms (Brusca and Brusca 2003), contamination is a concern. However, a number of observations argue against contamination. Firstly, we obtained different *POU* genes from different sponge clades, with *POU4* being exclusive to demosponges (including *Amphimedon*), while *POU3* was recovered in the hexactenellid *Rhabdocalyptus* and the homoscleromorph *Oscarella*. Secondly, we obtained *POU4* genes from both the saltwater demosponges *Acarnus* and *Thethya*, as well as the freshwater sponge *Spongilla*. Thirdly, although the sponge *POU* genes do not form monophyletic clades in phylogenetic analyses (Figure S8), they also show no consistent affinity to any other animal phyla across NCBI BLAST searches. Consequently, we infer that *POU1*, *POU3*, *POU4*, and *POU6* were all present in the last common ancestor of sponges.

As with any gene fishing expedition, one must be cautious about making hard conclusions regarding gene absence. For example, we did not recover any *POU2* or *POU6* genes, even though *POU6* homologs have been identified in every annotated metazoan genome (excluding nematodes) and *POU2* in every annotated bilaterian genome. Therefore, it is unclear how we should interpret our failure to recover *POU1* genes from the annelids *Alitta, Phragmatopoma,* or *Hydroides*, despite their presence in the three annelid genomes. A previous gene fishing study also failed to recover *POU1* in the earthworm *Lumbricus terrestris* (Shah et al. 2000), which suggests that *POU1* has

either been lost in many annelid lineages, or that annelid *POU1* genes are difficult to amplify with degenerate primers.

Still, the distribution of gene absences might provide some information. For example, our inability to find *POU3* or *POU4* homologues in *Pleurobrachia* supports the hypothesis that these families are absent from the two major ctenophore lineages (the Tentaculata and Nuda), and thus missing from the phylum altogether. Similarly, although *POU1* genes were recovered from cnidarians, sponges, and a ctenophore, none were recovered from flatworms, acoels, or molluscs, which is consistent with their absence in publically available genomes.

Ancestral state reconstruction supports a pattern of gene duplication followed by protein neo-functionalization

Resolving the topology and affinity of metazoan POU homologs allowed us to study the directionality of evolution within the POU_S and POU_{HD} domains. We used maximum likelihood-based ancestral state reconstruction on a species-tree-corrected gene tree to track all amino acid changes that occurred at each node up to the common ancestor of the extant POU classes (Figure 3). Out of 173 amino acid changes, 117 occurred within an α-helix domain, and 95 changes were "significant", which we define as a shift from one major type of amino acid to another (i.e. positively charged (K, R, H), negatively charged (D, E), hydrophilic (S, T, N, Q, C, G, P), and hydrophobic (A, I, L, M, F, W, V, Y)). Our results suggest that mutations are not distributed evenly across the tree; after most bifurcations, one lineage appears to accumulate more amino acid changes than the other.

To verify this pattern, we used the DIVERGE (v3.0) software package to perform pairwise comparisons between gene families (Gu et al. 2013; see Materials and Methods for more information). In our tests for differences in rates of significant amino acid substitutions, we determined that *POU4* was significantly different from *POU2*, *POU3*, or *POU5*, and that *POU5* was significantly different from *POU2*. Taken collectively, our analyses suggest three major times when a significant increase in amino acid substitutions occurred: (1) when *POU6* split from the last common ancestor of all other POU families, (2) when *POU4* split from the last common ancestor of *POU2/3/5*, and (3) when *POU5* and *POU2* split from *POU3*. These results appear consistent with neofunctionalization models of gene duplication, which predict purifying selection on one gene duplicate, and a release of purifying selection combined with the evolution of a new adaptive function in the other duplicate (Innan and Kondrashov 2010). This would be in contrast to subfunctionalization models that predict relaxed purifying selection on both gene duplicates (Innan and Kondrashov 2010). These results also lead to some unintuitive conclusions regarding the similarity between extant POU families and their ancestral nodes. For example, although *POU6* appears to be the earliest branching family, *POU1* has accumulated far fewer significant amino acid substitutions during its evolution, which is indicative of purifying selection and perhaps functional continuity. Similarly, although *POU2* appears to be sister to a *POU3*/*POU5* clade, *POU3* has accumulated far fewer significant substitutions since splitting from the common ancestor than either *POU2* or *POU5*. Similar to our presence-absence data described earlier, these results are consistent with the hypothesis that the last common ancestor of the *POU2/3/5* super-family was

POU3-like, and that *POU2* and *POU5* represent clade-specific duplications in bilaterians and vertebrates respectively.

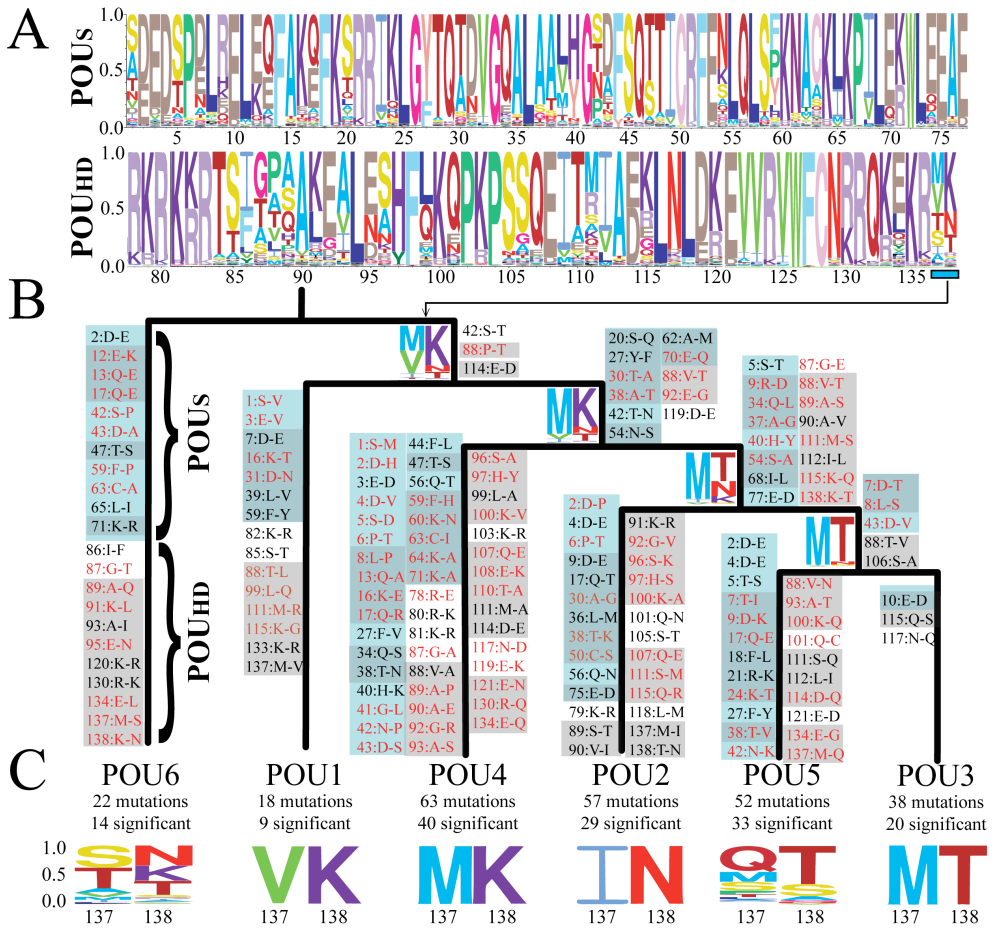

Figure 3: Ancestral sequence and evolutionary trajectory of the POU_S and POU_{HD} domains.

(A) Ancestral state reconstruction of the original POU_S and POU_{HD} domains. The probability of an amino acid being the ancestral state at each site is represented by the height of the letter, with the most probable peptide at the top. (B) Amino acid substitutions that occurred at each node, based on the most likely peptide at each node versus the ancestral node. Significant amino acid substitutions (i.e. moving between amino acids with positively charge, negatively charge, polar uncharged, or hydrophobic side chains) are colored in red. Mutations in the POU_S domain are highlighted in blue, and mutations occurring within an α-helix sub-domain are highlighted in grey. (C) Total number of mutations that occurred between the common ancestor of each POU class and the ancestral POU sequence. The probability of the final dipeptide for the ancestor of each POU class is visualized at the bottom of the figure, and at the bifurcation of each ancestral node.

Given the modular nature of POU genes, we were curious whether there was any evidence of some modules evolving at different rates than others. Mutations appear to be fairly evenly distributed between POU_S and POU_{HD} domains at every node, but more mutations occur in α-helices in the POU_{HD} domain (66 out of 87 amino acid changes for POU_{HD} versus 51 out of 86 changes for POU_S), even though the α-helix portion of POU_{HD} is smaller than in POU_S. As mentioned earlier, most amino acids known to play a direct role in DNA binding are largely invariant across the gene family. However, there are several significant amino acid substitutions in *POU4* (60K→N 64K→A) and *POU6* (91K→L) at positions involved in DNA binding in other POU classes. The consequences of these substitutions are unclear; the crystal structure has not been studied in *POU4* or *POU6*, so the impact that these substitutions have on DNA binding/bending are unknown. The protein folding prediction software I-TASSER (Roy et al. 2010) suggests that these substitutions have a minor impact on the shape of α-helices (Figure 4).

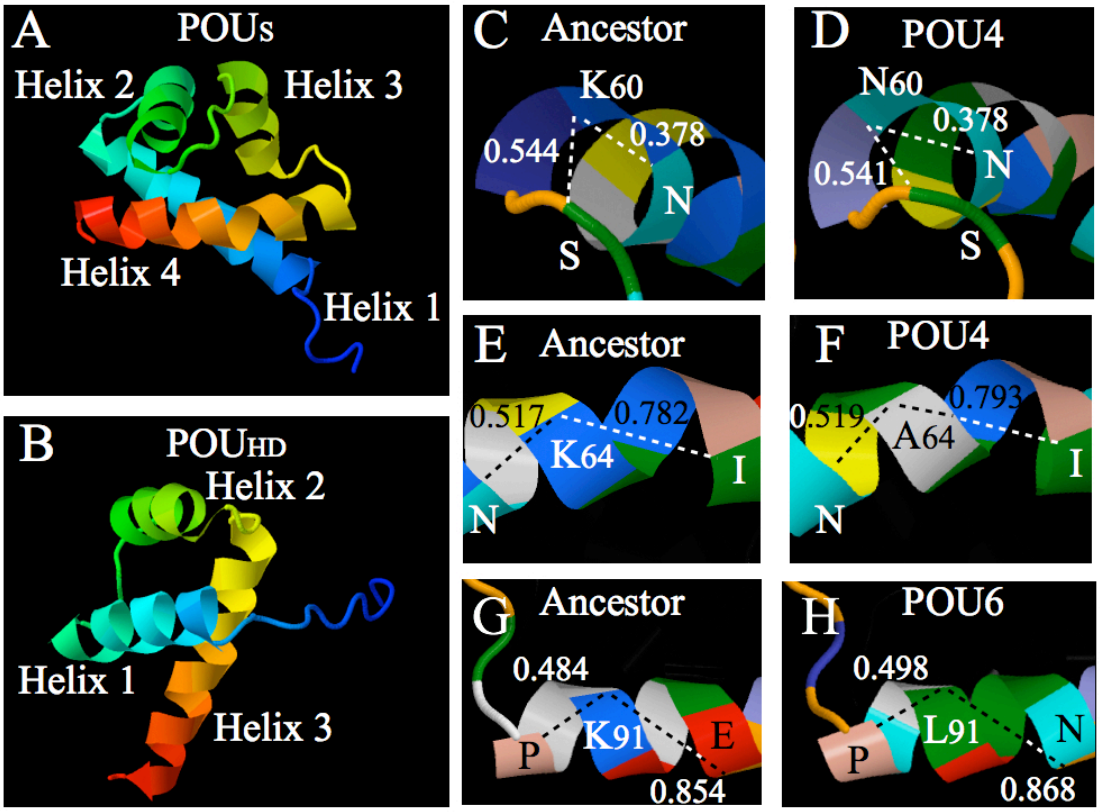

Figure 4: Predicted structure of the ancestral POU_S and POU_{HD} domains, and the effects of significant amino acid substitutions on protein folding.

Amino acid sequences were taken from the ancestral state reconstruction analysis, and folding was predicted using the I-TASSER server. The protein models were manipulated in Jmol. (A-B) Structure of the ancestral (A) POU_S and (B) POU_{HD} domains. (C-H) Comparisons of protein folding in the ancestral sequences versus ancestral POU family members for significant amino acid substitutions to known DNA binding sites. All measurements are in nm. (C) Ancestral condition of position 60. (D) Derived condition of position 60 in the last common ancestor of POU4. (E) Ancestral condition of position 64. (F) Derived condition of position 64 in the last common ancestor of POU4. (G) Ancestral condition of position 91. (H) Derived condition of position 91 in the last common ancestor of POU6.

The last two amino acids of the POU_{HD} domain are distinct at each family-level bifurcation, and in 4 of the 6 cases there is a conserved combination of an aliphatic

residue followed by a charged residue. This supports the hypothesis that this dipeptide is important in driving functional differentiation between the classes (Jauch et al. 2010). In *POU1*, amino acids 135-138 sit in an extended conformation beyond the terminus of the alpha helix (Jacobson et al. 1997), which is likely critical in driving dimerization in the final dipeptide. Thus, there might be an implicit loss of dimerization specification in *POU6* and *POU5*, the two families that have lost this aliphatic/charged motif in the final dipeptide. Position 134, two base pairs upstream of this final dipeptide, also exhibits an interesting evolutionary pattern; at each bifurcation, the ancestral peptide (glutamic acid) is retained in one lineage, while the other lineage exhibits a significant substitution (*POU6* E→L, *POU4* E→Q, *POU5* E→G). Protein folding predictions of the POU_{HD} domain suggest that these substitutions have impacted the conformation of the recognition helix C-terminus (Figure 5); *POU6*, *POU1*, and *POU3* have retained the structure of the ancestral POU_{HD} protein (see Figure 4B), while *POU2*, *POU4*, an *POU5* exhibit an unwinding of the final dipeptide.

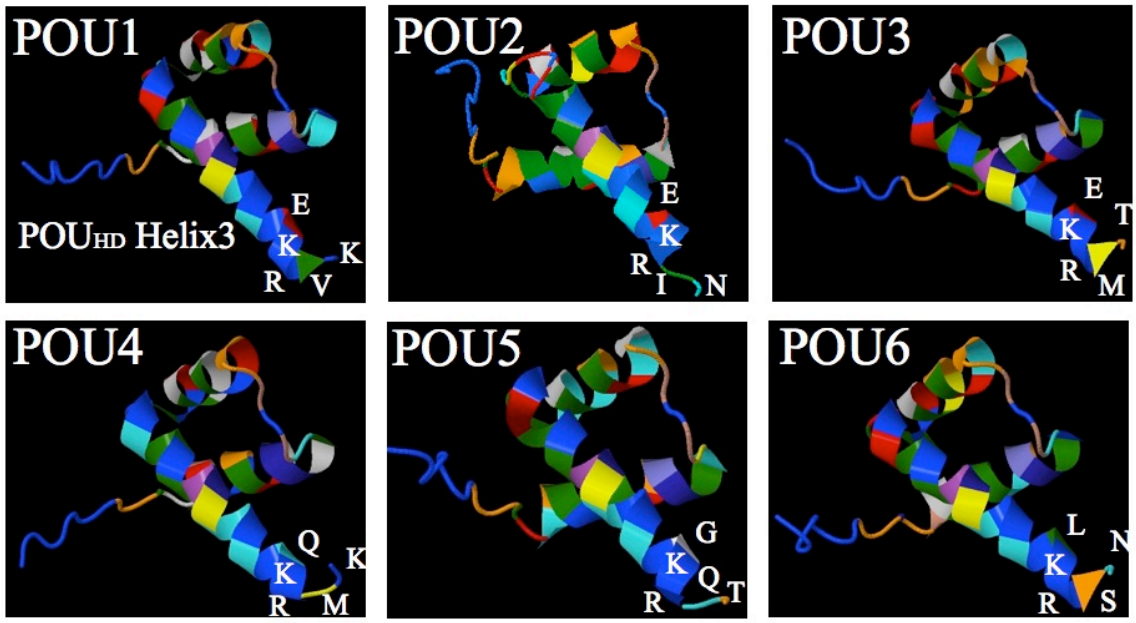

Figure 5: Predicted folding of POU_{HD} domains in the last common ancestor of each POU family, with a focus on the C-terminus.

The last common ancestors of *POU1*, *POU3*, and *POU6* exhibit C-termini that are similar to the ancestral POU protein (see Figure 4B). In contrast, the last common ancestors of *POU2*, *POU4*, and *POU5* display a unwinding of the final dipeptide from the recognition α-helix.

DISCUSSION

The results of this study are summarized in Figure 6. The POU gene tree is marked by an early diversification—with four out of six families evolving before the last common ancestor of living animals—which was followed by significant gene loss in certain clades. From a functional standpoint, these results are surprising. *POU6*, *POU4*, and *POU3*, which are best known for their roles in neurogenesis, appear to have first evolved in a clade without formal neurons, while *POU5*, which is best known for its role in regulating cell pluripotency, appears to be a vertebrate novelty. Hopefully the illumination of these paradoxes will lead to the acquisition of additional functional information, so that they might be resolved.

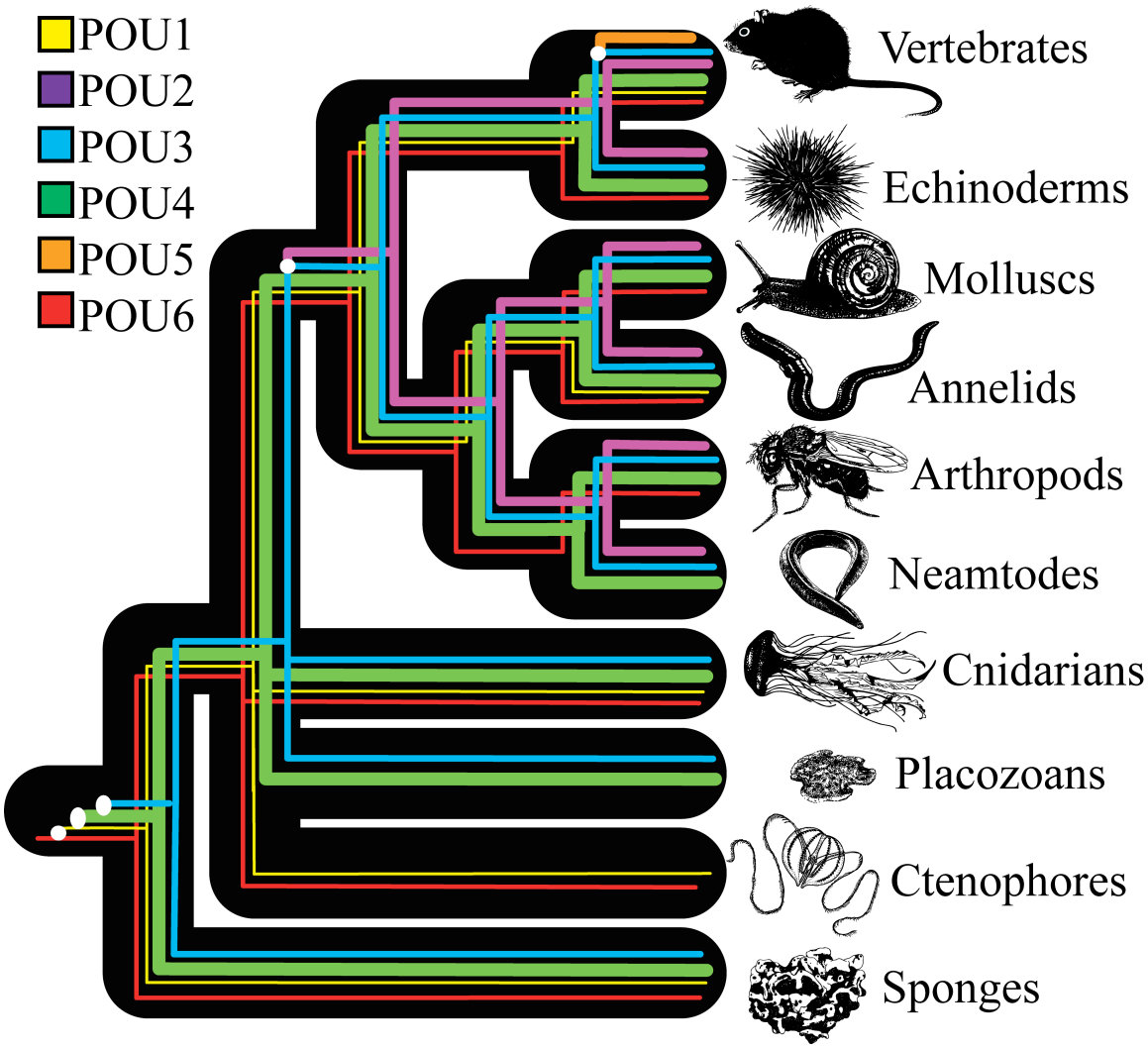

Figure 6: Reconciliation of the POU gene tree and our animal phylogeny.

This figure summarizes our hypothesis regarding how major POU families were gained and lost across the major animal phyla. The line thickness of each family indicates its relative degree of divergence from the ancestral POU sequence. Presence/absence results as they retain to each phylum were verified using BLAST searches on NCBI and through publically available genome datasets. Some of the animal images in this figure were modified under the creative commons agreement from the OpenLearn Tree of Life project (http://www.open.edu/openlearn/nature-environment/natural-history/tree-life). The base of the animal tree, particularly the placement of ctenophores and placozoans, is an active area of research. Opposing animal phylogenies to the one we present here have the potential to alter how rapid this initial expansion of POU

classes was, but all currently debated animal topologies would still require the divergence of the first four POU genes prior to the evolution of the *Eumetozoa*.

