nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0719-8

In the format provided by the authors and unedited.

The genome of the jellyfish *Aurelia* and the evolution of animal complexity

David A. Gold[1,2,12*], Takeo Katsuki[3,12*], Yang Li[4], Xifeng Yan[4], Michael Regulski[5], David Ibberson[6], Thomas Holstein[7], Robert E. Steele[8], David K. Jacobs[9] and Ralph J. Greenspan[3,10,11,12*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Department of Earth and Planetary Sciences, University of California Davis, Davis, CA, USA. [3]Kavli Institute for Brain and Mind, University of California San Diego, La Jolla, CA, USA. [4]Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA, USA. [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [6]Deep Sequencing Core Facility, Cell Networks, Heidelberg University, Heidelberg, Germany. [7]Department of Molecular Evolution and Genomics, Centre for Organismal Studies, Heidelberg University, Heidelberg, Germany. [8]Department of Biological Chemistry and Developmental Biology Center, University of California Irvine, Irvine, CA, USA. [9]Department of Ecology and Evolution, University of California Los Angeles, Los Angeles, CA, USA. [10]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [11]Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA. [12]These authors contributed equally: D. A. Gold, T. Katsuki, R. J. Greenspan. *e-mail: dgold@ucdavis.edu; takeo.katsuki@gmail.com; rgreenspan@ucsd.edu

SI Guide document

SUPPLEMENTARY MATERIALS AND METHODS

DNA Collection. Total DNA was extracted from ephyrae using the following salting-out protocol. ~200 ephyrae were collected in a microcentrifuge tube and homogenized in 500 µl of extraction buffer containing 200 mM Tris-HCl (pH 8.0), 20 mM disodium EDTA, 1% SDS, 2 mg/ml RNase A (QIAGEN), and 20 mg/ml Proteinase K (QIAGEN) by pipetting with a 200 µl pipette tip. The homogenate was incubated at 70°C for 20 min with occasional mixing. Protein was precipitated by adding 70 µl of 8 M potassium acetate. The samples were then incubated on ice for 20 min, and centrifuged at 15,000 rpm for 20 min at 4°C. The supernatant was collected in a fresh tube and DNA was precipitated by adding 340 µl of 2-propanol. The samples were kept at room temperature for 5 min, and centrifuged at 15,000 rpm for 10 min at 4°C. The precipitate was rinsed with 1 ml of 70% EtOH and air-dried for 10 min after a final 5 min centrifugation at 15,000 rpm. The DNA pellet was dissolved in 10 mM Tris-HCl, pH 8.5 and stored at -20°C. For PacBio sequencing, additional purification using a Qiagen Genomic Tip G-20 column was performed following the manufacturer's instructions.

DNA Library Preparation and Sequencing. Libraries for PacBio sequencing were generated using SMRTbell Template Preparation Reagent Kits (Pacific Biosciences). Libraries were sequenced via 120-minute movies on 19 SMRT cells using the DNA Polymerase Binding Kit Version 2 with Version 2 sequencing chemistry on a PacBio RS II sequencer (UCSD IGM Genomics Center, La Jolla, CA). A mate-pair library with 4 kbp inserts was prepared using an Illumina Mate Pair Library Prep Kit, and 50 bp paired-end sequencing was performed on one lane of an Illumina HiSeq2000. A mate-pair library with 8 kbp inserts was prepared using a Nextera Mate Pair Sample Preparation Kit (Illumina), and sequenced on one lane of a

HiSeq2500 (Illumina) at a read length of 2x100 bp paired-end in High Output mode. For paired-end short read sequencing, a PCR-free paired-end library was prepared using a Illumina TruSeq PCR-free DNA Library Preparation Kit (Illumina) with a 350 bp target insert size. 2x250bp paired-end sequencing was performed using a HiSeq2500 (Illumina) in Rapid Run mode.

Test for sequence contamination. The most likely sources of contamination in our datasets are the brine shrimp *Artemia* (the food for *Aurelia* in the lab), humans, and prokaryotes. We began by downloading all ESTs and nucleotide sequences for *Artemia* available at NCBI as of October 13,2016 (40,968 sequences total). These data were used to create a BLAST database, and the *Aurelia* genome and transcriptome were queried using BLASTn with 98% identity and 90% coverage cutoffs. No gene models or genomic scaffolds were rejected based on their similarity to *Artemia* sequences. To check for human contamination, we repeated the process above using the human transcriptome from Ensembl (release 83: ftp://ftp.ensembl.org/pub/release-83/fasta/homo_sapiens/cdna/). No gene models or genomic scaffolds were rejected based on their similarity to *Homo* sequences. Finally, we repeated this process to compare our scaffolds to all bacterial sequences in NCBI. Two scaffolds (Seg1657 and Seg1661) were removed from the assembly as probable bacterial contaminants, based on significant overlap with bacterial genomes. According to taxonomic profiling of the SPROT top BLASTX hit for each gene (see Additional File 2), only 5% of our final gene models had a best match with a non-eukaryote (41 viruses, 485 bacteria, and 23 archaea out of 10,863 annotations). None of the vetted genomic scaffolds consisted entirely of gene models with best BLAST hits to non-eukaryotes. So while some of these genes are likely to represent contaminants, we couldn't rule out horizontal gene transfer. Given the fact that these genes represent a small fraction of the total genes, we retained them in the models and downstream analyses.

Genome-wide polymorphism. The 250 bp paired-end reads were mapped to the draft assembly using BWA-MEM (version 0.5.7)[1] and PCR duplicates were removed with Picard MarkDuplicates (http://broadinstitute.github.io/picard). Variants were called using GATK HaplotypeCaller (version 3.5) in -ERC GVCF mode[2]. SNPs were filtered with the following setting: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0. A genome-wide SNP rate was calculated by the following formula: SNP/(number of examined site - non-scored sites), which was 1210521/(755635914 - 116175933) * 100 = 0.1893036.

Identification of repeats and analyses of transposable-element activity. De novo identification of repetitive elements was performed with RepeatScout[3], using an l-mer size of l = 15 bp. We used TRF and NSEG to filter out sequences less than 50 bp long, as well as anything with more than 50% low-complexity regions. We identified 10,231 distinct repeat motifs that occurred ≥10 times in the genome. To annotate these motifs, we created a BLAST database from Repbase (v.23.01). The *Aurelia* repeats were queried against the database using TBLASTX, with an e-value cutoff of 10e-5. Statistics about each type of repetitive element (number of unique copies, total number of copies, number of base pairs covered) were calculated using the BLAST results as well as the gtf file produced by RepeatMasker. These statistics are provided in Table S5.

Isolation of mRNA, library preparation, and *de novo* transcriptome sequencing. For transcriptome sequencing, planula larvae were collected from a brooding female raised at the Cabrillo Aquarium in San Pedro, California. RNA was immediately isolated from approximately 300 motile, ovoid larvae as a "planula" sample. *Aurelia* polyps derived from these planulae were

raised at UCLA on a diet of *Artemia*, and 25 animals were collected for RNA extraction. In contrast to the DNA extraction protocol, where we used 5-methoxy-2-methylindole to induce strobilation, for the RNA extraction experiments we induced strobilation by incubating polyps in 1 mL of iodine per gallon of seawater for five days, changing the water every day. Strobila began to develop from polyps one week following treatment. The "early strobila" sample was defined by the presence of oral tentacles, which degenerate during metamorphosis. The "late strobila" sample lacked oral tentacles, and the oldest (oral-most) developing ephyra exhibited pulsing behavior. 25 animals were collected for RNA extraction for each stage. 72 hours following the collection of strobilae, metamorphosis had completed, and 30 ephyrae were collected for RNA extraction. Three weeks following the collection of ephyrae, some animals had developed a complete bell, and were sampled as "juvenile" medusas. Ten individuals, each with a bell approximately 2 cm in diameter, were collected for RNA isolation.