Regarding the presence of "neurogenic" POU homologs in sponges, we suspect that this paradox could be resolved in two possible ways, which are not necessarily mutually exclusive. The first possibility is that some of the poorer-studied functions that "neurogenic" POU genes play in model organisms might represent the more ancestral roles of these families. For example, although *POU3* is best known for its neurogenic role across a diverse set of bilaterians (Bürglin and Ruvkun 2001; Ramachandra et al. 2002; Meier et al. 2006; Backfisch et al. 2013; Nomaksteinsky et al. 2013), it is also retains a function in the development of secretory/osmoregulatory organs (Chavez et al. 1999; Bürglin and Ruvkun 2001; O'Brien and Degnan 2002; Ramachandra et al. 2002; Zhang and Xu 2009), which could indicate its primitive role in animals that lack neurons. The second possibility is that the various cell types and organs regulated by POU genes share a deep common ancestry amongst the small number of cell types and structures that basally-branching animals possess (Jacobs et al. 2007; Jacobs et al. 2010). So for example, although *POU1* is restricted to the vertebrate pituitary (Ingraham et al. 1988; Li et al. 1990; Simmons et al. 1990; Niwa et al. 2000; Takahashi and Yamanaka 2006; Ng and Surani 2011), while *POU4* paralogs are involved in brain and/or ear development, both of these families are expressed in the rhopalia of the jellyfish *Aurelia*, which plays a role in mechanoreception, photoreception, and possibly growth (Cary 1916a; Cary 1916b; Nakanishi et al. 2010). Similarly, *POU6* and *POU4* homologs overlap in the mechanosensory statocysts of the cnidarian *Craspedacusta* (Hroudova et al. 2012). In the only study done on POU expression in a sponge, RTPCR in *Ephydatia* shows that several

POU genes are expressed during the period of canal formation, suggesting they might serve a purpose in choanocyte formation (Assa-Munt et al. 1993).

It worth noting that *POU3*, *POU4,* and *POU6* are just a subset of the many "synaptic signaling" genes found in the sponges (Sakarya et al. 2007), and a lack of formal neurons does not prevent these animals from signal propagation or producing coordinated responses (Leys and Meech 2006; Jacobs et al. 2007; Jacobs et al. 2010). For example, *Tethya* (one of the sponges from which we recovered a *POU4* homolog) is known to go through both rhythmic expansions of body size, as well more rapid responses to stimulations, which can be mediated by a wide range of neuroactive substances (Nickel 2004; Ellwanger and Nickel 2006). Hexactinellids (such as *Rhabdocalyptus*, the sponge we obtained a *POU3* homolog from) possess a distinct mode of transmitting non-neural impulses; their epidermis consists of a multinucleate syncytium, which can conduct electric signals across the body in response to touch, and perhaps light by using their spicules as optical glass fibers (Leys et al. 1999; Müller et al. 2006). The presence of distinct POU families in sponges with distinct means of impulse conduction deserves further study.

Similar explanations might elucidate the opposing trend, particularly why *POU2* and *POU5*, relative late-comers in the gene class, appear to play roles in very early and critical developmental processes, including epidermal stratification, cell apoptosis, and stem cell pluripotency. At face value, this appears contradictory to the popular idea of canalization in evo-devo, where more primitive developmental phenomena are harder to

modify than derived ones (Waddington 1942; Gibson and Wagner 2000; Flatt 2005). But in this instance, our results could suggest that certain developmental phenomena assumed to be ancient due to their critical importance at early stages of development might actually be derived. This issue is particularly germane to our understanding of the role of the *POU5* family in the evolution of stem cells. The *POU5f1/Oct4* paralog appears to be involved in cell pluripotency across vertebrates (Scully et al. 2000; Morrison and Brickman 2006; Lavial et al. 2007), although many other aspects of the well-described pluripotency gene regulatory network appear unique to mammals (Reményi et al. 2001; Fernandez-Tresguerres et al. 2010; Jauch et al. 2010; Esch et al. 2013). This variability within the vertebrates makes it difficult to interpret the significance of the observation that POU genes are involved in stem cell dynamics of the planarian worm *Schmidtea mediterranea* (Onal et al. 2012) as well as the cnidarian *Hydractinia echinata* (Millane et al. 2011). The results of our paper would suggest that these invertebrate genes do not represent genuine *POU5* orthologs. Indeed, adding these proteins to our phylogenetic alignment suggests that candidate *Schmidtea* and *Hydractinia* genes represent *POU4* and *POU3* paralogs respectively (Figure S9). Additionally, these invertebrate POU sequences lack the α-helix domain that exists in the linker of amniote *POU5* peptides (Figure S10), which is necessary for inducing pluripotency in mammalian cells (Esch et al. 2013). This could be interpreted as further evidence for the independent evolution of invertebrate and mammalian stem cells (Steele et al. 2011; Gold and Jacobs 2013), although additional regulatory and epigenetic similarities between planarian and mammalian stem cells suggest that there might still be deep underlying conservation of the pluripotency network, even if disparate POU paralogs are ultimately utilized in different animal

lineages (Onal et al. 2012). Such uncertainty only reinforces the point that we are just beginning to appreciate how dynamically evolving protein families become integrated into ancestral and novel genetic networks.

In an era of comparative and functional genomics, the elucidation of gene trees will prove just as important as the resolution of species trees. Our results suggest that POU genes have undergone a complex series of lineage-specific duplication and loss, which will only be fully clarified by using an extensive and diverse sampling of animals (see Frankenberg et al. 2010; Frankenberg and Renfree 2013 for similar results regarding the evolution of *POU5* paralogs within the vertebrates). Greater study of POU genes in animal clades such as sponges, cnidarians, ctenophores, and annelids should help elucidate the functional evolution of the POU class, and will be critical to determining cellular homologies between the invertebrates and vertebrates. This will likely prove important for establishing invertebrate model systems for a variety of developmental phenomena, including neurogenesis and stem cell dynamics.

MATERIAL AND METHODS

Data Collection and Alignment

For our phylogenetic analysis, we searched for POU sequences from the publically available genomes of *Amphimedon queenslandica* (demosponge)*, Ocsarella carmella* (homoscleromorph sponge)*, Hydra magnipapillata* (cnidarian), *Nematostella vectensis* (cnidarian), *Mnemiopsis leidiy* (ctenophore), *Trichoplax adherans* (placozoan)*, Capitella telata* (annelid)*, Lottia gigantea* (mollusc), *Caenorhabditis elegans* (nematode),

Tribolium castaneum (arthropod), *Drosophila melanogaster* (arthropod), *Strongylocentrotus purpuratus* (echinoderm), *Xenopus tropicalis* (vertebrate), *Mus musculus* (vertebrate), and *Homo sapiens* (vertebrate). We also included sequences based on our unpublished transcriptomic data for *Aurelia sp.1* (cnidarian). Databases were queried using the Human Pit-1 POU_S domain: DSPEIRELEKFANEFKVRRIKLGYTQT NVGEALAAVHGSEFSQTTICRFENLQLSFKNACKLKAILSKWL. Sequences from *Hydra*, *Nematostella*, *Lottia*, *Caenorhabditis*, *Tribolium*, *Drosophila*, *Strongylocentrotus*, *Xenopus*, *Mus*, and *Homo* were collected from Metazome (http://www.metazome.net/) using BLASTP against the predicted proteomes. For *Amphimedon queenslandica*, we used TBLASTN against the Spongezome Metazome database (http://spongezome.metazome.net). Sequences from *Capitella* and *Trichoplax* were collected from the Joint Genome Institute using BLASTP. *Oscarella* sequences were obtained from the predicted protein models (OCAR G-PEP) available on the Compagen website (Hemmrich and Bosch 2008). *Mnemiopsis* proteins were recovered using BLASTP against the protein models (v2.2) available at the NIH *Mnemiopsis* Genome Project Portal (http://research.nhgri.nih.gov/mnemiopsis/blast/). The proteins we recovered for *Amphimedon* and *Mnemiopsis* are not identical to those that have been previously published (Larroux et al. 2008; Ryan et al. 2010); we interpreted this as resulting from improvements in the respective genome/proteome assemblies, and chose to work with the POU proteins we recovered. For the *Capitella POU1* gene, we recovered an alternative transcript using TBLASTN against the genome, which contained part of the POU_S domain missing from the predicted peptide; this longer sequence was

used for subsequent analyses. Accession numbers for all genes are included in the alignment, available as Supplementary File S1.

Sequences were aligned using the MUSCLE algorithm (Edgar 2004) in Geneious (v.5.4.6., created by Biomatters and available from http://www.geneious.com/). The alignment was edited by hand and restricted to the POU_S and POU_{HD} domains. Redundant sequences, unalignable sequences, and uninformative (unique) insertions were manually removed. The final alignment is available as Supplementary File S1.

Phylogenic Analyses

We used ProtTest3 (Darriba et al. 2011) to determine the best-fitting model of amino acid evolution for our alignments. The program strongly preferred the LG model in conjunction with a gamma distribution and four substitution rate categories. We used PhyML (Guindon et al. 2010) to perform maximum likelihood estimates; node values were determined using aLRT SH-like support. We used PhyloBayes 3.3 (Lartillot et al. 2009) for our Bayesian analyses. PhyloBayes was ran with the commands "pb -d {*Alignment*} -lg -nchain 2 100 0.3 100 {*Output*}", which means that the program ran two chains in parallel, checking every 100 cycles to see if all discrepancies between the two chains were less than or equal to 0.3, and that all effective sizes were larger than 100. The runs were automatically stopped once these conditions were met.

Gene Fishing

Animals were starved for at least 48 hours prior to sampling. Genomic DNA was extracted using either a classic C-Tab protocol (Bebenek et al. 2004) or the DNeasy Kit (Qiagen). Degenerate PCR primers were designed to capture conserved regions of the POU_S and POU_{HD} domains (F1:CAA GCA GMG RMG VAT MAA RYT RGG; F2: CTB ACB YTB TCV CAY AAC AAC ATG; R1: CKY TTY TCN GGH GCV GCR ATR S; R2: RTT RCA RAA CCA SAC BCK MAC MAC). For each gene recovered, we used BLAST as well as phylogenetic analysis (Figure S8) to assign a family identity to each gene. These family identities were supported with MUSCLE-based alignments of the linker regions, performed in Geneious (Figure S7).

Ancestral State Reconstruction

Accurate ancestral state reconstruction requires a gene tree that is consistent with the species tree, which is not generally the result of a standard ML or Bayesian analysis. To generate a gene tree informed by the species tree, we created an additional topology using TreeBeST (Vilella et al. 2008). Because of uncertainties in the topology at the base of the animal tree, we removed *Oscarella*, *Mnemiopsis*, and *Trichoplax* from our ancestral state reconstruction. We invoked the commands "treebest best -f {*Input tree*} -o {*Output tree*} {*Alignment*}", which resulted in a gene tree that was reconciled with the species tree, rooted by minimizing the number of duplications and losses, and bootstrapped 100 times.

The output of TreeBeST did a good job at creating a gene tree that was consistent with the species tree, with one exception. It produced a topology in the *POU6* family where all bilaterian invertebrate *POU6* genes were derived from one of the two vertebrate

homologs (data not shown). This scenario would require a duplication of *POU6* at the base of the bilaterians, with the same paralog being lost in every invertebrate clade. A more likely scenario is that there was a single *POU6* gene in invertebrate bilaterians, and this gene duplicated in the vertebrates; a scenario that occurred in *POU2*, *POU3*, and *POU4* families. To modify the TreeBeST topology and get adjusted initial branch lengths, we ran the original POU alignment through BEUTi/BEAST (Drummond et al. 2012) for 500,000 generations, constraining every node as a prior to reflect the TreeBeST topology with our modification. For this analysis, we ultimately decided to exclude *Amphimedon POUB* and a *Nematostella POU3* paralog, since both sequences were highly derived, and we wished to avoid biasing our ancestral states with these sequences. However, it is worth noting that when *Amphimedon POUB* was included in the TreeBest analysis, it grouped with *POU4*. The final tree used for ancestral state reconstruction is available as Supplementary File S2.

The modified consensus tree and the relevant protein alignment were imported into the FastML server (Ashkenazy et al. 2012), using the LG substitution model, optimization of branch lengths, and gamma distribution options. The probabilities of the ancestral POU sequence were graphically exported using the WebLogo (Crooks et al. 2004) function in FastML, and re-colored in Adobe Illustrator to be consistent with MacClade-style amino acid coloration (as seen in Figure 1). The most probable ancestral state at each relevant node was exported from the FastML output, and amino acid substitutions were determined manually.

Tests of Asymmetric Functional Divergence

We tested for functional divergence following gene duplication using the DIVERGE (v3.0) package (Gu et al. 2013). The tree used for ancestral state reconstruction (Supplementary File S2) and the relevant sequences were imported into DIVERGE to calculate the coefficient of functional divergence (or θ) for each pairwise comparison between POU families. We performed tests for type-I functional divergence (differences in amino acid variability between POU families) and type-II functional divergence (differences in significant amino acid substitutions between families, using the "significance" criteria described earlier). Z-values were calculated by dividing θ by the standard error, and p-values were determined using a two tailed Z-score test (normal distribution test). The results of all tests are available in Figure S11.

ACKNOWLEDGMENTS

We would like to thank Janet Sinsheimer, David Plachetzki, and Ryan Ellingson for helpful advice on this project. This work was supported by a National Institutes of Health Training Grant in Genomic Analysis and Interpretation T32HG002536 (D.A.G.) and the National Aeronautics and Space Administration Astrobiology Program (D.K.J.). New sequences reported in this paper have been deposited in GenBank under the accession numbers KJ632362 - KJ632398.

REFERENCES

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40:W580–W584.

Assa-Munt N, Mortishire-Smith RJ, Aurora R, Herr W, Wright PE. 1993. The solution

structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage λ repressor DNA-binding domain. Cell 73:193–205.

Backfisch B, Veedin Rajan VB, Fischer RM, Lohs C, Arboleda E, Tessmar-Raible K, Raible F. 2013. Stable transgenesis in the marine annelid Platynereis dumerilii sheds new light on photoreceptor evolution. Proc Natl Acad Sci U.S.A. 110:193–198.

Bebenek IG, Gates RD, Morris J, Hartenstein V, Jacobs DK (2004). sine oculis in basal Metazoa. Dev Genes Evol. 214:342-351.

Brusca RC, Brusca GJ. 2003. Invertebrates. 2nd ed. Sinauer Associates

Bürglin TR, Ruvkun G. 2001. Regulation of ectodermal and excretory function by the C. elegans POU homeobox gene ceh-6. Development 128:779–790.

Candiani S, Holland ND, Oliveri D, Parodi M, Pestarino M. 2008. Expression of the amphioxus Pit-1 gene (AmphiPOU1F1/Pit-1) exclusively in the developing preoral organ, a putative homolog of the vertebrate adenohypophysis. Brain Res Bull. 75:324–330.

Cary LR. 1916a. The Influence of the Marginal Sense Organs on Metabolic Activity in Cassiopea Xamachana Bigelow. Proc Natl Acad Sci U.S.A. 2:709–712.

Cary LR. 1916b. The influence of the marginal sense organs on the rate of regeneration in Cassiopea xamachana. J Exp Zool. 21:1–32.

Chavez M, Landry C, Loret S, et al. 1999. APH-1, a POU homeobox gene expressed in the salt gland of the crustacean Artemia franciscana. Mech Dev. 87:207–212.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A Sequence Logo Generator. Genome Res. 14:1188–1190.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Degnan BM, Vervoort M, Larroux C, Richards GS. 2009. Early evolution of metazoan transcription factors. Curr. Opin. Genet. Dev. 19:591–599.

Dekker N, Cox M, Boelens R, Verrijzer CP, van der Vliet PC, Kaptein R. 1993. Solution structure of the POU-specific DNA-binding domain of Oct-1. Nature 362:852–855.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29:1969–1973.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Ellwanger K, Nickel M. 2006. Neuroactive substances specifically modulate rhythmic

body contractions in the nerveless metazoon Tethya wilhelma (Demospongiae, Porifera). Front Zool. 3:7.

Erkman L, McEvilly RJ, Luo L, et al. 1996. Role of transcription factors a Brn-3.1 and Brn-3.2 in auditory and visual system development. Nature 381:603–606.

Esch D, Vahokoski J, Groves MR, et al. 2013. A unique Oct4 interface is crucial for reprogramming to pluripotency. Nat Cell Biol. 15:295–301.

Fernandez-Tresguerres B, Cañon S, Rayon T, Pernaute B, Crespo M, Torroja C, Manzanares M. 2010. Evolution of the mammalian embryonic pluripotency gene regulatory network. Proc Natl Acad Sci U.S.A. 107:19955–19960.

Flatt T. 2005. The evolutionary genetics of canalization. Q Rev Biol. 80:287–316

Frankenberg S, Pask A, Renfree MB. 2010. The evolution of class V POU domain transcription factors in vertebrates and their characterisation in a marsupial. Dev Biol. 337:162–170.

Frankenberg S, Renfree MB. 2013. On the origin of POU5F1. BMC Biol. 11:56.

Friedrich RP, Schlierf B, Tamm ER, Bösl MR, Wegner M. 2005. The class III POU domain protein Brn-1 can fully replace the related Oct-6 during schwann cell development and myelination. Mol Cell Biol. 25:1821–1829.

Gibson G, Wagner G. 2000. Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22:372–380.

Gold DA, Jacobs DK. 2013. Stem cell dynamics in Cnidaria: are there unifying principles? Dev Genes Evol. 223:53–66.

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. 2013. An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family. Mol Biol Evol. 30:1713-1719.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 59:307–321.

Hemmrich G, Bosch TCG. 2008. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. Bioessays 30:1010–1018.

Herr W, Sturm RA, Clerc RG, et al. 1988. The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. Genes Dev. 2:1513–1516.

Hroudova M, Vojta P, Strnad H, Krejcik Z, Ridl J, Paces J, Vlcek C, Paces V. 2012.

Diversity, phylogeny and expression patterns of Pou and Six homeodomain transcription factors in hydrozoan jellyfish Craspedacusta sowerbyi. PLoS ONE 7:e36420.

Ingraham HA, Chen R, Mangalam HJ, Elsholtz HP, Flynn SE, Lin CR, Simmons DM, Swanson L, Rosenfeld MG. 1988. A tissue-specific transcription factor containing a homeodomain specifies a pituitary phenotype. Cell 55:519–529.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:4–108.

Jacobs DK, Gates RD. 2001. Evolution of POU/homeodomains in basal metazoa: Implications for the evolution of sensory systems and the pituitary. Dev Bio. 235: 241-241.

Jacobs DK, Gates RD. 2003. Developmental genes and the reconstruction of metazoan evolution—implications of evolutionary loss, limits on inference of ancestry and type 2 errors. Int Comp Bio. 43:11-18.

Jacobs D, Nakanishi N, Yuan D. 2007. Evolution of sensory structures in basal metazoa. Int Comp Bio. 47:712–723.

Jacobs DK, Gold DA, Nakanishi N, Yuan D, Camara A, Nichols SA, Hartenstein V. 2010. Basal Metazoan Sensory Evolution. In: DeSalle R, Schierwater B, editors. Key Transitions in Animal Evolution. Key Transitions in Animal Evolution. Science Publishers. pp. 175–196.

Jacobson EM, Li P, Leon-del-Rio A, Rosenfeld MG, Aggarwal AK. 1997. Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. Genes Dev. 11:198–212.

Jauch R, Choo SH, Ng CKL, Kolatkar PR. 2010. Crystal structure of the dimeric Oct6 (POU3f1) POU domain bound to palindromic MORE DNA. Proteins: Struct, Funct, Bioinf. 79:674–677.

Kamm K, Schierwater B. 2007. Ancient linkage of a POU class 6 and an anterior Hox-like gene in cnidaria: implications for the evolution of homeobox genes. J Exp Zool B Mol Dev Evol. 308:777–784.

Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. 1994. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell 77:21–32.

Konzak KE, Moore DD. 1992. Functional isoforms of Pit-1 generated by alternative messenger RNA splicing. Mol Endocrinol. 6:241–247.

Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. 2008. Genesis and Expansion of Metazoan Transcription Factor Gene Classes. Mol Biol

Evol. 25:980–996.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lavial F, Acloque H, Bertocchini F, et al. 2007. The Oct4 homologue PouV and Nanog regulate pluripotency in chicken embryonic stem cells. Development 134:3549–3563.

Lee M-H, Salvaterra PM. 2002. Abnormal Chemosensory Jump 6 Is a positive transcriptional regulator of the cholinergic gene locus in Drosophila olfactory neurons. J Neurosci. 22:5291–5299.

Leys SP, Mackie GO, Meech RW. 1999. Impulse conduction in a sponge. J Exp Biol. 202 (Pt 9):1139–1150.

Leys SP, Meech RW. 2006. Physiology of coordination in sponges. Can J Zool. 84:288–306.

Li S, Crenshaw EB III, Rawson EJ, Simmons DM, Swanson LW, Rosenfeld MG. 1990. Dwarf locus mutants lacking three pituitary cell types result from mutations in the POU-domain gene pit-1. Nature 347:528–533.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151–1155.

Meier S, Sprecher SG, Reichert H, Hirth F. 2006. ventral veins lacking is required for specification of the tritocerebrum in embryonic brain development of Drosophila. Mech Dev. 123:76–83.

Millane RC, Kanska J, Duffy DJ, Seoighe C, Cunningham S, Plickert G, Frank U. 2011. Induced stem cell neoplasia in a cnidarian by ectopic expression of a POU domain transcription factor. Development 138:2429–2439.

Morrison GM, Brickman JM. 2006. Conserved roles for Oct4 homologues in maintaining multipotency during early vertebrate development. Development 133:2011–2022.

Müller WEG, Wendt K, Geppert C, Wiens M, Reiber A, Schröder HC. 2006. Novel photoreception system in sponges? Unique transmission properties of the stalk spicules from the hexactinellid Hyalonemasieboldi. Biosens Bioelectron. 21:1149–1155.

Nakanishi N, Yuan D, Hartenstein V, Jacobs DK. 2010. Evolutionary origin of rhopalia: insights from cellular-level analyses of Otx and POU expression patterns in the developing rhopalial nervous system. Evol Dev. 12:404–415.

Ng H-H, Surani MA. 2011. The transcriptional and signalling networks of pluripotency. Nature 13:490–496.

Nickel M. 2004. Kinetics and rhythm of body contractions in the sponge Tethya wilhelma (Porifera: Demospongiae). J Exp Biol. 207:4515–4524.

Niwa H, Smith AG, Miyazaki J-I. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation, or self-renewal of ES cells. Nat Genet. 24:372–376.

Nomaksteinsky M, Kassabov S, Chettouh Z, Stoeklé H-C, Bonnaud L, Fortin G, Kandel ER, Brunet J-F. 2013. Ancient origin of somatic and visceral neurons. BMC Biol. 11:53.

O'Brien EK, Degnan BM. 2002. Pleiotropic developmental expression of HasPOU-III, a class III POU gene, in the gastropod Haliotis asinina. Mech Dev. 114:129–132.

Onal P, Grün D, Adamidi C, et al. 2012. Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. The EMBO Journal 31:2755–2769.

Phillips K, Luisi B. 2000. The virtuoso of versatility: POU proteins that flex to fit. J Mol Biol.

Raible F, Tessmar-Raible K, Osoegawa K, et al. 2005. Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. Science 310:1325–1326.

Ramachandra NB, Gates RD, Ladurner P, Jacobs DK, Hartenstein V. 2002. Embryonic development in the primitive bilaterian Neochildia fusca: normal morphogenesis and isolation of POU genes Brn-1 and Brn-3. Dev Genes Evol. 212:55–69.

Reményi A, Lins K, Nissen LJ, Reinbold R, Schöler HR, Wilmanns M. 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev. 17:2048–2059.

Reményi A, Tomilin A, Pohl E, Lins K, Philippsen A, Reinbold R, Schöler HR, Wilmanns M. 2001. Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. Mol Cell 8:569–580.

Rodda DJ, Chew J-L, Lim L-H, Loh Y-H, Wang B, Ng H-H, Robson P. 2005. Transcriptional Regulation of Nanog by OCT4 and SOX2. J Biol Chem. 280:24731–24737.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5:725–738.

Ryan A, Rosenfeld M. 1997. POU domain family values: flexibility, partnerships, and developmental codes. Genes Dev. 11:1207–1225.

Ryan JF, Pang K, Comparative Sequencing Program N, Mullikin JC, Martindale MQ, Baxevanis AD. 2010. The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa.

EvoDevo 1:9.

Sakarya O, Armstrong KA, Adamska M, Adamski M, Wang I-F, Tidor B, Degnan BM, Oakley TH, Kosik KS. 2007. A Post-Synaptic Scaffold at the Origin of the Animal Kingdom. PLoS ONE 2:e506.

Scully KM, Jacobson EM, Jepsen K, et al. 2000. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. Science 290:1127–1131.

Seimiya M, Watanabe Y, Kurosawa Y. 1997. Identification of POU-class homeobox genes in a freshwater sponge and the specific expression of these genes during differentiation. Eur J Biochem. 243:27–31.

Shah D, Aurora D, Lance R, Stuart GW. 2000. POU Genes in Metazoans: Homologs in Sea Anemones, Snails, and Earthworms. DNA Seq. 11:457–461.

Simmons D, Voss J, Ingraham H. 1990. Pituitary cell phenotypes involve cell-specific Pit-1 mRNA translation and synergistic interactions with other classes of transcription factors. Genes Dev. 4:695–711.

Steele RE, David CN, Technau U. 2011. A genomic view of 500 million years of cnidarian evolution. Trends in Genetics 27:7–13.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676.

Theodorou E, Dalembert G, Heffelfinger C, White E, Weissman S, Corcoran L, Snyder M. 2009. A high throughput embryonic stem cell screen identifies Oct-2 as a bifunctional regulator of neuronal differentiation. Genes Dev. 23:575–588.

Tichy AL, Ray A, Carlson JR. 2008. A new Drosophila POU gene, pdm3, acts in odor receptor expression and axon targeting of olfactory neurons. J Neurosci. 28:7121–7129.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.

Voss JW, Wilson L, Rosenfeld MG. 1991. POU-domain proteins Pit-1 and Oct-1 interact to form a heteromeric complex and can cooperate to induce expression of the prolactin promoter. Genes Dev. 5:1309–1320.

Waddington CH. 1942. Canalization of development and the inheritance of acquired characters. Nature 150:563-565.

Zhang T-Y, Xu W-H. 2009. Identification and characterization of a POU transcription factor in the cotton bollworm, Helicoverpa armigera. BMC Mol Biol. 10:25.

Chapter 2: The Homeodomain Complement of the moon jellyfish *Aurelia*: Differential Gene Duplication, and the Disconnect Between Life History and Genetic Complexity in the Cnidaria

ABSTRACT

Using genomic, transcriptomic, and RNA-Seq resources, we have annotated the homeodomain repertoire for the moon jellyfish *Aurelia sp.1*, and compared it to data from cnidarian relatives that lack a medusa life stage (the sea anemone *Nematostella*, the coral *Acropora*, and the hydroid *Hydra*). Cnidarian homeodomains can be subdivided into 66 bilaterian families encompassing nine classes, providing a significant upwards revision for the homeodomain complement of the last common ancestor of cnidarians and bilaterians. Despite having simpler life cycles and bodyplans, the anthozoans *Nematostella* and *Acropora* have far more homeodomains than *Aurelia* (149, 127, and 99 respectively). While each cnidarian lineage exhibits a unique pattern of gene gain and loss, these larger gene counts in the Anthozoa are primarily caused by clade-specific gene expansions. The one exception to this trend is the non-anterior Hox genes, where *Aurelia* has seven paralogs compared to *Nematostella* and *Acropora's* two. RNA-Seq analyses suggest that these non-anterior Hox genes are expressed dynamically through the *Aurelia* life cycle, and therefore represent candidate sequences for future studies in medusozoan bodyplan evolution. Comparisons of gene expression between *Aurelia* and *Nematostella* during polyp formation suggest that taxon-specific gene duplications often take on opposing expression patterns during development. Even in genes that lack paralogs, *Aurelia* and *Nematostella* can exhibit opposing expression dynamics, suggestive of cryptic differences in cnidarian development.

INTRODUCTION

A goal of comparative genomics is to decipher the causal connections between genome composition and animal form. To this end, the Cnidaria (sea anemones, corals, hydras, and jellyfish) hold a valuable place in comparative studies. Phylogenetic analyses consistently support the Cnidaria as the major sister clade to the bilaterians (protostomes plus deuterostomes), which encompasses 99% of living animals (Pick et al. 2010; Philippe et al. 2011; Nosenko et al. 2013; Ryan et al. 2013). The cnidarians represent over 10,000 species, with a wide range of morphologies, ecologies, and life histories. This disparity is generated from the combination of polyp and medusa bodyplans (Figure 1), the former representing a sessile and structurally simple life stage, and the later a free-swimming organism often equipped with neural and sensory structures that rival many bilaterians. Genetic and fossil evidence suggests that the medusa-bearing cnidarians (medusozoans) diverged from their morphologically simpler relatives (the anthozoans) before the Cambrian "explosion" of bilaterian animals ~542 million years ago, meaning the cnidarian radiation is one of the earliest examples of the evolution of complex animal forms (Figure 1A; Erwin et al. 2011).

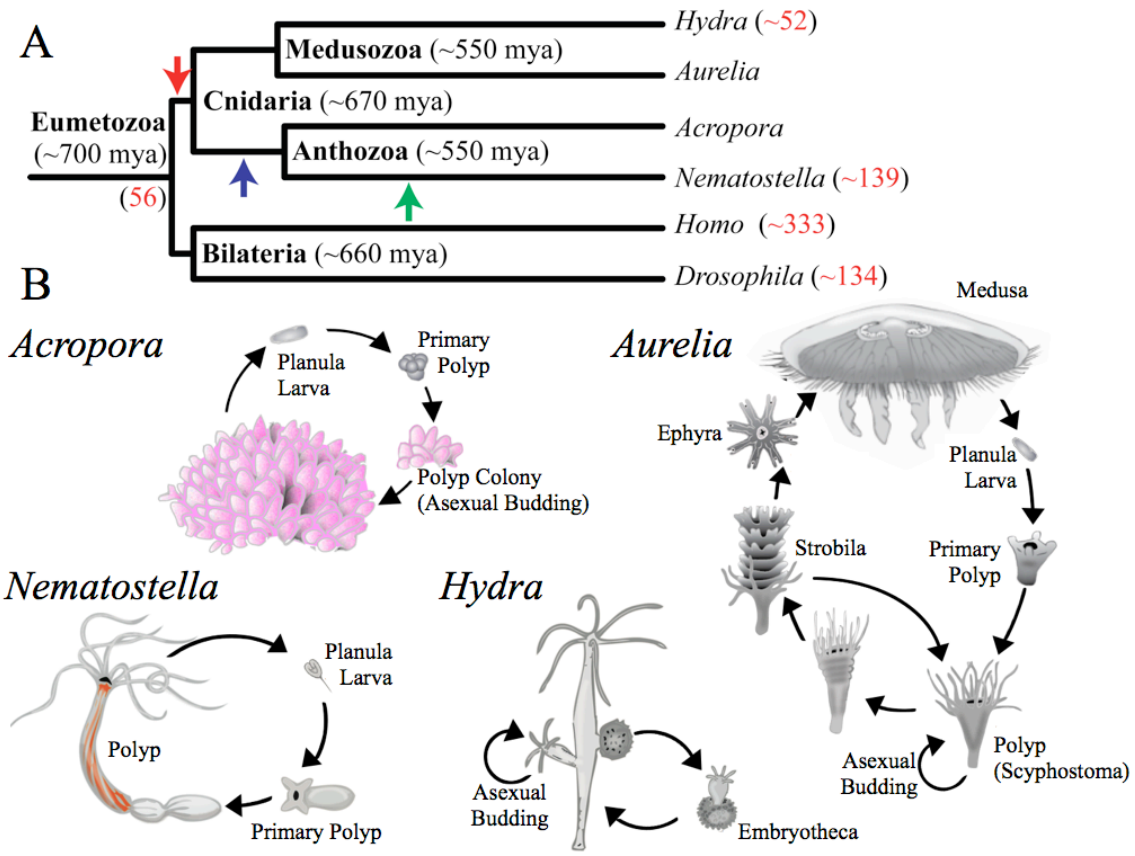

Figure 1: The phylogenetic placement and life cycle of cnidarians in this study.

(A) Time-calibrated phylogeny of the four cnidarians considered in this study. Homeodomain estimates from previous studies are presented in red, based on Chourrout et al. (2006), Ryan et al. (2006), and Steele et al. (2011). Arrows represent points in the phylogeny where potential homeobox expansions could have occurred: at the base of the Cnidaria (red arrow), before the divergence of anthozoan hexacorals (blue arrow), or following the divergence of *Nematostella* from *Acorpora* (green arrow). Divergence time estimates are based on (Erwin et al. 2011; Park et al. 2012). (B) Life cycles of cnidarians involved in this study.

Publically available cnidarian genomes include the sea anemone *Nematostella vectensis* (Putnam et al. 2007), the coral *Acropora digitifera* (Shinzato et al. 2011), and the hydroid *Hydra vulgaris* (formally *Hydra magnipapillata*; Chapman et al. 2010). Unfortunately, none of these taxa exhibit a medusa life stage. Although *Hydra* is a part of the

Medusozoa, it has undergone significant simplification over the course of its evolution, and has subsequently lost both planula and medusa stages (Figure 1B; Collins et al. 2006).

To address this gap, we are assembling the genome and developmental transcriptome of the moon jellyfish *Aurelia species 1* (*sensu* Dawson and Jacobs 2001). The genome of *Aurelia* has an estimated size of ~0.7 Gb (C-value = 0.73pg; Goldberg et al. 1975), which falls within the range of *Nematostella* (~0.45 Gb) and *Hydra* (~1 Gb) (Steele et al. 2011). Our current assembly of the *Aurelia* genome consists of 29,729 contigs with an N50 of 16,820bp. We have augmented this genome with an extensive transcriptome that covers the major life stages (described in Chapter 3). *Aurelia* offers a tractable laboratory model, and a valuable addition to comparative genomics. It is a member of the medusozoan class Scyphozoa, which represents the probable sister clade to *Hydra* and its relatives (the Hydrozoa) (Collins 2002; Dawson 2004; Collins et al. 2006; see Kayal et al. 2013 for an alternative phylogeny). The *Aurelia* medusa is a free-swimming carnivore, featuring complex neural and sensory system architecture that culminates in eight structures called rhopalia, which circle the medusa's bell. Rhopalia neurally integrate and coordinate several sensory structures—including an eye-cup, a mechanosensory touch plate, and a geosensory statocyst—and is patterned using several genes involved in bilaterian sensory organogenesis (Horridge 1956; Nakanishi et al. 2009; Nakanishi et al. 2010). No comparable sensory structures exist in *Nematostella*, *Acropora*, or *Hydra*. Thus, despite the inherent difficulties in defining biological complexity, *Aurelia* clearly meets McShea's (1996) definitions for increased morphological complexity (i.e. nonhierarchical

object complexity) as well as increased developmental complexity (i.e. nonhierarchical process complexity) when compared to *Nematostella*, *Acropora*, or *Hydra*.

It is less clear if *Aurelia's* morphological and developmental complexity correlates with genomic complexity. To begin addressing this question, we used phylognetic reconstruction and RNA-Seq to analyze the homeobox genes, a large clade of transcription factors that share a ~180 bp DNA binding region called the homeodomain (Scott and Weiner 1984). Early studies of *Drosophila* revealed that mutations to certain homeobox genes result in homeosis, or the transformation of one organ type into another (Schneuwly et al. 1987). Since then, the homeoboxes have been primary candidates in the study of animal body-plan evolution (Holland et al. 2007), and are a common starting point when analyzing the genomes of early-branching animal lineages (Ryan et al. 2006; Srivastava et al. 2008; Ryan et al. 2010; Srivastava et al. 2010). The homeobox complements of *Nematostella* and *Hydra* have been explored in previous studies (Chourrout et al. 2006; Ryan et al. 2006); *Nematostella* has many more homeodomains than *Hydra* (Figure 1), and since *Hydra* is missing many homeobox families that *Nematostella* shares with bilaterians, it is assumed that *Hydra* has experienced significant gene loss (Chourrout et al. 2006). However, *Nematostella* also has many paralogs within families, with upwards of 74 additional homeoboxes compared to the last common ancestor of cnidarians and bilaterians (Figure 1A; Ryan et al. 2006). It is currently unclear whether this expansion of homeoboxes occurred at the base of the cnidarian tree, or at some point during the divergence of *Nematostella* from other cnidarians (arrows in

Figure 1A). Determining when this gene radiation occurred will impact our interpretation of the role homeobox genes in this early phase of animal evolution.

RESULTS

The last common ancestor of cnidarians and bilaterians had at least 65 homeoboxes encompassing nine classes.

We queried the genomes and proteomes of *Nematostella*, *Acropora*, *Aurelia*, and *Hydra* using multiple homeodomains (see Materials and Methods). By incorporating both genomic and peptide datasets into our analyses, we recovered more homeoboxes from *Nematostella* than previously reported. We were also able to resolve the family-level affinity of more cnidarian genes than previous studies, in part because of our increased taxon sampling. Animal homeodomains are typically divided into eleven classes: ANTP, PRD, TALE, POU, CERS/LASS, PROS, ZF, LIM, HNF, CUT, and SINE (Zhong et al. 2008). We did not recover any ZF or PROS-like homeodomains in the cnidarians, consistent with previous studies (Chourrout et al. 2006; Ryan et al. 2006). Since these classes are also absent from sponge, placozoan, and ctenophore genomes (Ryan et al. 2010), it appears likely that ZF and PROS represent bilaterian novelties. There has been some uncertainty as to whether *Nematostella* has a genuine CUT gene (Ryan et al. 2006; Ryan et al. 2010), but our results confirm that *Nematostella*, *Acropora*, and *Aurelia* all have a homolog of the *onecut* family of CUT-class homeoboxes (Table 1). Not only do the homeodomains from these three genes clade with bilaterian *onecut* sequences, but all three proteins contain a recognizable Cut domain upstream of the homeodomain (Supplementary File 1, Part 4) We also recovered CERS/LASS-class genes from all four

cnidarians, suggesting that nine of eleven homeobox classes were present in the last common cnidarian ancestor. Most cnidarian genes could be further subdivided into 66 bilaterian families (Table 1), providing an upwards revision of the minimal homeobox complement of the last common ancestor for bilaterians and cnidarians (the Eumetaozoa; see Figure 1A).

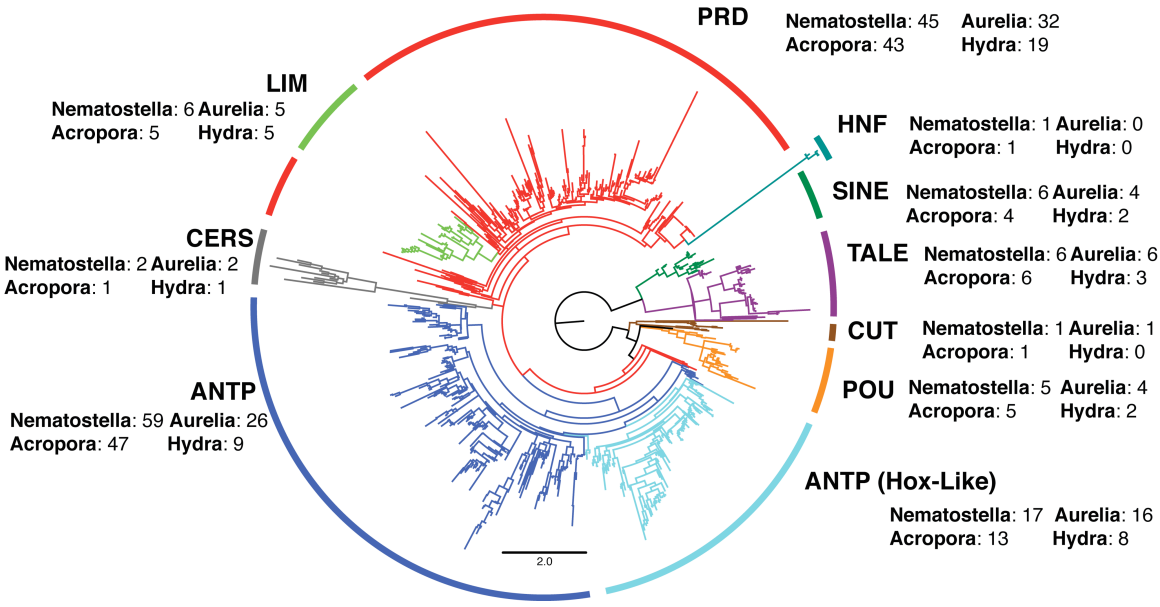

Figure 2. PhyML maximum likelihood tree of cnidarian and bilaterian homeoboxes. See Supplementary File 1 (Part 6) for full trees with support values.

Table 1: Homeobox annotations and counts for *Acropora*, *Nematostella*, *Aurelia*, and *Hydra*.

Class	Family (+ Drosophila homolog)	Acropora	Nematostella	Aurelia	Hydra
ANTP (HOX)					
	Cdx (*cad*)	0	0	1	1
	"CnidHox"	2	2	4	4
	Evx (*eve*)	1	1	1	0
	Gbx (*unpg*)	1	1	0	0
	Gsx (*ind*)	1	1	1	1
	HOX1 (*lab*)	1	1	1	1
	HOX2	1	2	0	0
	"Scox" / HOX9-13/15 (*Abd-B*)	0	0	3	0
	Meox (*btn*)	3	4	3	1
	Mnx (*exex*)	1	1	0	0
	Ro (*ro*)	0	1	0	0
	Xlox/Pdx	0	1	1	0
	Unclassified	2	2	1	0
	Total	**13**	**17**	**16**	**8**
ANTP (OTHER)					
	Barx/Bsx (*bsh*)	4	4	1	0
	Bari (*CG11085*)	9	6	3	0
	Dlx (*Dll*)	1	1	2	2
	EMX (*Es*, *ems*)	2	2	1	0
	Hhex (*CG7056*)	1	1	1	1
	Abox/Dbx/Hxlx	3	8	1	0
	Lbx/Ventx (*lbe*, *lbl*)	2	2	0	0
	Msx (*Dr*)	1	1	1	1
	Mxlx (CG1696)	2	2	1	0
	Nkx1 (*slou*)	1	1	1	1
	Nkx2 (*scro*, *vnd*)	5	6	2	1
	Nkx3 (*bap*)	1	1	2	0
	Nkx4 (*tin*)	1	1	1	1
	Nkx5 (*Hmx*)	1	1	1	0
	Nkx6 (*Hgtx*)	1	1	1	0
	Nkx7 (*Nk7.1*)	1	1	1	0
	Vax	1	2	1	0
	Nedx (*CG13424*)	2	2	2	0
	Noto (*CG18599*)	8	6	1	1
	Unclassified	0	10	2	1
	Total	**47**	**59**	**26**	**9**
CERS					
	Cers/Lag (*Lag1*)	1	2	2	1
CUT					
	Onecut (*ct*)	1	1	1	0
HNF					
	Hnf1/2	1	1	0	0
LIM					
	ISL	1	1	1	0
	Lhx1/5 (*Lim1*)	1	1	1	2
	Lhx2/9 (*ap*)	1	1	1	1
	Lhx6/8 (*Awh*)	1	1	1	1

	Lmx (*CG4328*, *CG32105*)	1	1	1	1
	Unclassified	0	1	0	0
	Total	**5**	**6**	**5**	**5**
POU					
	POU1	1	1	1	0
	POU3 (*vvl*)	2	2	1	0
	POU4 (*acj6*)	1	1	1	1
	POU6 (*pdm3*)	1	1	1	1
	Total	**5**	**5**	**4**	**2**
PRD					
	Alx	1	1	1	0
	Arx (*al*, *php13*)	1	2	0	0
	DMBX	7	7	2	0
	GSC (*Gsc*)	1	1	2	1
	Hbn (*hbn*)	1	1	2	1
	Leutx	2	1	0	0
	Nobox	1	1	2	2
	Otp (*otp*)	1	1	1	3
	Otx (*oc*)	6	4	6	3
	Pax3/7 (*gsb*, *prd*)	3	3	1	0
	Pax4/6 (*ey*, *toy*, *toe*, *eyg*)	1	1	1	0
	Pitx (*Ptx1*)	1	1	1	1
	Prop (*CG32532*)	1	1	0	0
	Rax (*Rx*)	1	1	1	0
	Repo (*repo*)	1	1	2	1
	Shox (*CG34367*)	1	1	1	0
	Uncx (*OdsH*, *unc-4*)	3	4	1	1
	Vsx (*Vsx1*, *Vsx2*, *tup*)	1	1	1	0
	Unclassified	12	13	5	4
	Total	**43**	**45**	**32**	**19**
TALE					
	Irx (*mirr*, *ara*, *caup*)	1	1	1	1
	Meis (*hth*)	1	1	2	1
	Pbx (*exd*)	1	1	1	1
	Pknox	1	1	1	0
	Tgif (*achi*, *vis*)	1	1	0	0
	Unclassified	1	1	1	0
	Total	**6**	**6**	**6**	**3**
SINE					
	Cnidarian-Sine	1	1	1	0
	Six1/2 (*so*)	1	1	1	0
	Six3/6 (*Optix*)	1	1	1	1
	Six4/5 (*Six4*)	1	3	1	1
	Total	**4**	**6**	**4**	**2**
UNCLASSIFIED		0	0	2	2
TOTAL		127	149	99	51

The largest expansion of homeodomains occurred in the Anthozoa

Consistent with previous analyses, our results suggest that *Hydra* has undergone significant gene loss; specifically, *Hydra* appears to be missing 39 homeobox families that can be identified in other cnidarians. *Aurelia* also has far fewer homeoboxes than *Acropora* or *Nematostella*, despite its complex life history. This lower count is partially due to gene loss; *Aurelia* appears to be missing eight families retained in *Acropora* and/or *Nematostella* (Gbx, Rough, Lbx, hnf1/2, Leutx, Arx, Prop, and Tgif). However, most of the difference appears to be the result of gene family expansions that occurred within the Anthozoa. Only a handful of gene expansions predate the anthozoan-medusozoan split, including Otx, Meox, and Uncx families. In contrast, probable family expansions that occurred before the split of scleractinian (*Acropora*) and actiniarian (*Nematostella*) anthozoans include Dmbx, POU3, Barx, Bari, Nk2, Noto, as well as a massive radiation of PRD-class paralogs that cannot be classified into any known family. Following its divergence from *Acropora*, *Nematostella* appears to have had its own gene radiation in ANTP-class genes. In the *Nematostella* genome, many of these paralogs are closely associated (Figure 3), supporting the hypothesis that these represent multiple rounds of gene duplication events, as opposed to sequence convergence.