Aurelia total RNA was extracted using a phenol-chloroform protocol. Approximately 50 μg of tissue per sample was homogenized in 200 μl of a lysis solution (100 mM Tris/HCl, pH 5.5, 10 mM disodium EDTA, 0.1 M NaCl, 1% SDS, 1% β-mercaptoethanol), after which 2 μl of Proteinase K (25 mg/ml) was added, and the samples were incubated at 55°C for 10 minutes. The samples were then placed on ice, and combined with 11 μl of 0.2 M sodium acetate (pH ≈ 4), and 250 μl of a phenol:chloroform:isoamyl alcohol mixture (25:24:1). The samples were kept on ice for 15 minutes, and centrifuged for 15 minutes at 16,000 rpm and 4°C. The upper phases were collected and concentrated in 2-3 1.2 ml tubes per life stage (creating the biological replicates). An equal volume of 2-propanol was added to each tube, and the samples were left to precipitate at -80°C overnight. The following day, the samples were spun for 15 minutes at 16,000 rpm and 4°C. The supernatants were removed, and the resulting pellets were washed in

70% ethanol. The sample was centrifuged a final time for 5 minutes at 9,000 rpm and 4°C, the supernatants were removed, and the pellets were air-dried for ten minutes. The pellets were dissolved in RNase-free water, and pooled to produce 2-3 biological replicates per life stage. To remove any genomic DNA from the RNA samples, the samples were extracted using TRI Reagent (Sigma-Aldrich) according to the manufacturer's instructions. Details about each sample and the relevant NCBI SRA accessions are provided in Table S7.

Gene Prediction and Annotation. The gene annotation pipeline is graphically described in Figure S1. All RNA reads with a quality score less than 20 were removed using the Filter FastQ tool in Galaxy[4]. Filtered reads from the paired-end data were assembled into transcripts *de novo* using Trinity[5]. The resulting transcriptome contained 191,123 transcripts and 118,376 putative genes. For genome-guided gene prediction, The Trinity assembly and unmasked genome were used as input for the PASA genome annotation pipeline[6], producing a revised set of 113,705 genes models. An initial "vetted" dataset was made from these predictions for use in downstream gene annotation programs. "Vetted" genes had to meet the following criteria to be retained: [1] have coding regions greater than 300 nucleotides long, [2] have no ambiguous nucleotides, [3] have four or more exons, [4] contain 5' and 3' UTR regions, and [5] produce a full-length protein with start and stop codons. The peptide translations of for these vetted sequences were compared in an all-against-all BLASTp (evalue 1e-10); if two or more peptides had a hit, only the largest peptide was retained. This process was repeated until no peptide shared sequence similarity with any other peptide, resulting in 2,521 sequences.

We next hard-masked the genome using RepeatMasker[7], with a filtered repeat library generated by RepeatScout[3]. *Ab initio* gene prediction of the masked genome was performed using

GeneMark- ES[8] glimmerHMM[9] and the AUGUSTUS web server[10] with default settings. The "vetted" PASA dataset was used to create the input training file for AUGUSTUS as well as the exon file for glimmerHMM. In addition to *ab initio* gene prediction, we mapped the Uniprot Swissprot protein dataset (accessed August 15, 2016) to the genome using exonerate[11], and re-mapped all Trinity *de novo* transcripts to the masked genome using GMAP[12]. Five sets of gene models (GeneMark, GlimmerHMM, AUGUSTUS, exonerate + GMAP, and the original PASA output) were passed into EVidenceModeler[6] to create a weighted consensus gene structure dataset. EVidenceModeler was run with default settings, with the PASA assembly given ten times the weight of *ab initio* predictions, and the "exonerate + GMAP" dataset given twice the weight of *ab initio* predictions. The EVidenceModeler gene models were passed back into PASA to create a final set of 33,134 consensus predictions[6]. Following vetting of the gene predictions described in the Materials and Methods of the main text, we ended up with 29,964 vetted gene models. An annotation report from this pipeline is included as Additional File 2.

Ortholog Predictions. To determine protein homology, proteomes were downloaded from the following genomes at NCBI: *Acropora digitifera* (Assembly ID: GCA_000222465.2), *Branchiostoma floridae* (GCA_000003815.1), *Capitella teleta* (GCA_000328365.1), *Exaiptasia pallida* (GCA_001417965.1), *Homo sapiens* (GRCh38.p11), *Hydra vulgaris* (Hydra_RP_1.0), *Lottia gigantea* (GCA_000327385.1), and *Nematostella vectensis* (ASM20922v1). The *Drosophila* predicted proteins was downloaded from Uniprot (ID: UP000000803). These datasets were passed to OrthoFinder[13]. A Venn diagram illustrating the overlaps of homologous between various datasets is provided in Figure S2. We used the output of OrthoFinder to produce a binary (present/absent) matrix. After excluding taxon-specific orthologous groups, the binary

and complete matrices were used to generate the correlation matrix heat maps in Figure 3a (using the corrplot library in R).

Gene expansion/reduction analysis. To analyze gene family expansions/reductions, we used CAFE v2.2[14]. CAFE requires two input files; a time-calibrated tree (created from the BEAST output, see the previous "Molecular Clock Analysis" section), and a matrix of gene copies per taxon per family. The matrix was modified from the OrthoMCL output with several modifications. First, as suggested by the program's documentation, we removed all orthologous gene sets that were not present in both branches of the last of common ancestor (i.e. each gene set had to include at least one cnidarian and one bilaterian sequence). Secondly, in order to get a non-infinite value for lambda (the gene birth-death parameter), we removed orthologous gene sets where a single taxon was represented by more than 100 genes. The full output of this analysis is provided in Additional File 1, part 5.

Microsynteny Analysis. Microsynteny analysis was performed using MCScanX[15]. This program requires two input files, a "gff" file that maps the genes from each species onto their respective chromosomes/contigs, and a "homology" file that identifies pairs of homologous genes. The gff files for each species were modified from the GFF files associated with each NCBI genome project (see "Ortholog Predictions" above for a list of relent NCBI IDs). The "homology" files were modified from the OrthoFinder output. MCScanX_h was run on pairs of organisms, with a minimum number of three genes required to assign syntenic blocks. Examples of both input files are provided in Supplementary File 1, part 3, alongside the detailed output of MCScanX.

Homeobox Analysis. Genome assemblies and predicted peptide models for *Nematostella*, *Hydra*, *Exaiptasia*, and *Acropora* were downloaded from NCBI (BioProjects cited previously). Genome assemblies from these four cnidarians plus *Aurelia* were concatenated into a single file and formatted into a nucleotide BLAST database using the standalone BLAST 2.4.0+ suite[16], while 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 Additional File 1, part 6.1. Protein sequences were queried using BLASTp, and genomic seqeunces were queried using tBLASTn (using an e-value cutoff of 10e-5 and the *-outfmt "6 sseqid sseq"* flag to recover database hits). Each BLAST hit was vetted using HMMER and the Pfam-A database; sequences were only retained if the program identified a homeodomain.

To remove redundant homeoboxes, the results were de-multiplexed by species, and subjected to substring de-replication using CD-Hit v4.6[17] with the *-c .98* flag. The remaining sequences were concatenated into a single dataset, combined with annotated homeodomain sequences for *Homo sapiens*, *Branchiostoma floridae*, *Caenorhabditis elegans*, and *Drosophila melanogaster* downloaded from HomeoDB2[18]. These sequences were aligned using the standalone version of MUSCLE[19]. An initial round of homeobox annotation for the cnidarian sequences was done using the ortholog clustering method described previously in the “Gene homology predictions

and gene family" section. RaxML tree reconstruction was performed using the BlackBox web server[20], with a gamma model of rate heterogeneity and an LG substitution matrix. Node probabilities were calculated using 100 bootstraps. A Bayesian tree was produced using MrBayes v.3.2.6[21]. The program was run using a GTR substitution model with gamma-distributed rate variation across sites and a proportion of invariable sites ("lset nst=6 rates=invgamma") with four chains for 2 million generations. The run was sampled every 1000th generation, and analyzed with a 25% burnin. The MUSCLE alignment, as well as final RaxML and MrByes trees are provided in Additional File 1, part 6.