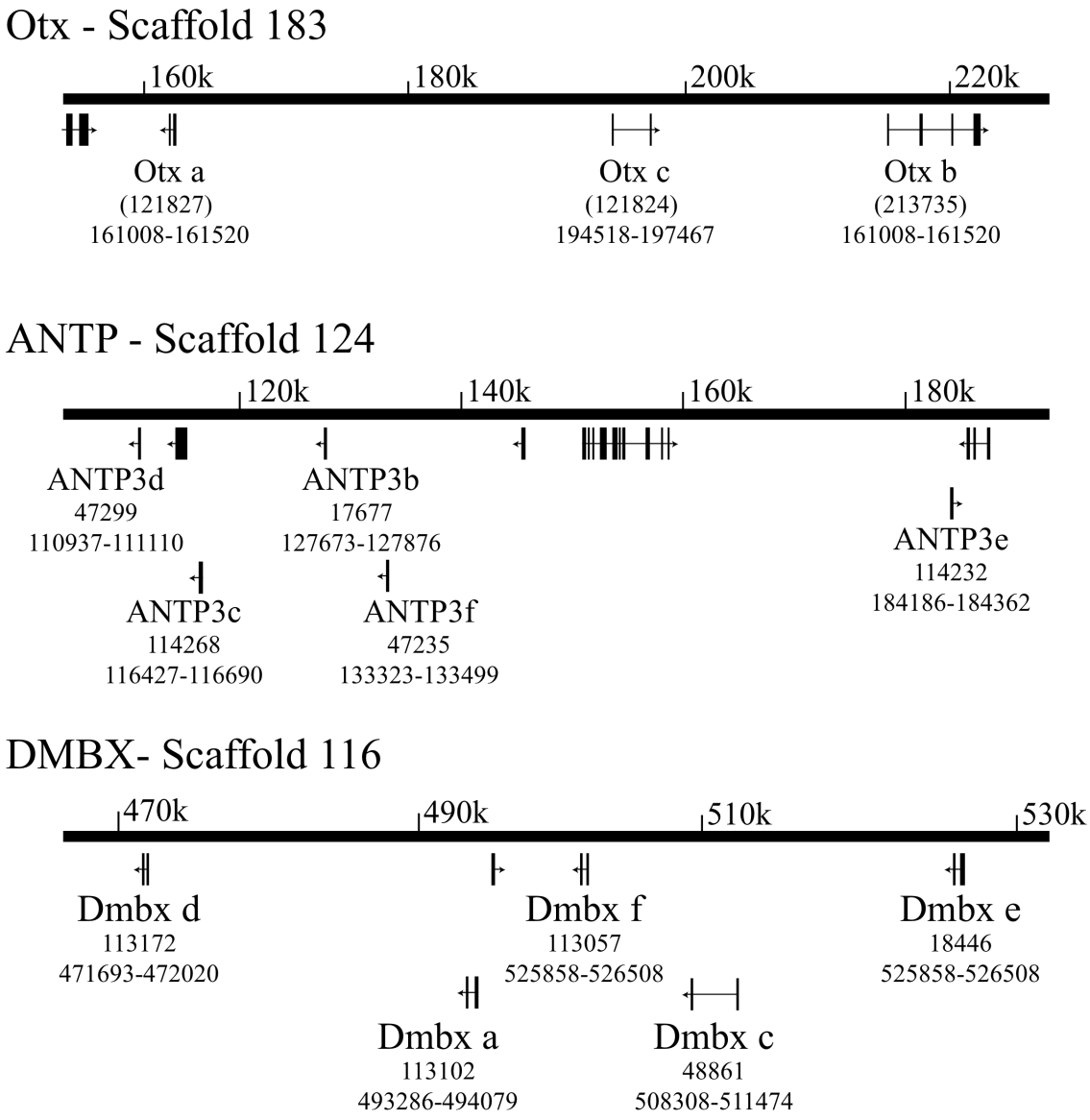

Figure 3. Several examples of paralog synteny in the *Nematostella* genome.

Non-anterior Hox genes represent the only major medusozoan gene expansion

The Hox-like ANTP genes were particularly volatile in our phylogenetic analyses, so we constructed a second alignment restricted to this clade, and including sequences from additional cnidarians and bilaterians (Figure 4, Supplementary File 1, Part 5). Some previous studies have suggested that cnidarians have anterior (Hox1-3) and posterior (Hox9-15) Hox genes, but no central Hox genes (Hox4-8) (Finnerty and Martindale

1999; Chiori et al. 2009). Cnidarians are also thought to have a proto-ParaHox cluster containing *Gsx* and a second gene, although there has been uncertainty as to whether that gene is homologous to *Xlox*, *Cdx*, or a common ancestor of the two (Chourrout et al. 2006). In contrast, our phylogenetic analyses unanimously support the presence of all three bilaterian Parahox genes in the Cnidaria (Figure 4 and Supplementary File 1). However, our topologies conflicted as to whether *Cdx* homologs are restricted to Medusozoa (Hox-specific trees suggest *Cdx* is medusozoan-specific, while full homeodomain trees suggest *Anthox6* genes are orthologous to *Cdx*). Additionally, all but one of our phylogentic trees support the hypothesis that the so-called Cnidarian posterior Hox-like genes (including *Nematostella Anthox1a/b*, *Clytia Hox9-14a/b/c*, and *Eleutheria Cnox1/3*) are in fact sister to a clade containing both posterior and central Hox genes. We do recover a cnidarian sister-clade for the posterior Hox genes, but it is restricted to the scyphozoans *Aurelia* and *Cassiopeia* (*Cassiopeia Scox1/4/5*; Kuhn et al. 1999). We recognize that this topology is phylogenetically implausible, for if we assume that the central and posterior Hox genes had a common ancestor, our topology suggests that the cnidarians had both the ancestral sequence and one of the daughter sequences. Two plausible evolutionary scenarios that are consistent with our topology: (1) non-anterior Hox genes evolved independently in Cndaria and Bilateria (i.e. the cnidarian "non-anterior" Hox genes in Chourrout et al. 2006), and a subset of scyphozoan homeodomains became “posteriorized” through convergent evolution, or (2) the last common ancestor of eumetozoa had both central and posterior Hox genes, and the posterior genes were lost multiple times in non-scyphozoan cnidarian clades. Given this uncertainty, we subsequently treat these genes collectively as “non-anterior” Hox genes, while

distinguishing between the “Scox” clade, which encompasses the six scyphozoan genes sister to the bilaterian Hox9-15 cluster in Figure 4, and the “CnidHox” clade, which encompasses all additional Cnidarian sequences sister to the bilateiran Cdx/Hox4-15 cluster. Whether or not the Scox and CnidHox genes are monophyletic, these results show that *Aurelia* has significantly more copies of non-anterior Hox genes compared to *Acropora* and *Nematostella* (seven, two, and two respectively).

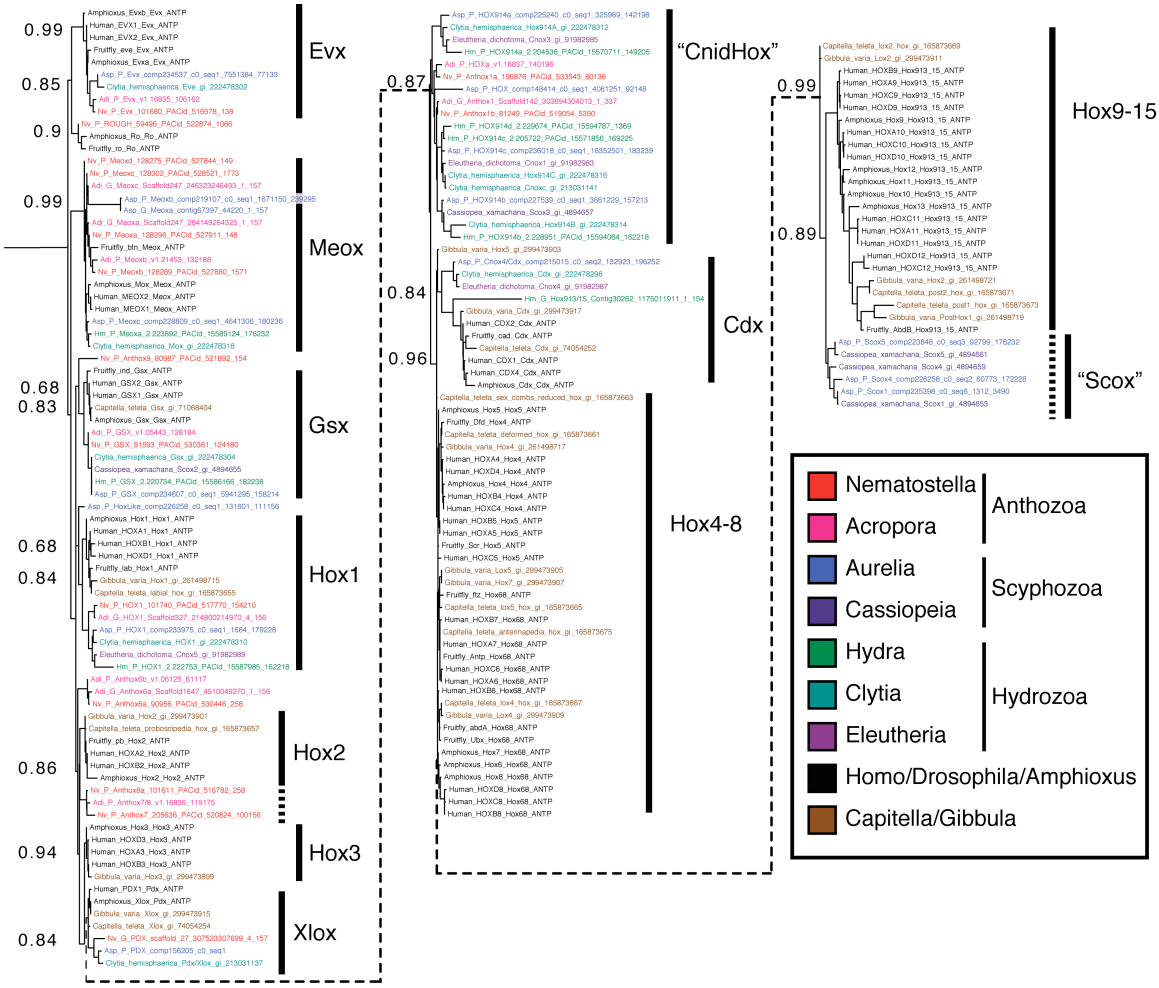

Figure 4: Phylogenetic tree of Hox and Parahox genes, with additional cnidarian and lophotrochozoan sequences. Select node probabilities are displayed, based on aLRT SH-Like scores in PhyML. The full tree with all node scores is available in Supplementary File 1 (Part 6).

Non-anterior Hox genes show dynamic changes in expression through the *Aurelia* life cycle

We used RNA-Seq to look at patterns of gene expression across seven time points in the *Aurelia* life cycle. Using the EdgeR/Bioconductor packages in Trinity (see Materials and Methods), we recovered 71 genes that exhibited at least one significant change in expression through time, which we clustered into six major clades based on the similarity of normalized expression patterns (Figure 5). Broadly, Clade I represents genes upregulated in the larval (pre-polyp) stages, Clade V genes show highest expression in the polyp, and Clade VI genes exhibit high expression in the medusa (post-polyp) stages. We also used EdgeR to look for significant differences in raw gene counts during the major transitions in the *Aurelia* life cycle (Figure 6).

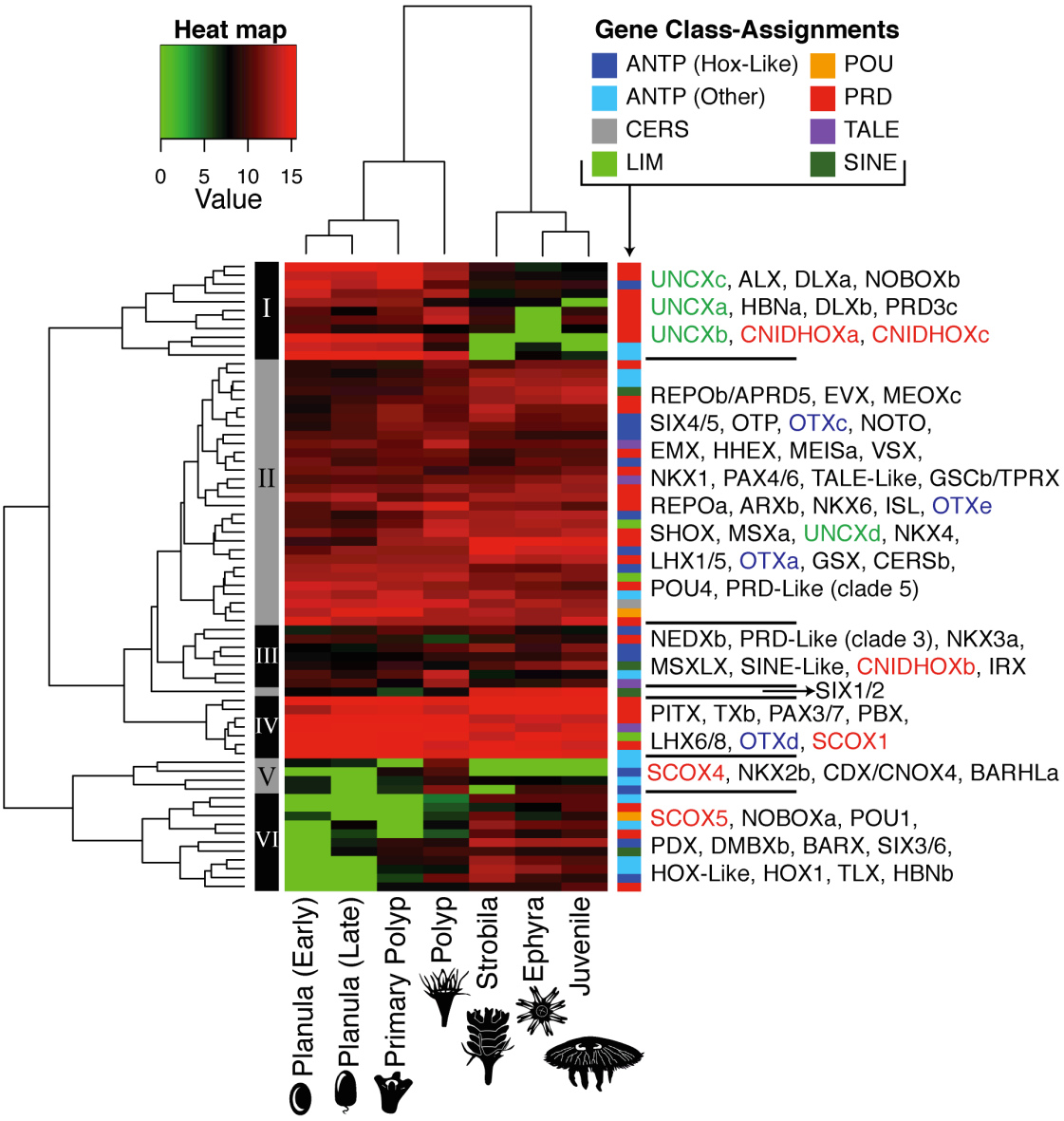

Figure 5: Heat map of FPKM-nromalized values (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) for *Aurelia* homeodomains. Gene names have been enlarged for legibility; for each cluster, genes are listed from left to right, starting with the top row. Non-anterior Hox paralogs are highlighted in red, Otx paralogs in blue, and Uncx paralogs in green. Genes were retained using an FDR-adjusted p-value cutoff of 0.05 and a minimum 2-fold change.

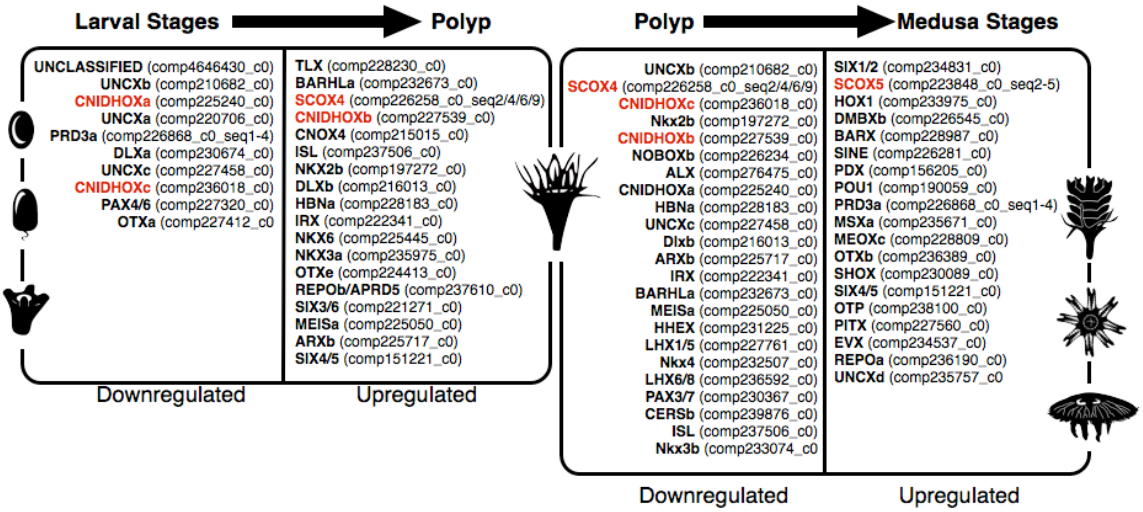

Figure 6. Significant differentially expressed genes during the metamorphosis of larval (pre-polyp) stages into the polyp, and the polyp into medusa (post-polyp) stages. Genes are listed in order of decreasing logfold change. Genes that are part of the "non-anterior" Hox-like clade are colored red. Significance is defined as FDR-adjusted p-values < 0.05, based on calculations performed in EdgeR.

We recovered nineteen genes that are significantly upregulated during *Aurelia's* medusa formation (Figure 6); many of these genes have a single paralog in other cnidarians, and may have been co-opted for novel functions during medusozoan bodyplan evolution. For example, several of the genes upregulated in the medusa—including *POU1/pit1*, *Otxb*, and *Six1/2*—have been shown to play a role in rhopalia formation using *in situ* hybridization (Nakanishi et al. 2010; Nakanishi et al. *submitted*).

Interestingly, the paralogous gene clades in *Aurelia* (Uncx, Otx, and the non-anterior Hox genes) show varying levels of expression pattern differentiation through the life cycle. Uncx paralogs show little temporal variation; three out of four paralogs are enriched in the early life stages (Clade I in Figure 5) and are significantly downregulated during

polyp formation (Figure 6). Otx genes tend to show high levels of expression across the life cycle according to the cluster analysis (Clade II in Figure 5), although the changes that occur are often significant (Figure 6). In contrast to Otx or Uncx, *Aurelia's* seven non-anterior Hox genes are scattered across five clades in the cluster analysis, and four out of six show significant up- or downregulation during the major transitions in *Aurelia's* bodyplan (red genes in Figure 6). If we accept the distinction between "CnidHox" and "Scox" clades (Figure 4), we still find a similar pattern of differential temporal expression of paralogs through the life cycle.

Comparison of gene expression between *Nematostella* and *Aurelia* suggests that multiple shifts in expression have occurred, particularly with taxon-specific paralogs

We also used EdgeR to compare gene expression during polyp formation in *Aurelia* with similar data previously reported from *Nematostella* (Helm et al. 2013; Figure 7). In some cases it is difficult to determine whether homologous genes in the two taxa show similar expression trends, as both occasionally use taxon-specific paralogs in contrasting roles (i.e. CnidHox and Otx paralogs in *Aurelia*, Nk2 paralogs in *Nematostella*; see the blue genes in Figure 7). There are a small number of differentially expressed genes with a single copy in both taxa (red genes in Figure 7), but these tell conflicting stories: while *Six3/6* is upregulated during polyp formation in both *Aurelia* and *Netmatostella*, *Irx* and *Hhex* show opposing regulatory patterns.

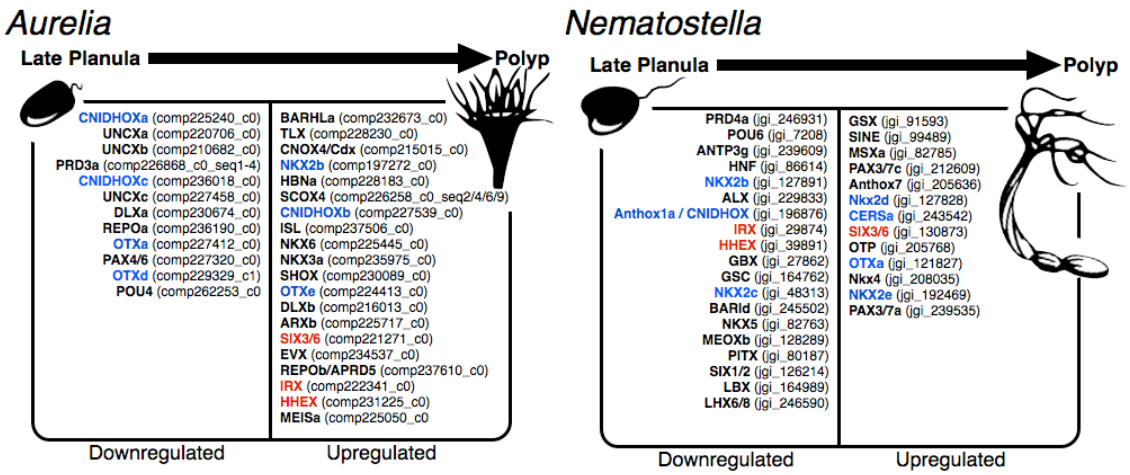

Figure 7. A comparison of differentially expressed genes during polyp formation between *Aurelia* and *Nematostella*. Genes found in both taxa but with multiple paralogs in at one taxon are labeled blue; genes found in both taxa without paralogs are labeled red. Because of low phylogenetic resolution, paralog annotations (i.e. Nkx2a versus Nkx2b) are assigned arbitrarily, and should not be used to infer paralog homology between taxa. Significance is defined as FDR-adjusted p-values < 0.05.

DISCUSSION

Figure 8 illustrates our working hypothesis regarding the evolution of animal homeoboxes. Assuming that the last common ancestor of Cnidaria had one copy of each family member, the phylum originated with more homeodomains than their earlier-branching cousins (the ctenohpore *Mnemiopsis leidyi*, the placozoan *Trichoplax adherans*, or the sponge *Amphimedon queenslandica*). Each cnidarian lineage exhibits a unique pattern of gene gain and loss, but the majority of gene expansions occurred within the *Anthozoa*. This result was unexpected, for despite a few aspects of the polyp endoderm (i.e. the presence of a pharynx/siphonoglyph, the subdivision of the coelenterons by mesenteries; see Daly et al. 2003), the Anthozoa are by most measures less complex than their medusozoan cousins. Given of the low number of paralogs, the

homeodomain repertoire of *Aurelia* appears more similar to the ancestral cnidarian than *Nematostella's*, even though *Aurelia* appears far more derived in other aspects, such as morphology (Marques and Collins 2005) and mitochondrial genome structure (Shao et al. 2006). Despite the fact that *Nematostella* is sometimes described as a "basal" cnidarian (Fritzenwanker and Technau 2002; Scholz and Technau 2003; Technau et al. 2005), this study is a powerful reminder that all living animals exhibit a mosaic of basal and derived traits, and that reconstructing the genomic evolutionary history of animal life will continue to require a broad, comparative approach.

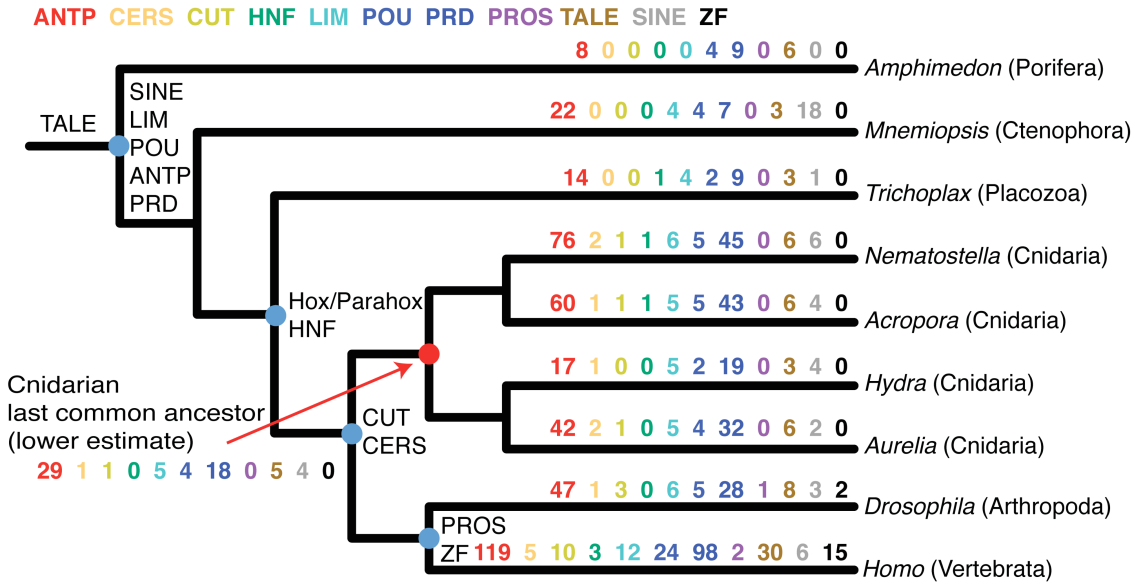

Figure 8. Homeodomain estimates for select animal taxa. The probable origination of important classes/subclasses is listed at relevant nodes. Gene counts for cnidarians are based on this study. Gene counts for non-cnidarians are taken from HomeoDB (Zhong and Holland 2011), Srivastava et al. (2008), Larroux et al. (2008) and Ryan et al. (2010).

The only gene radiation present in *Aurelia* to the exclusion of *Nematostella* and *Acropora* is the non-anterior Hox genes. Since the Hydrozoans *Clytia* and *Hydra* each have four

non-anterior Hox genes compared to *Acropora's* and *Nematostella's* two, the most parsimonious interpretation is that one round of gene duplications occurred before the split of anthozoans and medusozoans, followed by a second round of duplications in the medusozoa. Assuming CnidHox and Scox genes are monophyletic, a third round of duplications occurred in the Scyphozoa, leaving *Aurelia* with a total of seven non-anterior Hox paralogs. However, it is clear from our results that gene loss has been an important part of cnidarian evolution (and see Jacobs and Gates, 2003), so we cannot exclude the possibility that these higher counts in *Aurelia* are the result of a scyphozoan-specific retention of posterior Hox genes. Either way, the dynamic expression of these genes through the *Aurelia* life cycle, combined with their expansion through medusozoan evolution, makes them prime candidates for further study regarding the development of complex cnidarian bodyplans.

Finally, comparative RNA-Seq suggests many possible similarities in homeodomain expression between *Aurelia* and *Nematostella*, but these similarities are obfuscated by taxon-specific paralogs taking on opposing functions. Even with this uncertainty, some genes show conserved patterns of regulation (*Six3/6*), while others show opposing patterns (*Hhex*, *Irx*). These differences could reflect cryptic variation in cnidarian embryogenesis (for example, see the "secondary gastrulation" reported for *Aurelia* planula in Yuan et al. 2008). In *Nematostella*, *Six3/6* regulates early patterning of the planula's aboral domain, while *Irx* is expressed around tissue in the apical organ (Sinigaglia et al. 2013). Prior to the study of *Aurelia* planula with immunohistochemistry, apical organs had not been recognized in the medusozoa (Yuan et al. 2008). The lack of

Irx upregulation in *Aurelia* challenges the notion that the apical organ in *Aurelia* and *Nematostella* are homologous, and should be followed up with further work.

MATERIALS AND METHODS

All commands used in the bioinformatics of this paper have been reproduced in Supplementary File 1.

Animal Culture

Aurelia was sampled at seven points in its life cycle. Planula larva were collected from brooding females raised at the Cabrillo Aquarium in San Pedro, California. Motile, ovoid larvae were isolated, and approximately 300 individuals were sampled (this sample is subsequently referred to as "early planula"). The remaining planula were allowed to proceed with development, and 72 hours later, approximately 300 free-swimming, elongated planula ("late planula") were separated from approximately 300 animals that had settled on the surface tension of the water, and were metamorphosizing into primary polyps ("primary polyps"). *Aurelia* polyps were raised in 18°C seawater at UCLA, and 25 animals were collected for nucleic acid extraction. We induced strobilation in polyp colonies by incubating the animals in 1mL of iodine per gallon of seawater for five days, changing the water every day. Strobila were present one week following treatment, and 25 animals were collected for nucleic acid extraction. 72 hours after the collection of strobila, metamorphosis had completed in the remaining animals, and a significant number of ephyra were available. 30 ephyra were collected for nucleic acid extraction. Three weeks following the collection of ephyra, some of the remaining ephyra had

developed a complete bell, and were sampled as "juvenile" medusas. Ten individuals, each with a bell approximately 2cm in diameter, were collected for nucleic acid isolation.

RNA Isolation

Aurelia genomic DNA and RNA was isolated from the seven life stages using a phenol-chloroform protocol followed by a second cleanup with TRI Reagent (Sigma-Aldrich).

Library Preparation and Sequencing

The integrity of total RNA was verified using the Nanodrop 2000c (Thermo Scientific), Qbit 2.0 Fluorometer (Life Technologies), and Bioanalyzer 2100 (Agilent). The samples were converted into tagged cDNA libraries using the TruSeq RNA Sample Preparation Kit v2 (Illumina) according to the manufacturer's protocol. These libraries were sequenced using three lanes of a 100 base pair paired-end Illumina HiSeq 2000 run. The resulting data was cleaned and trimmed using the FASTX-Toolkit run on UCLA's Galaxy platform. Contigs were assembled into predicted genes and isoforms using the Trinity Package (Grabherr et al. 2011; Haas et al. 2013).

Data Collection and Mining

The Trinity-predicted transcripts were converted into best scoring peptides using the TransDecoder package included in Trinity. Genome assemblies and predicted peptide models for *Nematostella* and *Hydra* were downloaded from Metazome (http://www.metazome.net/); the relevant data for *Acropora* was downloaded from Compagen (Hemmrich and Bosch 2008). Genome assemblies from *Acropora*, *Aurelia*,

Hydra, and *Nematostella* were concatenated into a single file and formatted into a nucleotide BLAST database using the standalone BLAST package, and the protein models were converted into a protein database.

We queried our databases with candidate homeodomain sequences, representing each of the eleven major classes (ANTP, PRD, TALE, POU, CERS/LASS, PROS, ZF, LIM, HNF, CUT, and SINE), as well as four highly divergent *Mnemiopsis* homeodomains (HD07, HD141, HD31, and HD60) and sequences representing the major plant homeodomain families in *Arabidopsis* (ZIP I, ZIP II, ZIP III, ZIP IV, KNOX1, KNOX2, WOX, and DDX). All query sequences are available in Supplementary File 1 (Part 1). Protein sequences were queried using BLASTp with an e-value cutoff of 0.1. The results from all queries were combined, filtered for unique sequences, and the full-length peptides were retrieved from the database using Samtools**.** Genomic contigs were queried using tBLASTn with an e-value cutoff of 0.1; the matching DNA sequences were retrieved from the contigs using Samtools based on the start and stop coordinates recovered from the BLAST analysis. These genomic reads were translated in both forward and reverse directions using the Transeq tool (part of the EMBOSS package).

The presence of homeodmains in the BLAST results for peptides and translated genomic regions were assayed using the standalone PfamScan. Sequences were only retained if PfamScan identified a homeodomain. To remove genomic contigs that were already represented by longer peptide predictions, the results were demultiplexed by species, and subjected to substring dereplication, using USEARCH (http://www.drive5.com/usearch/).

Following substring dereplication, all reads were concatenated back together into a single dataset.

Tree Reconstruction

Using the coordinates from PfamScan, we extracted homeodomains from longer protein sequences using SAMTOOLS. We then combined our homeodomain dataset with annotated homeodomain data for *Homo sapiens*, *Branchiostoma floridae*, and *Drosophila melanogaster* downloaded from HomeoDB (Zhong et al. 2008; Zhong and Holland 2011). These sequences were aligned using the standalone version of MUSCLE. RaxML tree reconstruction was performed using the BlackBox web server (Stamatakis et al. 2008), using a gamma model of rate heterogeneity and an LG substitution matrix. PhyML tree reconstruction was performed on the PhyML 3.0 webserver (Guindon et al. 2010), using an estimated gamma shape with four substitution rate categories, and an LG substitution matrix. Node probabilities were calculated using 100 bootstraps (RaxML) or aLRT SH-Like estimation (PhyML).

RNA-Seq

Following Trinity *de-novo* assembly, we used the RSEM package (Li and Dewey 2011) to map the reads back to predicted transcripts to estimate gene counts. While vetting the data, we discovered that three pairs of homeodomains shared the same transcript ("comp") identification, which Trinity assigns to isoforms of the same predicted gene. After inspecting the Trinity output, we determined that the program had incorrectly grouped these sequences together as isoforms. The true counts for these genes were

determined by summing the counts only for the relevant isoforms, and the genes were given new IDs to reflect our edits: comp223848_c0_seq1 (*Msxlx*), comp223848_c0_seq2-5 (*Scox5*), comp226868_c0_seq1/2/3/4 (*PRD3a*), comp226868_c0_seq5 (*PRD3c*), comp226258_c0_seq1/3/10 (*Hox-Like*), and comp226258_c0_seq2/4/6/9 (*Scox4*). The modified counts were used to produce the final count matrix for edge R (available in Supplementary File 1, Part 7).

Differential gene expression was calculated using the EdgeR package (Robinson et al. 2010), and heat map clustering was performed using the Bioconductor wrapper provided by Trinity. In we treated pre-polyp (early planula, late planula, and primary polyp) and post-polyp (strobila, ephyra, juvenile) stages as representatives of broader "larva" and "medusa" stages, which allowed us to leverage the power of additional biological replicates.

To test for differentially expressed genes in *Nematostella*, we used the count data generated by Helm et al. (2013). Because this dataset used JGI identifiers as opposed to UniProt IDs, we BLASTed our *Nematostella* sequences against the database of JGI peptides (constructed from "Additional File 1" in Helm et al. 2013) All but six *Nematostella* genes had an exact match in the Helm database. Counts for homeodomains were extracted from Additional File 2 in Helm et al. (2013), and EdgeR-based differential expression was calculated as described above (See Supplementary File 1, Part 7).

WORKS CITED

Chapman JA, Kirkness EF, Simakov O, et al. 2010. The dynamic genome of Hydra. Nature 464:592–596.

Chiori R, Jager M, Denker E, Wincker P, Da Silva C, Le Guyader H, Manuel M, Queinnec E. 2009. Are Hox Genes Ancestrally Involved in Axial Patterning? Evidence from the Hydrozoan Clytia hemisphaerica (Cnidaria). PLoS ONE 4:e4231.

Chourrout D, Delsuc F, Chourrout P, et al. 2006. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature 442:684–687.

Collins A, Schuchert P, Marques A, Jankowski T, Medina M, Schierwater B. 2006. Medusozoan Phylogeny and Character Evolution Clarified by New Large and Small Subunit rDNA Data and an Assessment of the Utility of Phylogenetic Mixture Models. Syst. Biol. 55:97–115.

Collins AG. 2002. Phylogeny of Medusozoa and the evolution of cnidarian life cycles. J. Evolution. Biol. 15:418–432.

Daly M, Fautin DG, Cappola VA. 2003. Systematics of the Hexacorallia (Cnidaria: Anthozoa). Zoological Journal of the Linnean Society 139:419–437.

Dawson MN, Jacobs DK. 2001. Molecular evidence for cryptic species of Aurelia aurita (Cnidaria, Scyphozoa). Biol. Bull. 200:92–96.

Dawson MN. 2004. Some implications of molecular phylogenetics for understanding biodiversity in jellyfishes, with emphasis on Scyphozoa. Hydrobiologia 530-531:249–260.

Erwin D, LaFlamme M, Tweedt S, Sperling E, Pisani D, Peterson K. 2011. The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. Science 334:1091–1097.

Finnerty JR, Martindale MQ. 1999. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. Evol. Dev. 1:16–23.

Fritzenwanker JH, Technau U. 2002. Induction of gametogenesis in the basal cnidarian Nematostella vectensis(Anthozoa). Dev. Genes Evol. 212:99–103.

Goldberg RB, Crain WR, Ruderman JV, et al. 1975. DNA sequence organization in the genomes of five marine invertebrates. Chromosoma 51:225–251.

Grabherr MG, Haas BJ, Yassour M, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing

the Performance of PhyML 3.0. Syst. Biol. 59:307–321.

Haas BJ, Papanicolaou A, Yassour M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Helm RR, Siebert S, Tulin S, Smith J, Dunn C. 2013. Characterization of differential transcript abundance through time during Nematostella vectensis development. BMC Genomics 14:266.

Hemmrich G, Bosch TCG. 2008. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. Bioessays 30:1010–1018.

Holland PWH, Booth HAF, Bruford EA. 2007. Classification and nomenclature of all human homeobox genes. BMC Biol. 5:47.

Horridge A. 1956. The Nervous System of the Ephyra Larva of Aurellia Aurita. Quarterly Journal of Microscopical Science. 3:59-74.

Jacobs DK, Gates RD. 2003. Developmental genes and the reconstruction of metazoan evolution—implications of evolutionary loss, limits on inference of ancestry and type 2 errors. Integr. Comp. Biol. 43:11-18.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. BMC Evol. Biol. 13:5.

Kuhn K, Streit B, Schierwater B. 1999. Isolation of Hox genes from the scyphozoan Cassiopeia xamachana: implications for the early evolution of Hox genes. J. Exp. Zool. 285:63–75.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Marques AC, Collins AG. 2005. Cladistic analysis of Medusozoa and cnidarian evolution. Invertebr. Biol. 123:23–42.

McShea DW. 1996. Metazoan complexity and evolution: is there a trend? Evolution 50:477-492.

Nakanishi N, Hartenstein V, Jacobs DK. 2009. Development of the rhopalial nervous system in Aurelia sp.1 (Cnidaria, Scyphozoa). Dev. Genes Evol. 219:301–317.

Nakanishi N, Yuan D, Hartenstein V, Jacobs DK. 2010. Evolutionary origin of rhopalia: insights from cellular-level analyses of Otx and POU expression patterns in the developing rhopalial nervous system. Evol. Dev. 12:404–415.

Nosenko T, Schreiber F, Adamska M, et al. 2013. Deep metazoan phylogeny: when

different genes tell different stories. Mol. Phylogenet. Evol. 67:223–233.

Park E, Hwang D-S, Lee J-S, Song J-I, Seo T-K, Won Y-J. 2012. Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. Mol. Phylogenet. Evol. 62:329–345.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol. 9:e1000602.

Pick KS, Philippe H, Schreiber F, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27:1983–1987.

Putnam NH, Srivastava M, Hellsten U, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Ryan JF, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR. 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis. Genome Biol. 7:R64.

Ryan JF, Pang K, Comparative Sequencing Program N, Mullikin JC, Martindale MQ, Baxevanis AD. 2010. The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. EvoDevo 1:9.

Ryan JF, Pang K, Schnitzler CE, et al. 2013. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science 342:1242592–1242592.

Schneuwly S, Klemenz R, Gehring WJ. 1987. Redesigning the body plan of Drosophilaby ectopic expression of the homoeotic gene Antennapedia. Nature 325:816–818.

Scholz CB, Technau U. 2003. The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian Nematostella vectensis (Anthozoa). Dev. Genes Evol. 212:563–570.

Scott MP, Weiner AJ. 1984. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc. Natl. Acad. Sci. 81:4115–4119.

Shao Z, Graf S, Chaga OY, Lavrov DV. 2006. Mitochondrial genome of the moon jelly Aurelia aurita (Cnidaria, Scyphozoa): A linear DNA molecule encoding a putative DNA-dependent DNA polymerase. Gene 381:92–101.

Shinzato C, Shoguchi E, Kawashima T, et al. 2011. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature 476:320–323.

Sinigaglia C, Busengdal H, Leclère L, Technau U, Rentzsch F. 2013. The Bilaterian Head Patterning Gene six3/6 Controls Aboral Domain Development in a Cnidarian. PLoS Biol. 11:e1001488.

Srivastava M, Begovic E, Chapman J, et al. 2008. The Trichoplax genome and the nature of placozoans. Nature 454:955–960.

Srivastava M, Simakov O, Chapman J, et al. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466:720–726.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57:758–771.

Steele RE, David CN, Technau U. 2011. A genomic view of 500 million years of cnidarian evolution. Trends Genet. 27:7–13.

Technau U, Rudd S, Maxwell P, et al. 2005. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet. 21:633–639.

Yuan D, Nakanishi N, Jacobs DK, Hartenstein V. 2008. Embryonic development and metamorphosis of the scyphozoan Aurelia. Dev. Genes Evol. 218:525–539.

Zhong Y-F, Butts T, Holland PWH. 2008. HomeoDB: a database of homeobox gene diversity. Evol. Dev. 10:516–518.

Zhong Y-F, Holland PWH. 2011. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol. Dev. 13:567–568.

Chapter 3: The Developmental Transcriptome of the Moon Jellyfish *Aurelia*: Insights into the Evolution of Life History Complexity

ABSTRACT

We present the developmental transcriptome of the moon jellyfish *Aurelia sp.1*, a basal animal with a complex life cycle. We queried seven time points across *Aurelia's* life history, and discovered two major shifts in gene expression correlating with formation of the two "adult" morphs (the transition from primary polyp to polyp, and the production of medusa during strobilation). The morphologically complex medusa stage that distinguishes *Aurelia* from other model cnidarians is not enriched in novel (orphan) genes, but is enriched in many conserved cell-signaling pathways and transcription factors. Functionally similar G protein-coupled receptors and other neuroactive receptors are redeployed through the life cycle, correlating with major reorganizations of the *Aurelia* nervous system. Taken together, these observations suggest that the transcription regulatory equipment employed in bilaterian animal development is deployed successively in the two major stages of development. While many genes in the canonical mammalian stem cell pluripotency network are either absent in *Aurelia* or lack interesting expression patterns, many tumor suppression genes associated with *P53* are significantly up- or downregulated through the life cycle, suggesting that the need for tumor suppression changes in different life stages. This study offers a broad, first-order analysis of the *Aurelia* transciptome through the complete life cycle, providing a suite of candidate genes for future research, and a backbone for additional gene expression studies at more refined time scales.

INTRODUCTION

The first scientific description of *Aurelia's* early life stages were made by Michael Sars in 1829; the two morphs he described were so disparate that he classified them as separate genera (recounted in Agassiz 1860). The first organism, which Sars named *Scyphistoma filicorne*, was sedentary and reminiscent of a sea anemone or hydroid. The second—Sars' *Strobila octoradiata*—was also sessile, but was divided into a series of discs that would eventually detach and swim about freely, each reminiscent of small jellyfish. It wasn't until 1835 that Sars was to "rectify some major faults in the preceding observations" ("*berigtige nogle væsentlige Feil i de foregaaende Observationer*"; Sars 1835, pg.16), and recognize the two "genera" as developmental stages of the same animal. By this time, he came to suspect that the freely swimming animals were allied to a tiny, previously described jellyfish (Johann Eschscholtz' *Ephyra*; Eschscholtz 1829), but still failed to connect these forms with the much larger moon jellyfish (Sars' *Medusa aurita*) for at least another two years (Agassiz 1860).