Molecular clock analysis. To create a molecular clock, we used all single-copy orthologs (SCOs) recovered from all species, based on our OrthoFinder analysis. Each SCO group was aligned using MUSCLE[19], and each alignment was cleaned using TrimAL[22]. These alignments were combined into a supermatrix using the FASconCAT software package[23]. Our final dataset included 10,787 amino acids representing 31 genes. A tree produced by this supermatrix using PhyML[24] recapitulated the expected relationships between species (Figure S5). This supermatrix was passed into BEAUTi[25] to generate a dataset for molecular clock analysis. The best model of amino acid evolution for each gene was determined using ProtTest 3[26]. We chose an uncorrelated relaxed clock model with a lognormal distribution, and a "speciation: yule process" tree prior for the pattern of speciation/extinction events. Because our tree is so taxon poor, we provided as many node calibrations as possible. Seven calibration points were set based on fossil constraints, described in Table S11. An additional uniform prior (250-480 Mya) was set for the divergence of *Nematostella* and *Exaiptatsia*, based on Schwentner and Bosch[27]. Posterior probability distributions from these calibration points are provided in Figure S4. The resulting BEAUTi XML file was used as input into BEAST v1.8[25], and MCMC analysis was performed for ten

million generations. A consensus tree from the output was made using the TreeAnnotator program packaged with BEAUTi/BEAST. The BEAUTi XML file (which includes the final amino acid alignment) and the consensus BEAST tree are provided in Additional File 1.

Phylogenetic annotation of differentially expressed genes. To produce Figure 5, we performed a stepwise series of BLAST comparisons. We first translated all differentially expressed *Aurelia* StringTie transcript models into putative proteins using TransDecoder[5]. We then downloaded all proteins from the Swissprot Uniprot database, excluding all sequences from cnidarians (accessed 3/25/18; Uniprot query: "*NOT taxonomy:"Cnidaria [6073]" AND reviewed:yes*"). We used BLASTP to query our protein models against this database (e-value cutoff 10e-5); any gene with a match was annotated as "Eukaryotic". All genes without a BLAST hit were queried against a second database that included all NCBI anthozoan proteins (accessed 3/25/18; NCBI query: "*"anthozoans"[porgn:__txid6101]*"). Proteins with a hit were annotated as "Cnidarian-specific". Finally, all remaining proteins without a BLAST hit were queried against a third database containing all medusozoan proteins in NCBI (accessed 3/25/18; NCBI query: "*"cnidarians"[porgn:__txid6073] NOT "anthozoans"[porgn:__txid6101]*"). Proteins with a hit were annotated as "Medusozoan-specific". All remaining proteins were annotated as *Aurelia*-specific "orphan" genes.

SUPPLEMENTARY REFERENCES

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
2. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491 (2011).
3. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
4. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
5. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
6. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
7. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* 4–10 (2009).
8. Lukashin, A. V. & Borodovsky, M. GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
9. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
10. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
11. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

12. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
13. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
14. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
15. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
16. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
17. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
18. Zhong, Y. & Holland, P. W. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol. Dev.* **13**, 567–568 (2011).
19. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
20. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **57**, 758–771 (2008).
21. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
22. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

23. Kück, P. & Meusemann, K. FASconCAT: Convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).

24. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

25. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

26. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).

27. Schwentner, M. & Bosch, T. C. Revisiting the age, evolutionary history and species level diversity of the genus Hydra (Cnidaria: Hydrozoa). *Mol. Phylogenet. Evol.* **91**, 41–55 (2015).

28. Chapman, J. A. *et al.* The dynamic genome of Hydra. *Nature* **464**, 592 (2010).

29. Shinzato, C., Inoue, M. & Kusakabe, M. A snapshot of a coral "holobiont": a transcriptome assembly of the scleractinian coral, Porites, captures a wide variety of genes from both the host and symbiotic zooxanthellae. *PLoS One* **9**, e85182 (2014).

30. Bellis, E. S., Howe, D. K. & Denver, D. R. Genome-wide polymorphism and signatures of selection in the symbiotic sea anemone Aiptasia. *BMC Genomics* **17**, 160 (2016).

31. Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *science* **317**, 86–94 (2007).

32. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**, 10 (2013).

33. Shinzato, C. *et al.* Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* **476**, 320 (2011).

34. Gold, D. A., Runnegar, B., Gehling, J. G. & Jacobs, D. K. Ancestral state reconstruction of ontogeny supports a bilaterian affinity for Dickinsonia. *Evol. Dev.* **17**, 315–324 (2015).

35. Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontol. Electron.* **18**, 1–106 (2015).

36. Chen, J.-Y., Huang, D.-Y. & Li, C.-W. An early Cambrian craniate-like chordate. *Nature* **402**, 518 (1999).

37. Cartwright, P. *et al.* Exceptionally preserved jellyfishes from the Middle Cambrian. *PloS One* **2**, e1121 (2007).

SUPPLEMENTARY FIGURES

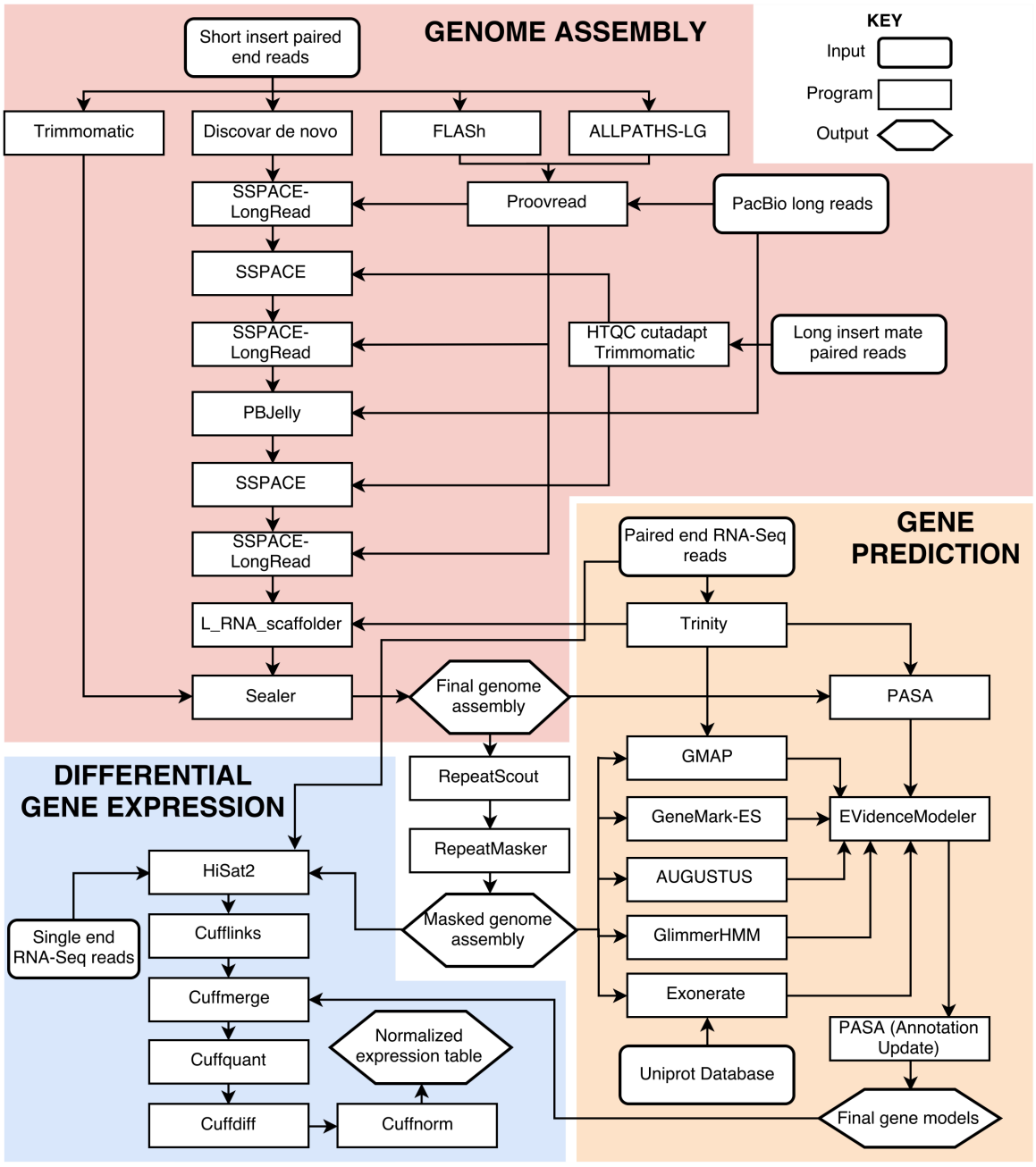

Figure S1. A flow chart describing the processes used for genome assembly, gene prediction, and differential gene expression analysis. The details of this pipeline are described in the Supplementary Materials and Methods.