The nearly decade-long process of piecing together the *Aurelia* life cycle is testament to the remarkable changes in form that occur over the animal's life. The life cycle (illustrated in Figure 1B) begins with a planula-type larva. As the larva matures, neurites extend from the middle of the animal anteriorly and posteriorly, creating a plexus of anterior neurites that have been hypothesized to function as an apical organ (Nakanishi et al. 2008). This organ presumably helps the larva determine where to settle, and begin its metamorphosis into a polyp. Antibody staining suggests that metamorphosis of the primary polyp involves a dramatic reorganization of the nervous system, as well as the

destruction and redevelopment of the endoderm (Nakanishi et al. 2008; Yuan et al. 2008). By the end of this process, the posterior end of the animal has developed into a mouth, surrounded by four tentacle buds and leading to a blind gut. The polyp (still often referred to as a scyphistoma after Sars) grows to a determinate size, and develops a ring of feeding tentacles that circle the mouth. When healthy, the polyp continuously produces clones through asexual budding, which appears to be driven by constant cellular proliferation (Gold and Jacobs 2013; Takashima, Gold, and Hartenstein 2013). This could explain why individual polyps show no signs of senescence after years in the lab, and how animals can be regenerated from isolated tentacles or disassociated cells (reviewed in Gold and Jacobs 2013). Proper environmental conditions can trigger the polyp to begin strobilation (after Sars' *Strobila*) wherein each polyp undergoes transverse fission, and produces a series of ephyra depending on the polyp's size (Kroiher et al. 2000). Strobilation is regulated by retinoic acid signaling, as well as several transcription factors unique to *Aurelia* (Fuchs et al. 2014). Following strobilation, the polyp returns to its original morphology, and can proceed with either form of asexual reproduction (budding or strobilation). By the time an ephyra has detached from the strobila, it has begun to develop eight sensory structures called rhopalia, which circle the animal's bell. Each rhopalia neurally integrates a mechanosensory touch plate, a geosensory statocyst, and a photoreceptive eye-cup, although the later does not fully develop until late in the ephyra's development (Nakanishi et al. 2009). Ephyrae eventually develop into mature medusas, which sexually reproduce and generate planula larvae that the female broods in grooves along the underside of the bell. Like the polyp, the medusa has impressive, albeit restricted, regenerative capabilities; it can regrow lost rhopalia, and can even "de-grow"

many tissues under times of stress (Hamner and Jenssen 1974). But unlike the polyp, the medusa undergoes a clear process of senescence; in the wild new medusas bloom and die off annually, and even in captivity the medusa exhibits morphological degradation within a few years (Hamner and Jenssen 1974; Moller 1980). The connection between longevity, sexual reproduction, and morphological complexity is unclear, but *Aurelia* offers a rare opportunity to study an animal species with two diametrically opposed "adult" forms.

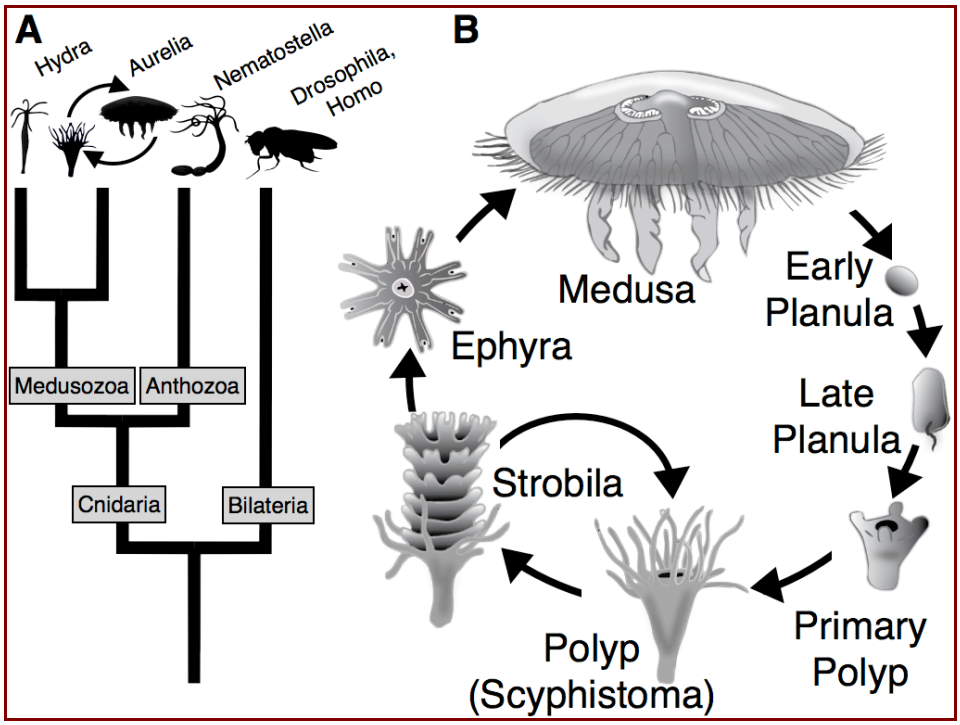

Figure 1. Life cycle and phylogenetic position of *Aurelia*. (A) The relationship between *Aurelia, Hydra, Nemastostella,* and the bilaterians. (B) The life cycle of *Aurelia*.

Aurelia is part of the phylum Cnidaria, which also includes sea anemones, corals, and hydras (Figure 1A). While there is still uncertainty regarding the relationships between the earliest-branching animal lineages, most phylogenetic studies support the cnidarians as the major sister clade to the bilaterians (protostomes plus deuterostomes), which

encompasses 99% of all living animal species (Pick et al. 2010; Erwin et al. 2011; Ryan et al. 2013). The cnidarians *Hydra* and *Nematostella* have become emerging model organisms for development and comparative genomics, but neither has a medusa life stage, and both therefore lack complex sensory structures or a multistage life cycle. The two organisms also have dramatically different life histories (*Nematostella* undergoes typical embryological development, while most laboratory strains of *Hydra* only produce by asexual budding), which can make it difficult to synthesize results between the two. As a species that shares aspects of both *Hydra's* and *Nematostella's* development, adding *Aurelia* to the list of cnidarian model organisms will aid in the interpretation of developmental data, and provide insight into one of the earliest examples in the evolution of life history complexity.

RESULTS AND DISCUSSION

We used next-generation sequencing to assemble and analyze the *Aurelia* transcriptome at seven time points (illustrated in Figure 1B). *De novo* assembly of ~320,000,000 100 base pair paired-end reads using Trinity (Haas et al. 2013) produced 191,123 transcripts, which clustered into 117,320 genes. After vetting the data (see Figure 2), we retained 52,986 transcripts that clustered into 24,308 genes. This later gene count is comparable to estimates for other cnidarians (Putnam et al. 2007; Chapman et al. 2010), which suggests our vetting process was effective at retaining true *Aurelia* transcripts.

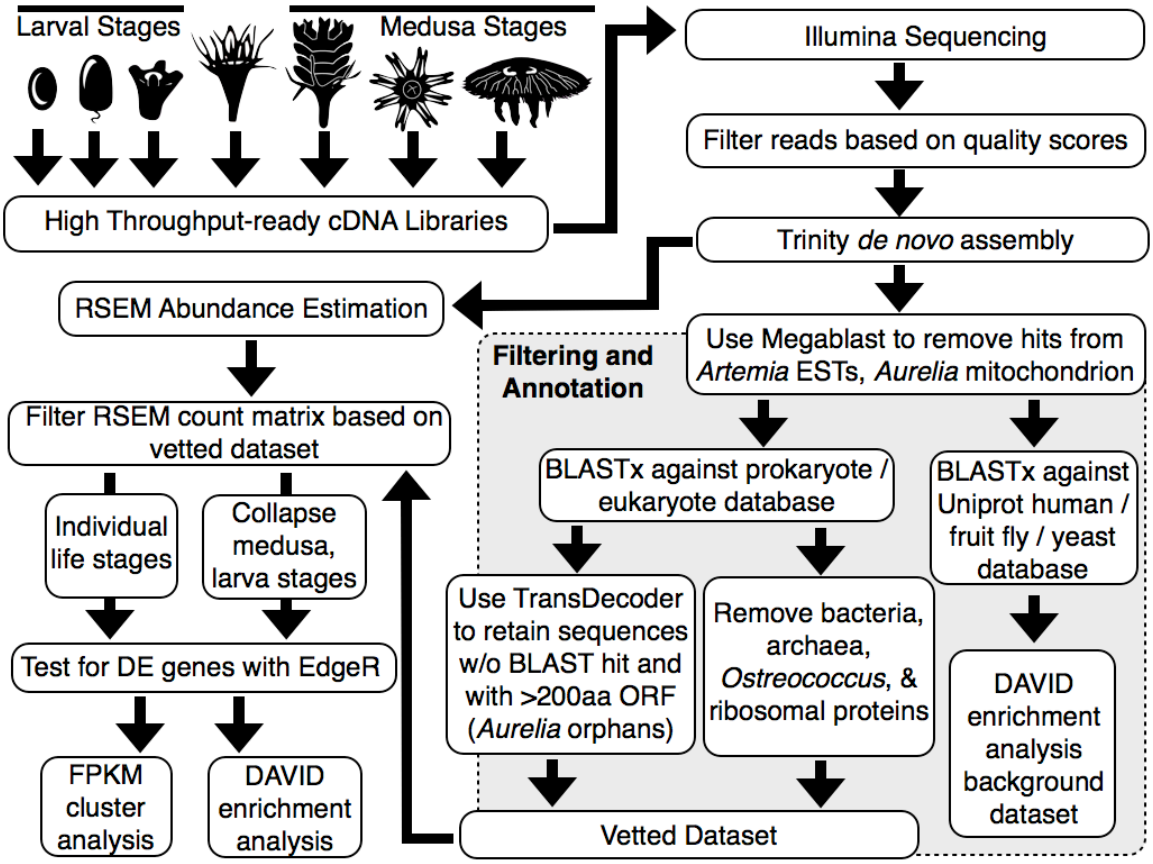

Figure 2: Graphical summary of analyses performed in this study.

As part of the filtering and annotation process, we used BLASTx to query our transcripts against a custom-built protein database consisting of ten animals, three non-metazoan opisthokonts, one green alga, and 34 prokaryotes (see Materials and Methods for details). Besides helping us filter out probable algal and prokaryotic contaminants (although we currently cannot rule out lateral gene transfer), this analysis allowed us to quantify the best hits for each *Aurelia* transcript relative to the 13 opisthokonts (Figure 3A).

The sequence similarity of the *Aurelia* transcriptome broadly reflects its phylogenetic relationships to these 13 opisthokonts. About half (47.7%) of all vetted *Aurelia*

transcripts had a best match with a *Nematostella* or *Hydra* protein, while another 4.7% of genes appear to be unique to *Aurelia*. Following cnidarians, the majority (31.7%) of top BLAST hits came from deuterostomes (*Strongylocentrotus*, *Branchiostoma*, and *Homo*). This can be probably be explained by the divergent rates of evolution in bilaterian lineages, where chordates show relatively low rates of sequence evolution compared to protostome model systems such as *Drosophila* and *Caenorhabditis*. Despite their collective assignment as "basal metazoans", few *Aurelia* transcripts share their closest identity to the sponge *Amphimedon* or the placozoan *Trichoplax*; in this analysis, there are approximately as many *Aurelia* genes most similar to humans as to *Amphimedon.* If the vetted *Aurelia* transcripts are queried against a database consisting of only human and *Amphimedon* proteins (Figure 3B), 61% of transcripts share a closer identity to humans, suggesting that at the peptide sequence level, the transcriptome of *Aurelia* is much more similar to our own than to the more basal sponges.

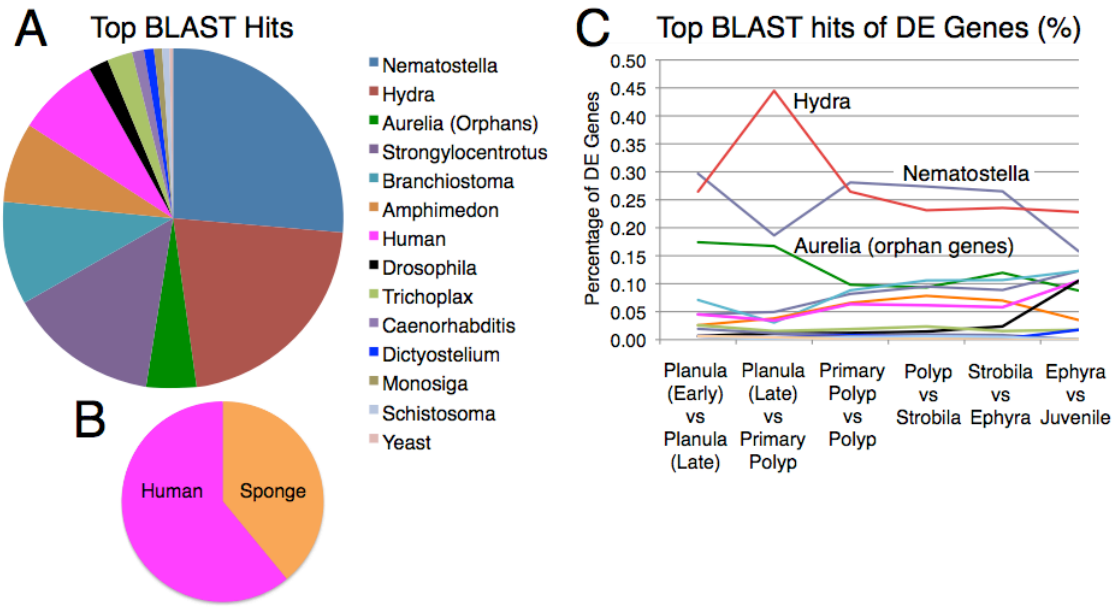

Figure 3. Comparisons of *Aurelia* transcripts to other opisthokonts. (A) The total fraction of best BLASTx hits for all vetted *Aurelia* transcripts against a database consisting of 13 opisthokonts. (B) An analysis of the same *Aurelia* dataset against a database restricted to sponge (*Amphimedon*) and human proteins. (C) The best BLAST hit assignments for all DE genes (upregulated and downregulated) at each life stage transition, converted into percentages. See figure 4 for information on the total number of differentially expressed genes for each pairwise comparison.

When best BLAST assignments for differentially expressed (DE) genes are plotted through the life cycle, the ratios of top organisms remain relatively stable over time (Figure 3C). At most life stages, the majority of DE genes share highest sequence similarity with *Nematostella* as opposed to *Hydra*, but there is a dramatic reversal during the formation of the polyp; this could make sense since the scyphistoma is morphologically much more similar to *Hydra's* polyp than *Nematostella's*. This reversal also occurs during the development of the ephyra into a juvenile medusa, but is based on a small number of DE genes (see Figure 4 for DE gene counts), which makes the significance of those results suspect. Interestingly, there appears to be no enrichment of *Aurelia* orphan genes in the medusa stages; these findings are similar to those found in the adult stage of *Amphimedon* (Conaco et al. 2012), and suggest that the evolution of clade-specific life stages in early-branching animals does not require the evolution of novel genes.

The largest changes in gene expression occur during polyp and medusa formation

Following transcript abundance estimation at each life stage using RSEM (Li and Dewey 2011), we used EdgeR (Robinson et al. 2010) to perform digital gene expression analyses

(Figure 2, and see Materials and Methods). The major changes in gene expression occur during the transition from the primary polyp into the polyp, and the polyp into the strobila (Figure 4). As the heat map in figure 4A illustrates, the seven life stages hierarchically cluster into three major clades based on the overall similarity of gene expression patterns: one clade encompasses the pre-polyp stages (early planula, late planula, and primary polyp); a second clade is restricted to the polyp, and a third clade encompasses the post-polyp stages (strobila, ephyra, and juvenile). While each sample in our time series was generated by pooling RNA from multiple animals (thereby accounting for variation between individuals, see Materials and Methods for more information) we only had one biological replicate (the polyp) available to estimate the variation within transcript abundance estimates. While the model used in EdgeR can provide robust estimates of biological variation with a single replicate (Robinson et al. 2010), the similarity within pre- and post-polyp stages (which in some cases is comparable to the variation between polyp biological replicates), encouraged us to re-run our analyses, treating the pre- and post-polyp samples as replicates of broader "larval" and "medusa" stages respectively. In these later analyses, we sacrificed temporal resolution for the increased statistical power that comes with greater biological replication. Subsequently, many of the downstream analyses in this study take advantage of both methods of partitioning the data.

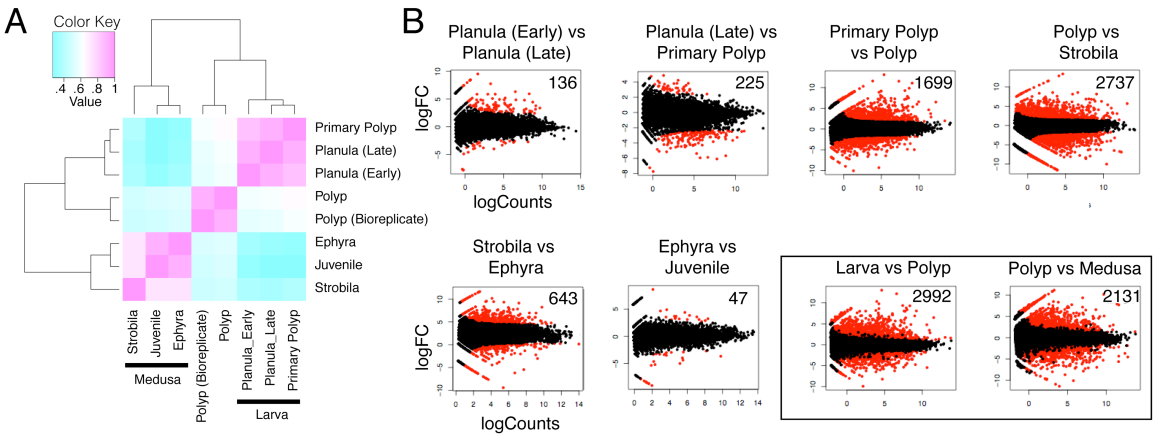

Figure 4. (A) Heat map illustrating the euclidean correlation matrix (with complete gene clustering) that results from comparing transcript expression values (TMM-normalized FPKM) between each pair of samples used in this study. (B) MA plots for pairwise comparisons through the *Aurelia* life cycle. The total number of DE genes (visualized with red dots) are provided in the upper-right corner of each box. Significance was calculated as a false discovery rate-adjusted p-value < 0.05.

Even with low levels of replication, each pairwise comparison of life stages produced DE genes with strong statistical support (FDR-adjusted p-value cutoffs of 0.05). Interestingly, the transition from ephyra to juvenile medusa produced the smallest number of DE genes in our study, despite the fact that significant changes in bell shape and the production of marginal tentacles occur during this time. Jacobs et al. (2010) hypothesized that the tentacles of *Aurelia* function as sensory organs, and could be developmental homologs of the rhopalia that form during the strobila and ephyra life stages. Subsequently, if genetic regulation of the medusa's tentacles is largely a recapitulation of rhopalia development, we might not expect to see major shifts in gene expression when these stages are compared at such broad time scales. Morphological changes to the bell might be best understood by research suggesting that changes in *Aurelia's* bell morphology are not driven by age, but by biomechanical pressures related

to changes in water viscosity as the ephyra increases in size (Nawroth et al. 2010); our results could therefore suggest that morphological changes to the bell shape have limited genetic regulation.

Gene Enrichment Analysis

Following digital gene expression analysis, we used the DAVID Bioinformatics Database (v6.7; Huang et al. 2008; Huang et al. 2009), to query DE genes against the vetted transcriptome, looking for the enrichment of protein domains and gene pathways as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database (Kanehisa 2000; Kanehisa et al. 2014). To generate the gene lists for the DAVID server, we used the Uniprot accession IDs recovered when our transcripts were queried against a database consisting of Human, *Drosophila*, and yeast proteomes using BLASTx (Figure 2). This method could lead to two areas of potential bias: firstly, the best BLAST hit for any *Aurelia* transcript might not reflect its true identity, which can only be fully resolved through detailed phylogenetic analyses of the relevant gene; secondly, only a subset of Uniprot accession IDs were recognized by the server (the relevant DAVID IDs are included in Supplementary File 1), and subsequently included in the enrichment analysis. Despite these limitations, this method provides a robust first-order look at gene and pathway enrichment through the *Aurelia* life cycle. Our results are summarized in Table 1. A number of protein domains reoccur at multiple life stages, notably epidermal growth factors (EGFs), 7-transmembrane G protein-coupled receptors (GPCRs), signaling pathway candidates (e.g. Notch, serrate, TGF-β), and transcription factor domains (e.g. homeodomains, HLH domains). Many of these gene families exhibit complex patterns of

up- and downregulation through the life cycle, and will be explored in further detail below.