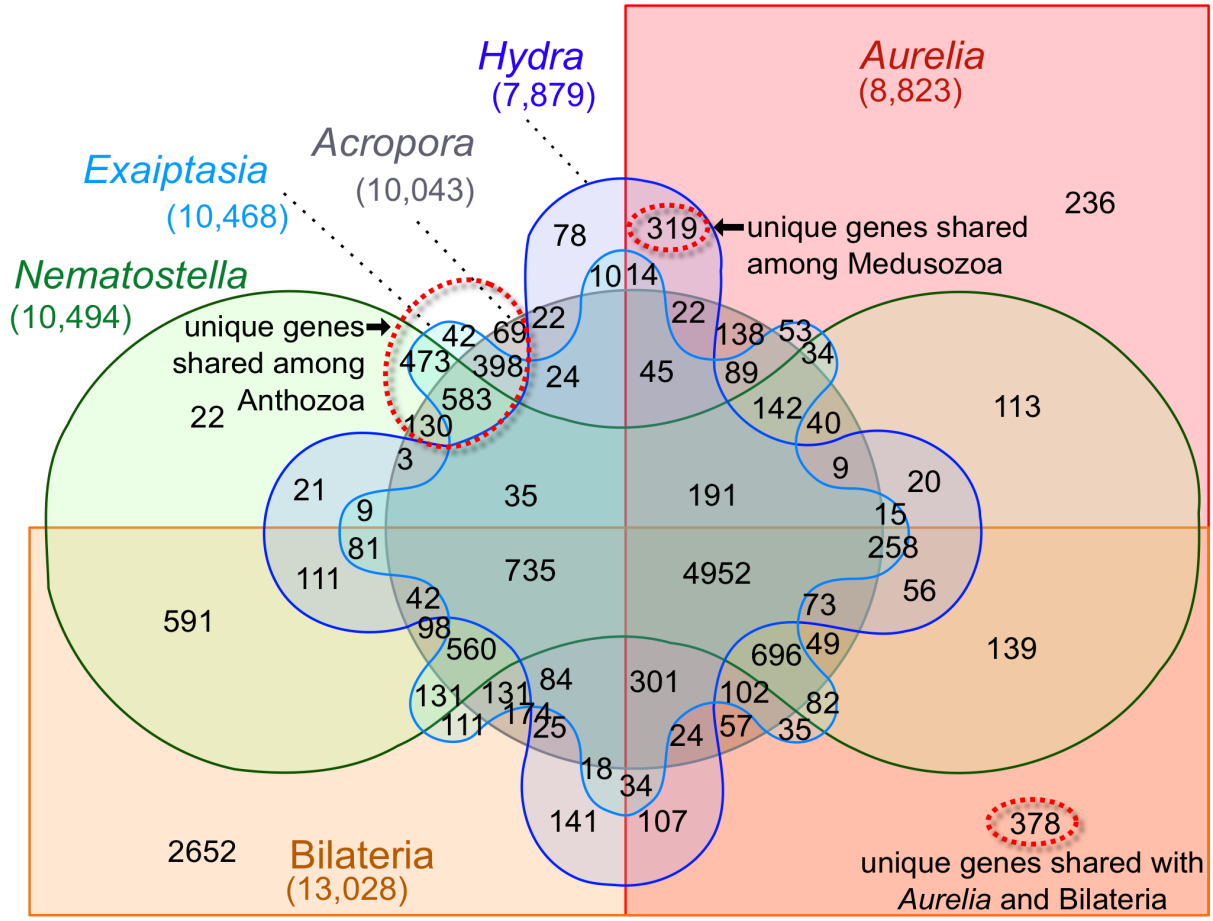

Figure S2. A Venn diagram showing orthologous gene overlap between organisms in this study. Ortholog groups were determined using OrthoFinder, and treated as binary "present/absent" data. For this image, all bilaterians have been combined into a single sample (i.e. if at least one bilaterian species has an ortholog, it is treated as "present").

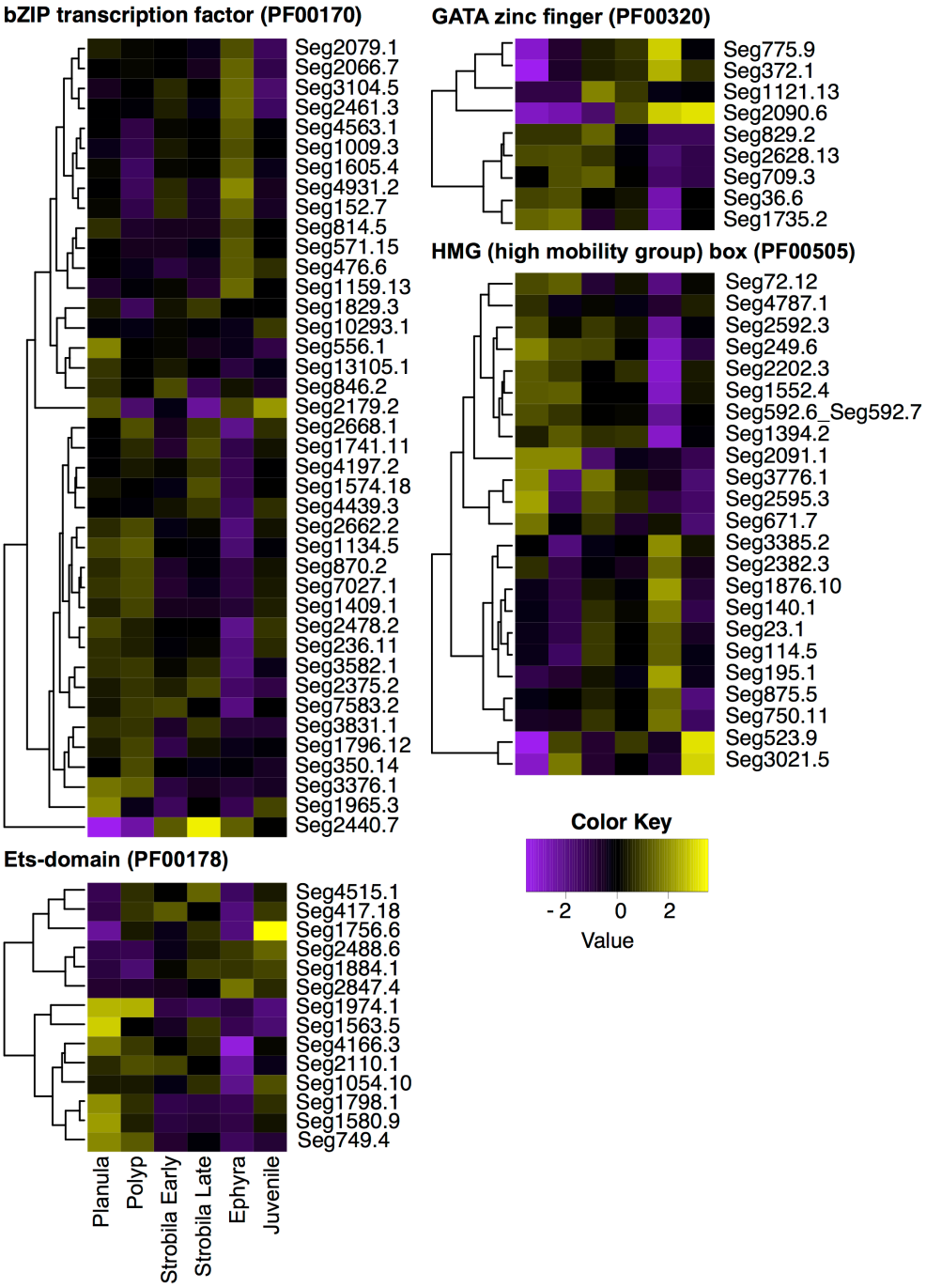

Figure S3. Heat maps for differentially expressed genes featuring overrepresented transcription factor domains. Overrepresentation was determined by comparing the *Aurelia* genome to other cnidarians (see Table S8). Differential expression was determined using EdgeR (see Additional File 1, part 8). Heat maps were produced using the PtR program packaged with Trinity[79]. TMM-normalized read counts were averaged across biological replicates, and then log transformed and centered around the row mean for each gene. Heat maps were produced using the following commands: "*PtR --matrix {matrix.TMM.averaged.FPKM} --heatmap --log2 --center_rows --sample_clust none*".

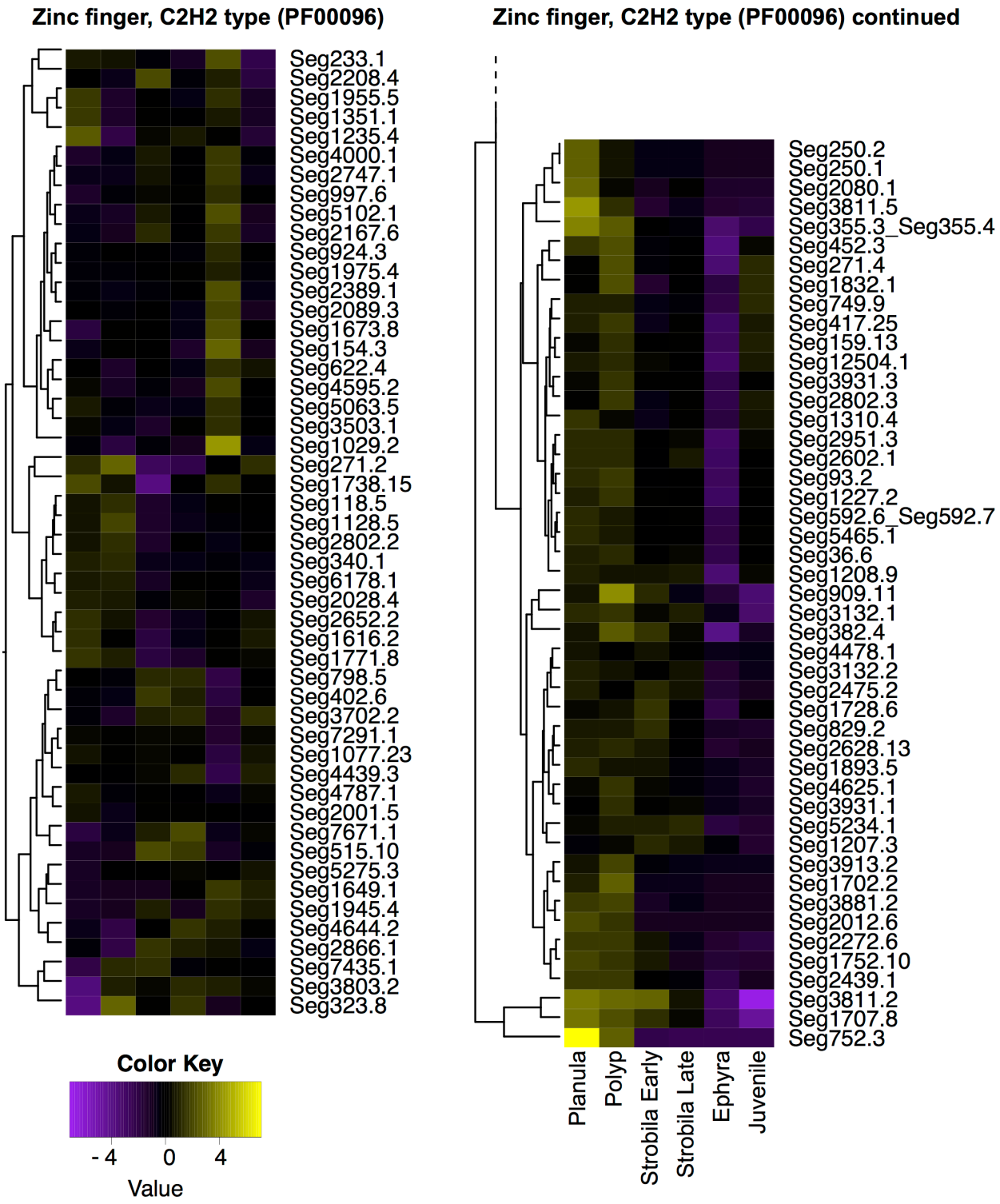

Figure S4. Heat map for differentially expressed genes featuring a zinc finger C2H2-type domain. The methodology for producing this heat map is described in the legend of Figure S3.

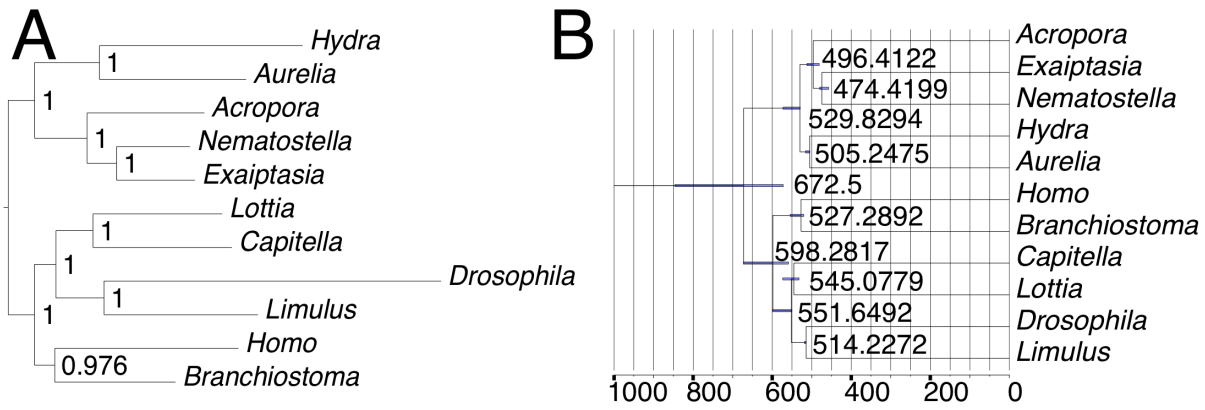

Figure S5. Animal trees derived from the single copy ortholog matrix. (A) A maximum likelihood tree made PhyML. Nodes represent likelihood values based on 100 bootstraps. Note that this tree gives the accepted species relationships. (B) Molecular clock generated from the same data using BEAST.