	Enriched Domains in DE Genes (InterPro)		**Enriched Pathways Based on DE Genes (KEGG)**	
Planula (Early) versus Planula (Late)	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	none	none	none	none
Planula (Late) versus Primary Polyp	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	none	SCP-like extracellular	none	none
		Allergen V5/Tpx-1 related		
Primary Polyp versus Polyp	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	7TM GPCR * ✝	7TM GPCR * ✝	TGF-beta signaling pathway	Neuroactive ligand-receptor interaction ✝
	Allergen *	Fibronectin *	Neuroactive ligand-receptor interaction ✝	Chondroitin sulfate biosynthesis
	BMP1/tolloid-like	Immunoglobulin *	ECM-receptor interaction	
	Cadherin ✝	Peptidase * ✝		
	Coagulation factor 5/8 type, C-terminal ✝			
	Concanavalin A-like lectin/glucanase, subgroup ✝			
	CUB			
	Cysteine-rich flanking region, C-terminal			
	Delta/Serrate/lag-2 (DSL) protein			
	EGF * ✝			
	Epithelial sodium channel			
	Fibrinogen*			
	Fibronectin *			
	Homeobox *			
	Immunoglobulin *			
	Laminin G * ✝			
	Leucine-rich repeat *			
	Lipase *			

	Na+ channel, amiloride-sensitive			
	Notch ligand, N-terminal			
	Peptidase *			
	SCP-like extracellular			
	von Willebrand factor, type C			
Polyp versus Strobila	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	bHLH dimerisation region	Allergen V5/Tpx-1 related	Neuroactive ligand-receptor interaction †	none
	EGF *	Ankyrin	Vascular smooth muscle contraction	
	Extracellular ligand-binding receptor	Coagulation factor 5/8 type, C-terminal		
	Fibrinogen, alpha/beta/gamma chain, C-terminal globular	Concanavalin A-like lectin/glucanase, subgroup		
	Helix-loop-helix DNA-binding	CUB		
	Homeobox *	Cysteine-rich flanking region, C-terminal		
	Peptidase (M12A, S1/S6, S1A, metallopeptidases) †	Cytochrome P450 *		
		DEATH-like		
		EGF *		
		Endoglin/CD105 antigen subgroup		
		Fibrinogen, alpha/beta/gamma chain, C-terminal globular		
		Fibronectin, type III *		
		GCC2 and GCC3		
		Immunoglobulin *		
		Leucine-rich repeat *		
		Peptidase S8 and S53		
		PHR		
		Proprotein convertase, P		
		SCP-like extracellular		

		Transcription factor jumonji/aspartyl beta-hydroxylase		
		von Willebrand factor, type A and C		
Strobila versus Ephyra	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	Extracellular ligand-binding receptor	CUB	none	p53 signaling pathway
		EGF * †		
		Hyalin		
		Peptidase* (S1, S6, S1A) †		
		Proprotein convertase, P		
Ephyra versus Juvenile	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	none	Carotenoid oxygenase	none	none
Larva versus Polyp Only	BTB/POZ-like	none	none	none
	Complement control module			
	Nicotinic acetylcholine receptor			
	Potassium channel, voltage dependent, Kv, tetramerisation			
	Sushi/SCR/CCP			
Polyp versus Medusa Only	**Upregulated**	**Downregulated**	**Upregulated**	**Downregulated**
	7TM GPCR *	Coagulation factor 5/8 type	none	none
		Fibronectin, type III		
		Immunoglobulin I-set		

Table 1: Overview of enriched protein domains and gene pathways (with a Benjamini FDR-adjusted p-value cutoff of 0.05), based on DAVID enrichment analysis. The occurrence of multiple related domains is indicated with an asterisk (*). Genes / pathways that were also recovered after pre- and post-polyp stages were reanalyzed as biological replicates of larval and medusa stages are labeled with a cross (†). The complete output from DAVID is available in Supplementary File 1.

Cell Signaling Pathways

Aurelia has a full suite of candidate genes involved in the major signaling pathways (Figure 5A). Despite recovering several pathways in the enrichment analyses, individual

genes within each pathway show little evidence of co-expression. This is not particularly surprising, since these signaling pathways are interconnected to each other, and to more complex gene networks. There is one notable gene absent from our dataset; while phylogenetic analyses suggest that cnidarians have both *delta* and *serrate/jagged*-like ligands (Gazave et al. 2009), we only recovered a single homolog from our transcriptome. Interestingly, *Aurelia's delta/serrate* has characteristics of both paralogs (Figure 5C); like *delta*, the protein lacks a von Willebrand type C (VWC) domain, but it has 17 EGF domains, which is more similar to *serrate*, which on average has 14 EGF domains compared to *delta's* 7 (Gazave et al. 2009).

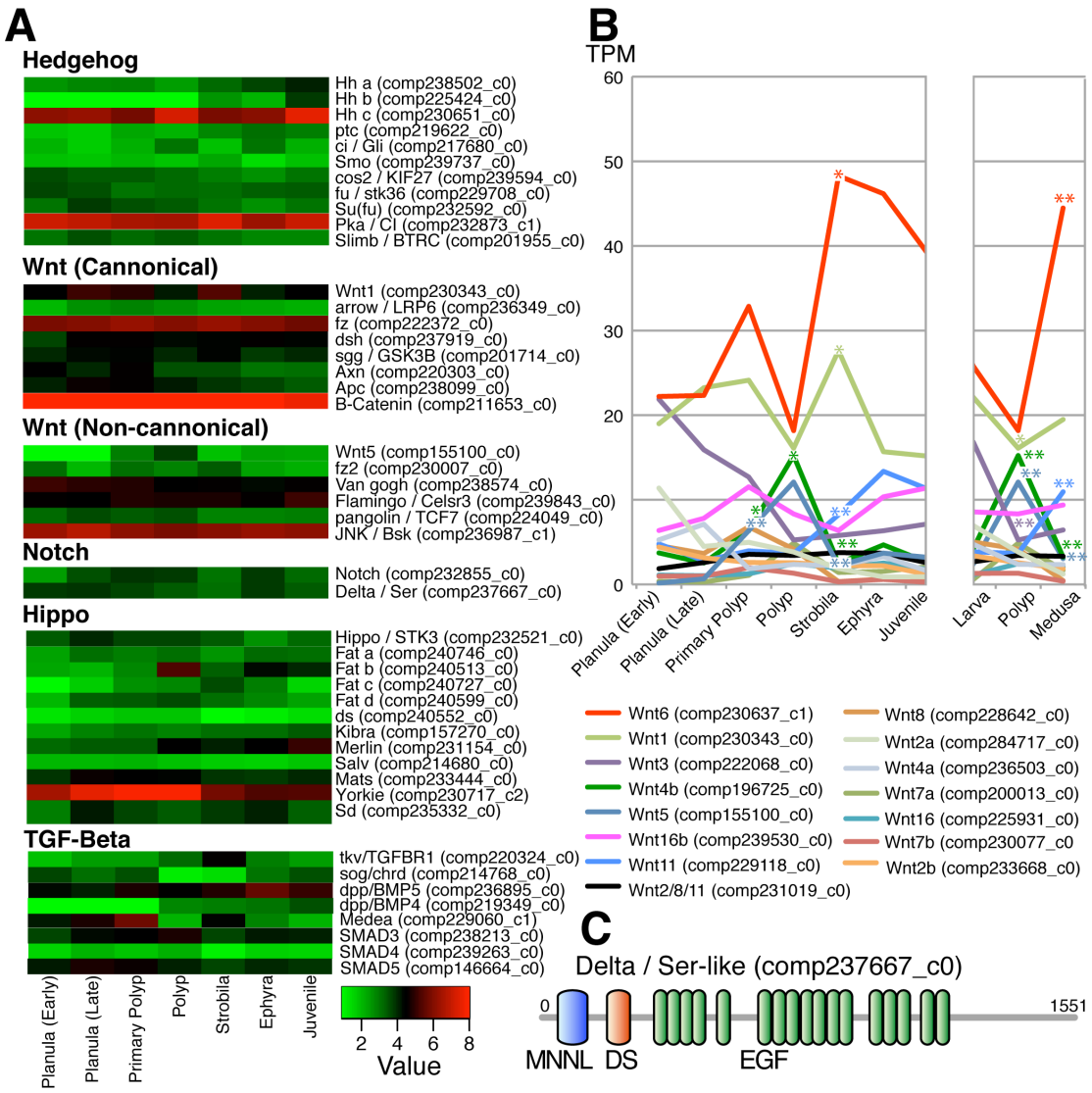

Figure 5. Gene signaling pathways. (A) FPKM-normalized heat map of candidate genes involved in hedgehog, notch, wnt, hippo, and TGF-β signaling pathways. (B) TPM-normalized gene counts for fifteen Wnt paralogs recovered from the transcriptome. Significant changes in gene count (p-value < 0.05) are labeled with an asterisk; highly significant changes (FDR-adjusted p-value < 0.05) are labeled with two asterisks. (C) Distribution of conserved domains in the *Aurelia Delta/Serrate* protein. Translation of domain abbreviations: MNNL = N terminus of Notch ligand, DS = Delta serrate ligand, EGF = epidermal growth factor-like.

We recovered fifteen wnt paralogs in *Aurelia*, one of the few times that an increase in cnidarian life history complexity correlated with an increase in gene paralogs (see also the non-anterior Hox genes in Chapter 2). There is broad evidence that wnt signaling plays a role in patterning and maintaining the body axis in both *Nematostella* and *Hydra* (reviewed in Lee et al. 2006). *Hydra* has a single *Wnt* gene (a parlog of *Wnt3*) that acts as an organizer for development and maintenance of the head, while *Nematostella* has twelve paralogs that exhibit overlapping but distinct expression patterns in the developing embryo, reminiscent of the *hox* code in bilaterian animals (Kussarow et al. 2005). Several *Wnt* paralogs exhibit dynamic shifts during the life cycle (Figure 5B); *Wnt4b* and *Wnt5* are upregulated in the polyp, while *Wnt1* and *Wnt3* are downregulated. In the post-polyp stages, *Wnt6* and *Wnt11* increase in expression, while *Wnt4b* and *Wnt5* are dramatically downregulated, suggesting that these later two genes play a specific role in maintaining the polyp bodyplan. Given the connection between wnt paralogs and cnidarian axial patterning, further study of these genes is warranted.

DNA-Binding Transcription Factor Domains

Running PfamScan against our predicted peptide dataset, we recovered all candidate proteins with conserved DNA-binding domains (Figure 6; see Materials and Methods for more details). We excluded the homeodomains from this analysis, as they are studied in detail in Chapter 2.

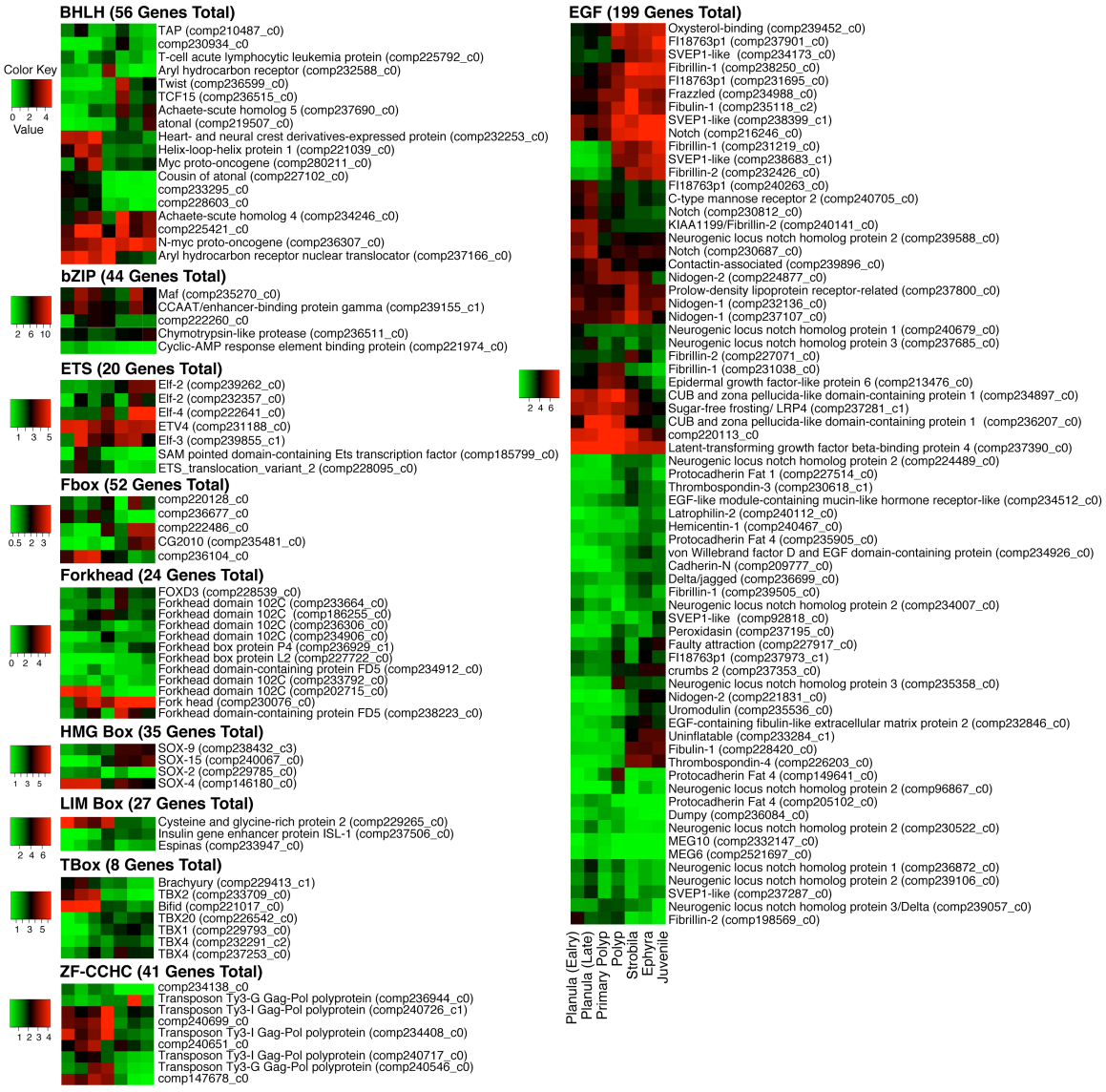

Figure 6. FPKM-normalized heat maps of DE genes based on the presence of conserved transcription factor binding domains (DE defined here as an FDR-adjusted p-value < 0.001, and a minimum of 4-fold change in expression for one or more comparisons). EdgeR and heat map generation was performed on each clade of genes independently. Genes were labeled using the best BLASTx hit against the human/*Drosophila*/yeast proteome database.

We recovered a large number of DE genes, some which have been extensively studied in other cnidarians, and some that have only been looked at in classical laboratory model organisms. Consistent with our DAVID enrichment analysis (Table 1), the majority of

DE transcription factors recovered were those containing EGF domains. Almost half of all EGF-containing genes were identified as *notch*, *nidogen*, *fibrillin*, or *svep* (*sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1*) -like. However, we do not necessarily interpret these results as implying a radiation of these gene families in *Aurelia*. For example, out of all the genes identified as *Notch*, only one (comp2328550_c0) had the full set of conserved domains that a canonical *Notch* protein possesses. It is likely that these other *Notch*-like genes were identified based on the number and length of EGF domain-repeats (similar issues with EGF-repeats were noted in an analysis of *Notch* in *Nematostella*, see Marlow et al. 2012). This issue highlights the limitations of using BLAST searches to annotate genes; the exact identity of these genes will ultimately need to be resolved with robust analyses of domain structure and gene phylogeny. Similar issues occurred with the Forkhead domains (which are dominated by *Forkhead domain 102c*-like genes) and ZF-CCH domains (dominated by *Transposon Ty3-G Gag-Pol polyprotein*-coding trasncripts). However, our BLAST searches provide unique annotations for most genes. There are too many genes to discuss in full detail, but we will highlight some of the transcripts that have the most dynamic expression patterns, and those we consider the most interesting for future work.

The BHLH clade is notable for a cluster of genes that exhibit strobila-specific upregulation: including putative *achaete-scute*, *twist*, and *TCF15* homologs. In many cnidarians, *achaete-scute*-like genes drive the differentiation of neural cell types, including the cnidocyte stinging cells unique to this group (Grens et al. 1995; Hayakawa et al. 2004; Seipel et al. 2004; Layden et al. 2012). Our analysis recovered a second

achaete-scute paralog that is highly expressed at multiple life stages, which could suggest that the strobila-specific gene is an important player during the reorganization of the polyp nerve net during metamorphosis. While *twist* plays a role in regulating endodermal development during *Nematostella* embryogenesis (Martindale 2004), the upregulation of *Aurelia twist* during strobilation is consistent with observations from the hydrozoan jelly *Podocoryne*, where *twist* shows strong but transient expression in the proliferating undifferentiated cell mass during medusa formation (Spring et al. 2000). In contrast to *achaete-scute* or *twist*, *TCF15* expression has not been studied in cnidarians. However, the gene is know to drive differentiation of embryonic stem cells in mammals and neoblast stem cells in planarians (Wagner et al. 2012; Davies et al. 2013), which might imply a connection with *twist* in driving cell pluripotency and transdifferentiation during this period of metamorphosis. Another neurogenic gene, *atonal*, is one of the few genes with juvenile-specific upregulation. This is particularly interesting, since *atonal/Math5* drives eye formation in diverse bilaterians, and is upregulated around the time that the optic cups of the *Aurelia* rhopalia are forming (Nakanishi et al. 2009).

Analysis of the ETS-domain containing transcription factors suggests an enrichment of dynamically expressed ELF-like genes. The upregulation of multiple putative ELF homologs during the ephyra and juvenile stages provides one of the rare examples in this study of genes restrictively enriched in these stages. The role of these ELF genes will require additional research, but at least one of these putative ELF paralogs (*ELF-4/MEF*) functions as a tumor suppressor by regulating cell quiescence (Yamada et al. 2009). The

hypothesis that the morphologically complex and short-lived medusa might require additional protection against tumorigenesis will be discussed later in this paper.

DE HMG-Box genes are enriched in putative Sox homologs, which have been studied in a variety of basal metazoans. The Sox genes show discrete patterns of expression: *Sox4* is upregulated in the larval stages, *Sox15* in the polyp and medusa stages, and *Sox9* in the medusa stages. *Sox4* (Sox Group C) plays a role in embryogenesis in bilaterians; it is also expressed in the oral pole of the anthozoans *Acropora* and *Nematostella* (Shinzato et al. 2008), and in the apical organ of the ctenophore *Mnemiopsis leidyi* (Schnitzler et al. 2014). *Sox9* (Sox Group E) related genes have been implicated in gonad development (Phochanukul and Russell 2010). The upregulation of *Sox9* in post-polyp stages could suggest a conserved function in *Aurelia*, but the expression of *Sox9* occurs well before the physical presence of gonads, which are first detectable in larger, mature medusas. The expression of this gene could imply that cryptic cell populations are set aside for gonad development as early as strobilation.

The T-Box clade is notable for a cluster of phylogenetically related genes (candidate *brachyury*, *TBX2*, and *Bifid/optomotor-blind* homologs), which show high expression in the larval stages. In *Nematostella*, *brachyury* is expressed in a circle around the blastopore, and persists in mature polyps in the elaborated endodermal tissue called mesenteries (Scholz and Technau 2003). However, mesenteries are unique to the sea anemones, which could explain the rapid downregulation of *brachyury* in *Aurelia* polyps.

Nervous and Sensory System Development

DAVID enrichment analysis suggests that the *Aurelia* transcriptome exhibits multiple shifts in G protein-coupled receptor (GPCR) domains, as well as members of the broader neuroactive ligand-receptor interaction pathway, over time. GPCRs are embedded in the membranes of cells, and activate signal transduction using a variety of ligands that can include light or odor-sensitive compounds, as well as hormones and neurotransmitters. Given the importance of the GPCRs to animal sensory systems, and the complexity of sensory structures in the medusa's rhopalia, we looked at these candidate genes in detail.

DE GPCRs are illustrated in Figure 7. Because the GPCRs represent a massive family with high levels of sequence similarity (in excess of the EGF domain-containing genes), we were less confident in the BLASTx annotations, and opted for broader categorizations as defined in GPCR SARfari, a subset of the ChEMBL database (Gaulton et al. 2012; see Materials and Methods for more information, and Supplementary File 1 for full annotations). As Figure 7 illustrates, DE GPCRs are enriched in receptors for short peptides and small molecules. A more detailed annotation for a subset of GCPRs is illustrated in part B of Figure 7. As Figure 7B suggests, the peptide receptors primarily recognize short peptides (e.g. neuropeptide, melanocortin, opioid), while the small molecules receptors primarily recognize monoamine derivatives (e.g. dopamine, seratonin, trace amines). Chemosensory peptides (chemokines) and light-sensitive opsins are rare in this analysis, and when they are present, they tend to be up- and downregulated several times throughout the life cycle. While we anticipated the upregulation of light-sensitive compounds during the formation of the medusa's eye-cup,

this does not appear to be an obvious pattern in the data; in fact, the opsin with the strongest DE in our analysis exhibits decreasing expression starting at the strobila life stage. Cnidarians without eyes—such as *Hydra*—are known to use opsins to regulate photosensitive behaviors, such as cnidocyte discharge (Plachetzki et al. 2012). Our results suggest that the complex sensory organs of *Aurelia* medusa might develop through the co-optation of GPCRs that are also used in less complex sensory systems in the earlier stages.

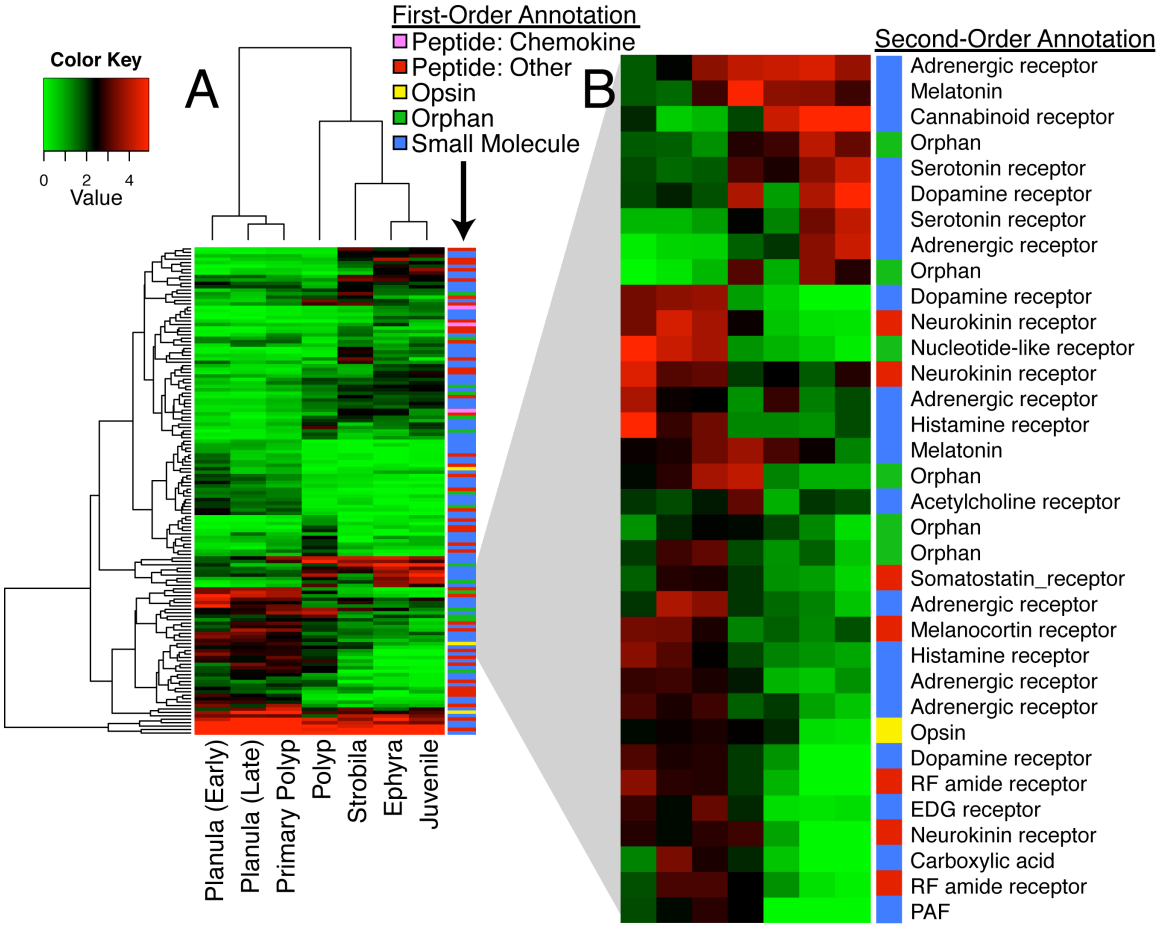

Figure 7. Heat map illustrating GPCRs in *Aurelia* that exhibit differential expression (DE defined as FDR-adjusted p-value < 0.001, and a minimum of 4-fold change in expression for one or more comparisons). (A) All DE opsins, with broad (first-order) annotation as determined by metadata from the GPCR SARfari. (B)

An expanded look at a subset of the more dynamic DE genes, with more detailed (second-order) annotation as determined by GPCR SARfari.