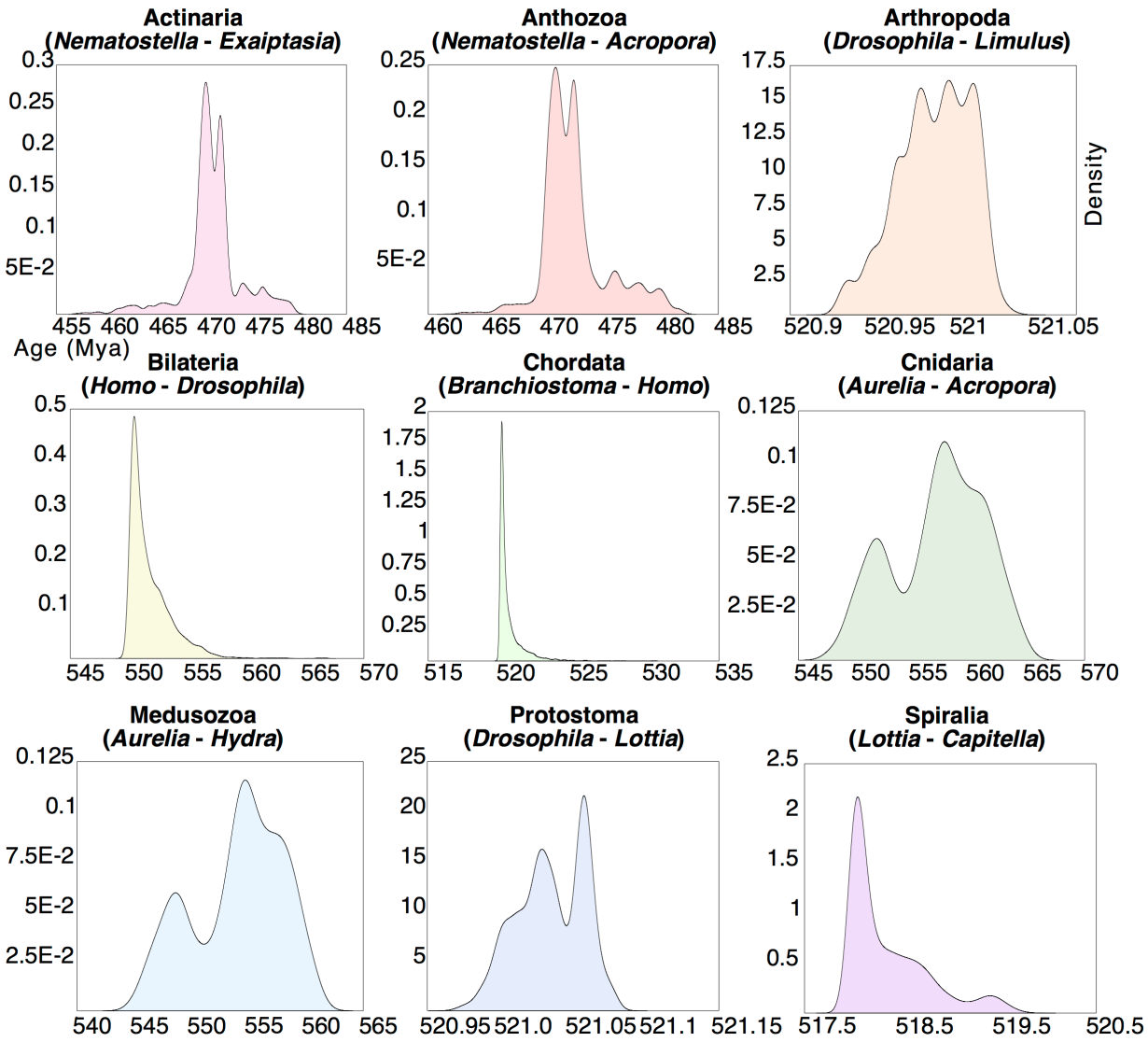

Figure S6. Marginal posterior probabilities of calibrated nodes. These are derived from the BEAST molecular clock analysis (see Additional File 1, part 4). These visualizations were produced in Tracer (v1.7).

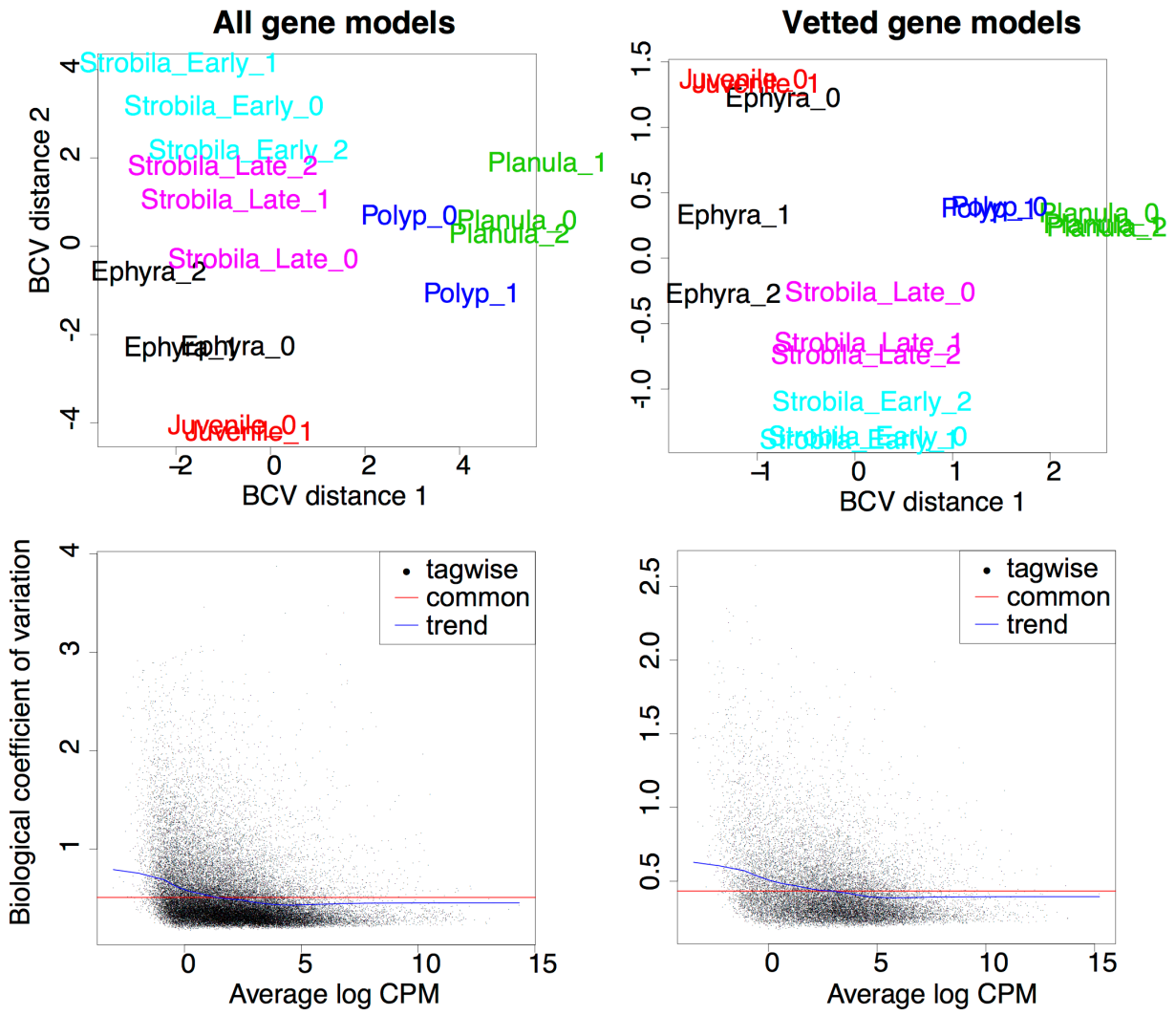

Figure S7. Variation within and across biological replicates. Top: Multidimensional scaling plots of distances between samples. Bottom: biological coefficient of variance (BCVs) plotted against gene abundance. These analyses are based off of the gene counts produced in StringTie and performed in EdgeR (see Additional File 1, part 7-8).

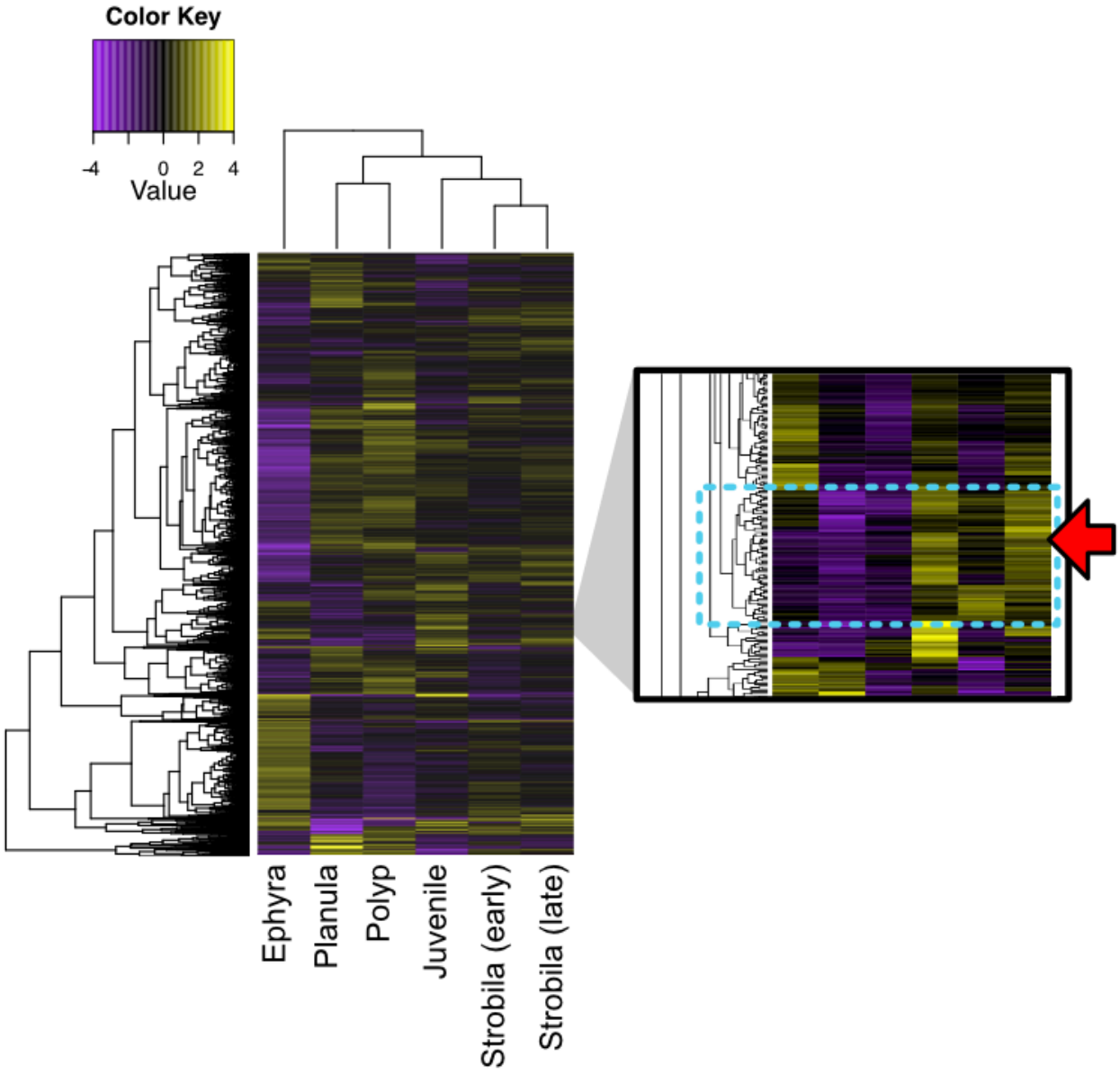

Figure S8. Heat map used for gene profile clustering. The window to the right expands the region associated with the *eyes absent* gene (marked with a red arrow). Genes with a similar gene expression profile (shown with a blue box) were used for the downstream analyses in Figure 6 of the main text. The list of relevant genes are provided in Additional File 1, part 11.2.

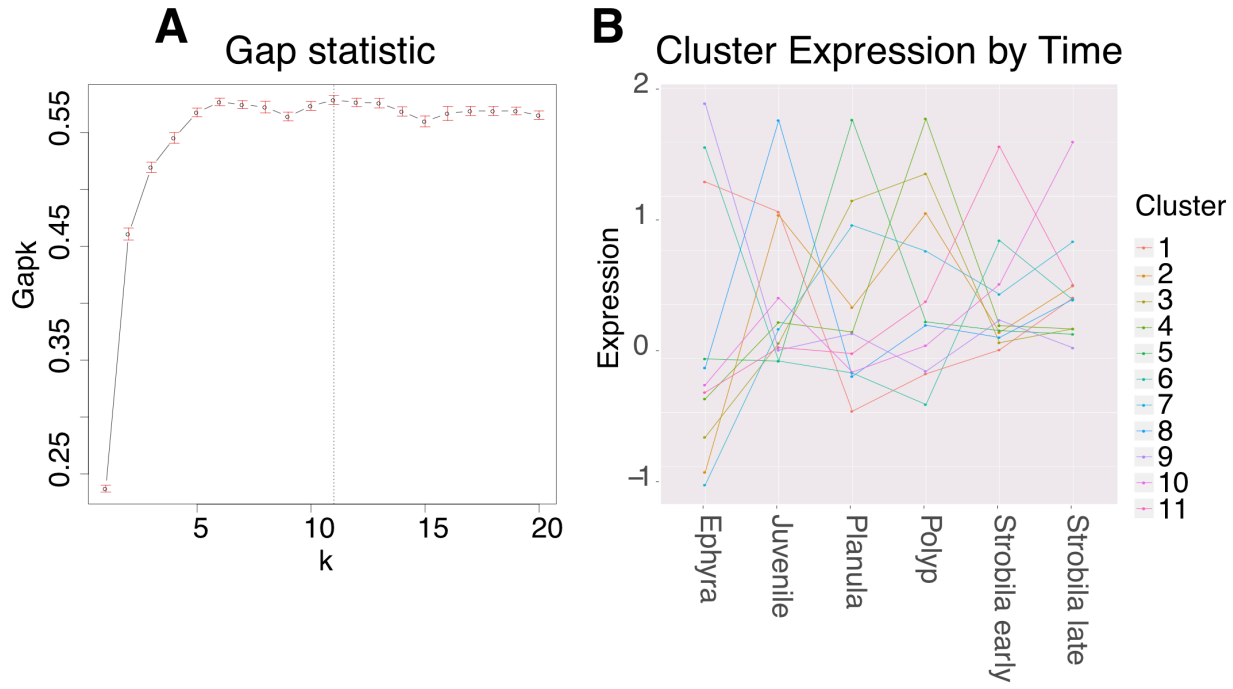

Figure S9. Clustering analysis of differentially expressed genes with putative homologs in other organisms. See Additional File 9 for the code used to produce this analysis, as well as the results of enrichment analysis based on these clusters. **a**, Optimal number of clusters based on "gap statistic" analysis. **b**, Expression profiles of the 11 clusters. Note that this output includes clusters that are upregulated in every life stage.

SUPPLEMENTARY TABLES

Table S1. Summary statistics for cnidarian genomes. Data collected from the relevant NCBI genomes (accessed March 2018), with the exception of SNP levels, which are taken from multiple references[28–31]. Statistics for *Aurelia* include all scaffolds and contigs greater than 2 kbp.

	Aurelia	***Acropora***	***Exaiptasia***	***Hydra***	***Nematostella***
Total sequence length (Mb)	747	447	256	1,260	357
Number of scaffolds	16,793	2,421	4,312	132,858	10,804
Scaffold N50 (kb)	124	484	442	65	473
Number of contigs	67,055	54,401	29,750	236,667	59,149
Contig N50 (kb)	20	11	14	13	20
GC Content (%)	32.6	40.5	29.8	27.6	41.9
Genome-wide SNP level (%)	0.19	0.4	0.3-0.5	0.7	6.5
Repetitive DNA (%)	50	13	26	57	26
Number of gene models	29,964	32,338	26,789	22,183	27,173

Table S2. Sequencing libraries generated from *Aurelia* gDNA, with associated statistics.

Library type	**Insert size (bp)**	**Library size (Gb)**	**Sequencer**	**Read length (bp)**	**Coverage**	**SRA accession number**
Paired-end[a]	220	379.7	HiSeq 2500	100	500X	SRR7889280-SRR7889284
Paired-end	400	88.2	HiSeq 2500	250	123X	SRR7866920
Mate-pair	4000	6.7	HiSeq 2000	50	9.4X	SRR7834587
Mate-pair	8000	37.8	HiSeq 2500	100	53X	SRR7866321
PacBio	1573	2.9	PacBio RSII	5471	4.1X	SRR7866923

Table S3. NG50 measures and number of scaffolds at each step of the assembly pipeline. Statistics were calculated as in the Assemblathon2 study[29] using "assemblathon_stats" Perl script (https://github.com/ucdavis-bioinformatics/assemblathon2-analysis) with its default options except that the genome size was set to 713 MB.