The re-use of GPCR classes through the *Aurelia* life cycle can also be visaulized by illustrating the results from DAVID enrichment of the receptor-ligand signaling pathway (Figure 8). In Figure 7B, many receptors in the same class show opposing patterns through development (e.g. melatonin and serotonin receptors); this is reflected in Figure 8 with arrows pointing in both directions. These results provide strong evidence for major reorganizations in neural coordination through the life cycle, which has been suggested by previous developmental studies (Nakanishi et al. 2008; Nakanishi et al. 2009).

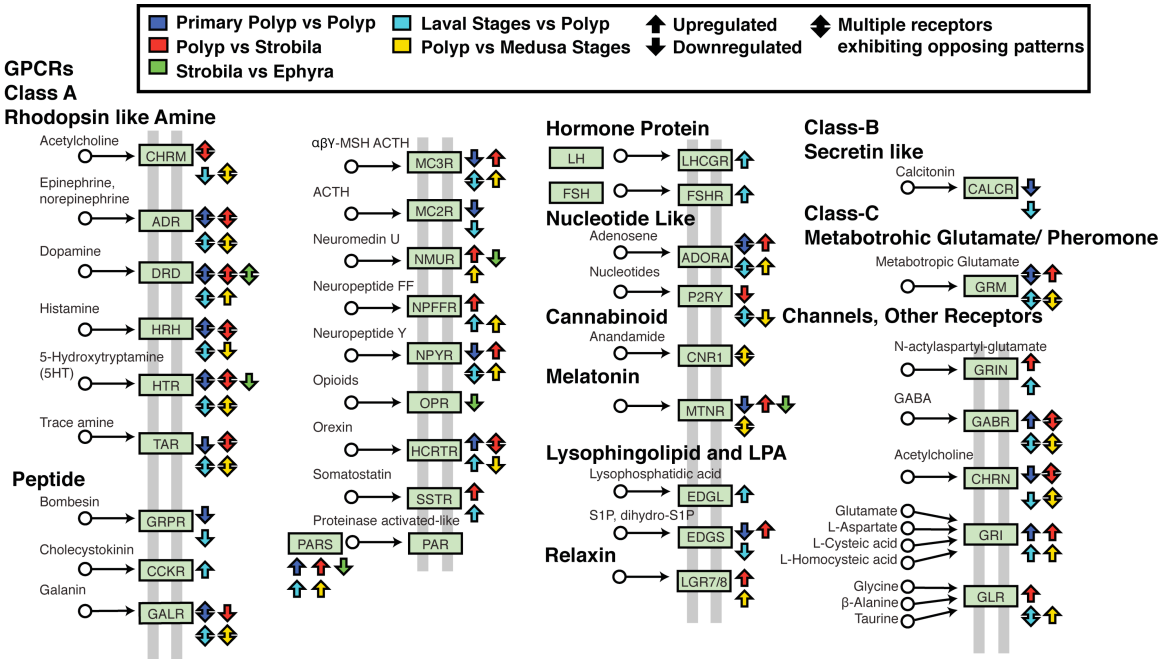

Figure 8. Neuroactive ligand-receptor interactions recovered in *Aurelia* using DAVID functional annotation. Colored boxes next to each receptor illustrate which comparison(s) produced significant results. Image modeled after KEGG pathway hsa04080.

Tumor Suppression and Stem Cell Dynamics

In many of our analyses, mammalian tumor suppressors and/or oncogenes were identified as candidate DE genes. There are no published cases of cancer in a cnidarian, and it has been proposed that *Hydra* is immune to tumorigenesis because of high cellular turnover, driven by the constant proliferation of several stem cells lines (Bosch 2008; Bosch et al. 2010). While *Aurelia* polyps also exhibit continual cell turnover, they do not appear to have interstitial stem cells, which we have previously argued are a derived feature of hydrozoans (Gold and Jacobs 2013). How *Aurelia* polyps maintain cellular proliferation without I-cells, and whether the short-lived medusa stage is at risk of cancer are currently unknown.

We tested the conservation of a canonical tumor-suppression network and a stem-cell pluripotency network in Figure 9. We used the STRING protein interaction database (v9.1; Franceschini et al. 2013) to recover the ten proteins most closely associated with *POU5f1*/*Oct-4* (a canonical mammalian stem cell gene) and *P53* (a canonical tumor suppressor and cell-cycle regulator) in humans. The results are widely divergent. Most of the genes in the *POU5f1* network do not have a candidate homolog in *Aurelia*. When the closest paralog is used, most genes show no significant shifts in expression, and those that do show no clear pattern of association. This supports our previous assertions that the *POU5f1/Oct-4* pluripotency network in mammalian stem cells is not conserved in cnidarians (Gold and Jacobs 2013). Conversely, all of the genes in the *P53* pathway have at least one homolog in *Aurelia*. DE genes appear to interact in expected ways; for example, *MDM2*—a suppressor of *P53*—is strongly downregulated as *P53* expression

strongly increases. Other DE genes show interesting changes through the life history; the serine/threonine protein kinase *ATM* is upregulated during medusa formation, while one *BRCA* homolog is significantly upregulated in the polyp. In mammalian cells, *ATM* protects cells against UV radiation (Shiloh 2001), which is likely important for *Aurelia* medusa that often congregate near the ocean surface. What *BRCA* genes—infamous for their role in human breast cancer—do in *Aurelia* is unclear, but the significant changes in expression through the life cycle make these interesting candidates for future work. Overall, these results suggest that the pressures of tumorigenesis might shift during *Aurelia's* life cycle according to changes to its ecology and morphology.

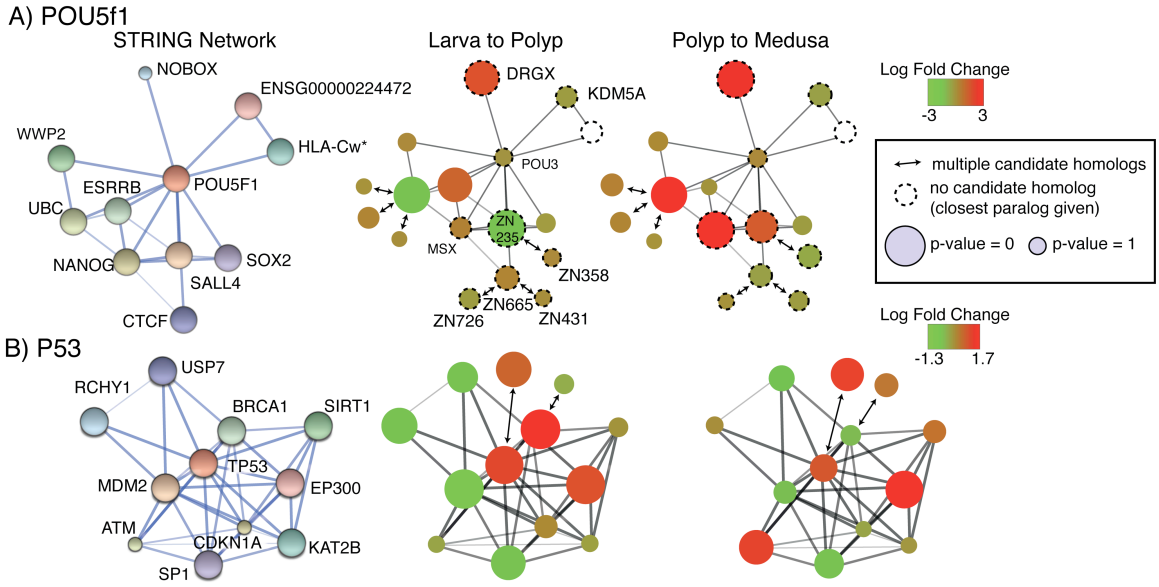

Figure 9. Gene expression of candidate proteins associated with *POU5f1* and *P53*. From left to right, the protein association network recovered by STRING, changes in gene expression from the larva to polyp transition, and changes in gene expression from the polyp to medusa. The size of each circle represents the p-value associated with change in expression; larger circles signify smaller p-values. Circles with a dashed line indicate that the *Aurelia* gene is probably not the closest homolog to the human query; in these cases, the more likely homolog is listed in or next to the circle.

CONCLUSIONS

This project represents our initial analysis of the moon jellyfish transcriptome through the animal's life cycle. Ultimately, these results will be refined as better gene models are built from the emerging genome (unpublished data), and many of the patterns and hypotheses presented in this study will need to be followed up with developmental, gene expression, and gene tree studies. Still, this analysis offers a suite of candidate genes and pathways for future research into the evolution of life history complexity in the animals.

MATERIALS AND METHODS

Animal culuture and collection, RNA isolation, and library perperation/sequencing was performed as described in Chapter 2. All commands and select scripts used in these analyses are available in Supplementary File 1.

Transcriptome Assembly

All bioinformatic analyses were performed on UCLA's Hofmann2 server, or using the Data Intensive Academic Grid (DIAG; http://diagcomputing.org/). Reads with a quality score less than 20 were removed using the Filter FastQ tool in Galaxy (Blankenberg et al. 2010). Filtered reads were assembled into predicted genes and isoforms using the Trinity Package (Grabherr et al. 2011; Haas et al. 2013).

Generation of Vetted Dataset

Using BLAST to compare all of our transcripts to the NCBI database would have been prohibitively time consuming, and runs the risk of matching our sequences to poor and

mis-annotated sequences. Subsequently, we constructed our own BLAST databases for vetting and annotation.

Sequences for the mitochondrial genome and ribosome (encompassing the 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, internal transcribed spacer 2, 28S ribosomal RNA gene, and intergenic spacer 1) were extracted from the *Aurelia* draft genome. We also downloaded all ESTs and nucleotide sequences for the brine shrimp (*Artemia*) available on NCBI, as we use this organism to feed our *Aurelia*. These sequences were formatted into BLAST databases, and the Trinity transcriptome was queried against these databases using MEGABLAST. All query sequences that had a hit in MEGABLAST were removed from the dataset.

Following removal of rRNA and *Artemia* contaminants, we constructed a BLAST database of prokaryote and eukaryote proteins for annotation and further vetting. Proteome datasets were downloaded from NCBI's FTP server. Prokaryotes were chosen in an attempt to capture the major clades of bacteria and archea. Bacteria incorporated into the database include *Aquifex aeolicus* (VF5), *Chlamydia muridarum* (str. Nigg), *Fusobacterium nucleatum* (subsp. nucleatum ATCC 25586), *Helicobacter hepaticus* (ATCC 51449), *Rhodopirellula baltica* (SH 1), *Leifsonia xyli* (subsp. xyli str. CTCB07), *Bacteroides fragilis* (YCH46), *Synechococcus elongatus* (PCC 6301), *Burkholderia ambifaria* (AMMD), *Coxiella burnetii* (RSA 331), *Leptospira biflexa* (serovar Patoc strain), *Agrobacterium vitis* (S4), *Deinococcus deserti* (VCD115), *Bacillus anthracis* (str. A0248), *Kosmotoga olearia* (TBF 19.5.1), *Dehalococcoides sp.* (GT),

Thermodesulfobacterium geofontis (OPF15), *Thermovirga lienii* (DSM 17291), *Corynebacterium argentoratense* (DSM 44202), and *Mycoplasma hyorhinis* (DBS 1050). Archaea incorporated into the database include *Methanocaldococcus jannaschii* (DSM 2661), *Pyrococcus horikoshii* (OT3), *Methanopyrus kandleri* (AV19), *Nanoarchaeum equitans* (Kin4-M), *Pyrobaculum arsenaticum* (DSM 13514), *Methanocella paludicola* (SANAE), *Archaeoglobus profundus* (DSM 5631), *Sulfolobus islandicus* (L.D.8.5), *Halalkalicoccus jeotgali* (B3), *Acidilobus saccharovorans* (345-15), *Cenarchaeum symbiosum* (A), *Methanobacterium sp*. (SWAN-1), *Pyrolobus fumarii* (1A), and *Ferroplasma acidarmanus* (fer1).

Eukaryotes were selected to capture major animal clades, and to check for algal contamination. The Eukaryote proteomes were downloaded from Uniprot. Eukaroytes used in this study include *Amphimedon queenslandica*, *Branchiostoma floridae*, *Caenorhabditis briggsae*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Hydra vulgaris* (formally *Hydra magnipapillata*), *Homo sapiens*, *Monosiga brevicollis*, *Nematostella vectensis*, *Ostreococcus tauri*, *Saccharomyces cerevisiae*, *Schistosoma mansoni*, *Strongylocentrotus purpuratus*, and *Trichoplax adherans*.

All prokaryote and eukaryote proteomes were concatenated into a single fasta file, and then formatted into a BLAST database. The Trinity-assembled *Aurelia* transcripts were BLASTed against this database. Sequences with a best BLAST hit that matched a prokaryote or the alga *Ostreococcus* was removed from the dataset. The transcripts that did not receive a BLAST hit were extracted and translated using the Transdecoder

wrapper in Trinity. Any transcript without a BLAST hit that produced a protein with a minimum length of 200 amino acids was retained; these were annotated as "*Aurelia* Orphan" sequences. This produced our final, vetted dataset of *Aurelia* transcripts.

Digital Gene Expression and Heat Map Generation (Cluster Analysis)

Following assembly, the transcripts were demultiplexed based on life history stage. To estimate counts of gene expression at each life stage, we used the RSEM package (Li and Dewey 2011). Since RSEM appears to produce the most accurate gene-level abundance estimates when large numbers of short single-end reads are used (Li and Dewey 2011), only the forward reads were used for this analysis. Differential gene expression was calculated using the EdgeR package (Robinson et al. 2010), and heat map clustering was performed using the Bioconductor wrapper provided by Trinity. All heat maps were produced using a euclidean gene distribution and complete gene clustering.

WORKS CITED

Agassiz L. 1860. Contributions to the Natural History of the United States of America: pt. 1. Acalephs in general. pt. 2. Ctenophoræ. 1860.

Blankenberg D, Gordon A, Kuster Von G, Coraor N, Taylor J, Nekrutenko A. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics, 26:1783-1785.

Bosch TCG, Anton-Erxleben F, Hemmrich G, Khalturin K. 2010. The Hydra polyp: nothing but an active stem cell community. Dev. Growth Differ. 52:15–25.

Bosch TCG. 2008. Stem Cells in Immortal *Hydra*. (Bosch TCG, editor.). Dordrecht: Springer Netherlands. pp. 37-57.

Chapman JA, Kirkness EF, Simakov O, et al. 2010. The dynamic genome of Hydra. Nature 464:592–596.

Conaco C, Neveu P, Zhou H, Arcila M, Degnan SM, Degnan BM, Kosik KS. 2012. Transcriptome profiling of the demosponge Amphimedon queenslandica reveals

genome-wide events that accompany major life cycle transitions. BMC Genomics 13:209.

Davies OR, Lin C-Y, Radzisheuskaya A, Zhou X, Taube J, Blin G, Waterhouse A, Smith AJH, Lowell S. 2013. Tcf15 primes pluripotent cells for differentiation. Cell Rep. 3:472–484.

Erwin D, LaFlamme M, Tweedt S, Sperling E, Pisani D, Peterson K. 2011. The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. Science 334:1091–1097.

Eschscholtz JF. 1829. System der Acalephen : eine aus führliche Beschreibung aller medusenartigen Strahlthiere. Berlin: F. Dummler

Franceschini A, Szklarczyk D, Frankild S, et al. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41:D808-D815.

Gaulton A, Bellis LJ, Bento AP, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40:D1100–D1107.

Gazave E, Lapébie P, Richards GS, Brunet F, Ereskovsky AV, Degnan BM, Borchiellini C, Vervoort M, Renard E. 2009. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. BMC Evol. Biol. 9:249.

Gold DA, Jacobs DK. 2013. Stem cell dynamics in Cnidaria: are there unifying principles? Dev. Genes Evol. 223:53–66.

Grabherr MG, Haas BJ, Yassour M, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Grens A, Mason E, Marsh JL, Bode HR. 1995. Evolutionary conservation of a cell fate specification gene: the Hydra achaete-scute homolog has proneural activity in Drosophila. Development 121:4027–4035.

Haas BJ, Papanicolaou A, Yassour M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Hamner WM, Jenssen RM. 1974. Growth, Degrowth, and Irreversible Cell Differentiation in Aurelia aurita. Int. Comp. Bio. 14:833–849.

Hayakawa E, Fujisawa C, Fujisawa T. 2004. Involvement of Hydra achaete-scute gene CnASH in the differentiation pathway of sensory neurons in the tentacles. Dev. Genes Evol. 214:486–492.

Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.

Jacobs DK, Gold DA, Nakanishi N, Yuan D, Camara A, Nichols SA, Hartenstein V. 2010. Basal Metazoan Sensory Evolution. pp. 175-193 in Key Transitions in Animal Evolution, B. Schieirwater and R. DeSalle eds. CRC Press.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42:D199–D205.

Kanehisa M. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28:27–30.

Kroiher M, Siefker B, Berking S. 2000. Induction of segmentation in polyps of Aurelia aurita (Scyphozoa, Cnidaria) into medusae and formation of mirror-image medusa anlagen. Int J Dev Biol 44:485–490.

Kussarow A, Pang K, Sturm C, Hrouda M. 2005. Unexpected complexity of the Wnt gene family in sea anemone. Nature 433:156-160.

Layden MJ, Boekhout M, Martindale MQ. 2012. Nematostella vectensis achaete-scute homolog NvashA regulates embryonic ectodermal neurogenesis and represents an ancient component of the metazoan neural specification pathway. Development 139:1013–1022.

Lee P, Pang K, Matus D, Martindale M. 2006. A WNT of things to come: Evolution of Wnt signaling and polarity in cnidarians. Semin. Cell Dev. Biol. 17:157–167.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Marlow H, Roettinger E, Boekhout M, Martindale MQ. 2012. Functional roles of Notch signaling in the cnidarian Nematostella vectensis. Dev. Biol. 362:295–308.

Martindale MQ. 2004. Investigating the origins of triploblasty: `mesodermal' gene expression in a diploblastic animal, the sea anemone Nematostella vectensis (phylum, Cnidaria; class, Anthozoa). Development 131:2463–2474.

Moller H. 1980. Population dynamics of Aurelia aurita medusae in Kiel Bight, Germany (FRG). Mar Biol 60:123–128.

Nakanishi N, Hartenstein V, Jacobs DK. 2009. Development of the rhopalial nervous system in Aurelia sp.1 (Cnidaria, Scyphozoa). Dev. Genes Evol. 219:301–317.

Nakanishi N, Yuan D, Jacobs DK, Hartenstein V. 2008. Early development, pattern, and reorganization of the planula nervous system in Aurelia (Cnidaria, Scyphozoa). Dev.

Genes Evol. 218:511–524.

Nawroth JC, Feitl KE, Colin SP, Costello JH, Dabiri JO. 2010. Phenotypic plasticity in juvenile jellyfish medusae facilitates effective animal-fluid interaction. Biol. Lett. 6:389–393.

Phochanukul N, Russell S. 2010. No backbone but lots of Sox: Invertebrate Sox genes. Int. J. Biochem. Cell Biol. 42:453-464.

Pick KS, Philippe H, Schreiber F, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27:1983–1987.

Plachetzki DC, Fong CR, Oakley TH. 2012. Cnidocyte discharge is regulated by light and opsin-mediated phototransduction. BMC Biol. 10:17.

Putnam NH, Srivastava M, Hellsten U, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Ryan JF, Pang K, Schnitzler CE, et al. 2013. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science 342:1242592–1242592.

Sars M. 1835. Beskrivelser og iagttagelser over nogle mærkelige eller nye i havet ved den bergenske kyst levende dyr af polypernes, acalephernes, radiaternes, annelidernes, og molluskernes classer: med en kort oversigt over de hidtil af forfatteren sammesteds fundne arter og deres forekommen. T. Hallager.

Schnitzler CE, Simmons DK, Pang K, Martindale MQ, Baxevanis AD. 2014. Expression of multiple Sox genes through embryonic development in the ctenophore Mnemiopsis leidyi is spatially restricted to zones of cell proliferation. EvoDevo 5:15.

Scholz CB, Technau U. 2003. The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian Nematostella vectensis (Anthozoa). Dev. Genes Evol. 212:563–570.

Seipel K, Yanze N, Schmid V. 2004. Developmental and evolutionary aspects of the basic helix-loop-helix transcription factors Atonal-like 1 and Achaete-scute homolog 2 in the jellyfish. Dev. Biol. 269:331–345.

Shiloh Y. 2001. ATM and ATR: networking cellular responses to DNA damage. Curr. Opin. Genet. Dev. 11:71–77.

Shinzato C, Iguchi A, Hayward DC, Technau U, Ball EE, Miller DJ. 2008. Sox genes in the coral Acropora millepora: divergent expression patterns reflect differences in developmental mechanisms within the Anthozoa. BMC Evol. Biol. 8:311.

Spring J, Yanze N, Middel AM, Stierwald M, Gröger H, Schmid V. 2000. The mesoderm specification factor twist in the life cycle of jellyfish. Dev. Biol. 228:363–375.

Wagner DE, Ho JJ, Reddien PW. 2012. Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. Cell Stem Cell 10:299–311.

Yamada T, Park CS, Mamonkin M, Lacorazza HD. 2009. Transcription factor ELF4 controls the proliferation and homing of CD8+ T cells via the Krüppel-like factors KLF4 and KLF2. Nat. Immunol. 10:618–626.