Assembler	Contig NG50	Scaffold NG50	Number of scaffolds
Discovar de novo	6934	7147	145165
SSPACE-LongRead	10128	14820	86306
SSPACE	13333	52625	41671
SSPACE-LongRead	13461	56749	38512
PBJelly	16567	57073	38400
SSPACE	16762	81615	28654
SSPACE-LongRead	16802	83255	27878
L_RNA_scaffolder	16802	122151	25455
Sealer	19245	121917	25455

Table S4. Gene model statistics.

All gene models	33,134
Gene models with protein prediction	25,858
Gene models with complete protein prediction (start and stop codons)	16,929
Gene models with RNA-Seq reads (sum count >10 across samples)	26,476
Dubious models (RNA-Seq count < 10 AND lack of "complete" peptide prediction AND lack of BLAST/HMMER annotation)	2,239
Putative rRNA (95%+ Megablast identity with rRNA sequence from RNAMMER OR BLASTP annotation)	330
Puative repetitive Element? (95%+ Megablast identity with repetative element from RepeatScout AND (BLASTP annotation OR no protein model)	617
Vetted gene models	29,964
Vetted gene models with protein predictions	23,986
Vetted gene models with RNA-Seq reads (sum count >10 across samples)	27,712

Table S5. Repetitive elements statistics, based on Repbase annotation.

Repetitive Elements	Unique Copies	Total Copies	BP Covered
Transposable Element			
DNA transposon			
Mariner/Tc1	32	6,044	1,931,031
hAT	63	4,282	1,697,719
MuDR	6	191	42,184
EnSpm/CACTA	8	115	49,338
piggyBac	13	238	103,874
P	19	396	162,266
Merlin	10	173	76,093
Harbinger	23	982	300,331
Transib	4	543	402,385
Novosib	0	0	0
Helitron	24	750	408,776
Polinton	3	47	18,431
Kolobok	77	3,619	1,124,057
ISL2EU	28	1,075	480,267
Crypton	2	25	9,645
CryptonA	0	0	0
CryptonF	0	0	0
CryptonI	0	0	0
CryptonS	0	0	0
CryptonV	5	145	66,603
Sola	0	0	0
Sola1	1	75	13,396
Sola2	28	461	188,484
Sola3	0	0	0
Zator	0	0	0
Ginger1	3	34	18,730
Ginger2/TDD	3	182	42,147
Academ	84	4,902	2,122,348
Zisupton	1	10	6,332
IS3EU	7	90	30,514
Dada	1	12	2,342
LTR Retrotransposon			
Gypsy	240	6,174	3,746,009

Copia	13	552	204,118
BEL	166	3,280	1,953,533
DIRS	110	3,079	1,260,516
Endogenous Retrovirus			
ERV1	2	26	7,993
ERV2	0	0	0
ERV3	0	0	0
Lentivirus	0	0	0
ERV4	0	0	0
Non-LTR Retrotransposon			
SINE	0	0	0
SINE1/7SL	0	0	0
SINE2/tRNA	0	0	0
SINE3/5S	0	0	0
SINE4	0	0	0
SINEU/snRNA	0	0	0
CRE	0	0	0
NeSL	0	0	0
R4	98	4,804	1,811,726
R2	0	0	0
L1	7	269	78,598
RTE	124	5,527	1,973,762
I	0	0	0
Jockey	7	401	111,782
CR1	418	79,489	30,969,508
Rex1	0	0	0
RandI	0	0	0
Penelope	176	19,563	7,006,078
Tx1	2	96	22,918
RTEX	90	3,793	1,427,488
Crack	180	17,805	8,072,548
Nimb	0	0	0
Proto1	0	0	0
Proto2	1	12	5,373
RTETP	0	0	0
Hero	13	296	86,297
L2	86	7,393	2,423,540
Tad1	0	0	0
Loa	2	21	5,327
Ingi	1	10	4,194

	Outcast	0	0	0
	R1	0	0	0
	Daphne	125	13,028	6,288,424
	L2A	239	35,849	16,017,735
	L2B	8	805	261,873
	Ambal	0	0	0
	Vingi	1	32	16,577
	Kiri	15	718	203,736
	Simple Repeat			
	Satellite	1	216	42,416
	SAT	4	714	233,444
	MSAT	1	67	21,344
	Multicopy gene			
	rRNA	1	32	8,916
	tRNA	0	0	0
	snRNA	1	20	3,202
	Integrated Virus			
	DNA Virus	0	0	0
	Caulimoviridae	0	0	0
	Nematostella-specific			
	NVREP236	1	136	34,949
	NVREP286	1	1,073	336,199
	NVREP353	2	193	57,215
	NVREP379	1	155	51,390
	NVREP444	3	250	65,458
	NVREP514	8	522	118,912
	NVREP537	1	147	46,113
	NVREP538	1	107	19,944
	NVREP579	1	50	19,441
Total		2,596	231,095	94,315,889

Table S6. Number of collinear gene blocks (>3 orthologous genes) between pairs of cnidarians, based on microsynteny analysis.

	Aau	**Adi**	**Epa**	**Hvu**	**Nve**
Aau	N/A	1,269	518	70	94
Adi	1,269	N/A	15,530	762	5,176
Epa	518	15,530	N/A	715	7,015
Hvu	70	762	715	N/A	163
Nve	94	5,176	7,015	163	N/A

Table S7. Basic statistics of RNA-Seq.

Sample	**SRA Accession**	**Type**	**Read length**	**Read number***	**% Map to Genome ^**
Planula 0	SRR7879514	Paired	100	56,875,671	53.40
Planula 1	SRR7879513	Paired	100	84,905,800	58.50
Planula 2	SRR7879512	Paired	100	50,701,454	65.27
Polyp 0	SRR7879511	Paired	100	227,510,586	58.00
Polyp 1	SRR7879510	Paired	100	100,957,192	70.39
Strobila ^	SRR7879507	Paired	100	31,550,545	--
Strobila Early 0	SRR7879506	Single	50	72,675,109	73.70
Strobila Early 1	SRR7879509	Single	50	94,730,961	71.37
Strobila Early 2	SRR7879508	Single	50	80,594,517	72.33
Strobila Late 0	SRR7879504	Single	50	48,806,129	62.14
Strobila Late 1	SRR7879503	Single	50	77,497,681	73.48
Strobila Late 2	SRR7879505	Single	50	84,544,140	73.63
Ephyra 0	SRR7879519	Single	50	60,846,424	53.40
Ephyra 1	SRR7879518	Single	50	69,611,757	37.29
Ephyra 2	SRR7879517	Single	50	50,199,336	36.35
Juvenile 0	SRR7879516	Paired	100	30,988,926	70.52
Juvenile 1	SRR7879515	Paired	100	12,764,628	63.85

* For paired-end data, number refers to forward reads only.

^ Only used for *de novo* transcript assembly. Not used for differential gene expression.

Table S8: Comparison of the number of proteins with transcription factor-related domains of *Aurelia aurita/sp.1* (*Aau*), *Acropora digitifera* (*Adi*), *Nematostella vectensis* (*Nve*) and *Hydra vulgaris* (*Hvu*). The counts for cnidarians other than *Aurelia* come from the *Acropora* genome paper[33]. Counts with an asterisk were noted as overrepresented in *Aurelia*, and further analyzed for differential gene expression (Figures S3-S4).

domain name	**accession**	**description**	***Aau***	***Adi***	***Nve***	***Hvu***
HLH	PF00010	Helix-loop-helix DNA-binding domain	57	52	72	32
Homeobox	PF00046	Homeobox domain	99	97	155	43
Hormone_recep	PF00104	Ligand-binding domain of nuclear hormone receptor	11	9	21	7
Pou	PF00157	Pou domain - N-terminal to homeobox domain	5	4	6	3
bZIP_1	PF00170	bZIP transcription factor	98*	24	38	24
Ets	PF00178	Ets-domain	23*	12	16	9
Forkhead	PF00250	Fork head domain	32	22	34	15
PAX	PF00292	'Paired box' domain	2	8	9	27
SRF-TF	PF00319	SRF-type transcription factor (DNA-binding and dimerisation domain)	4	1	4	2
GATA	PF00320	GATA zinc finger	18*	5	5	5
HMG_box	PF00505	HMG (high mobility group) box	48*	26	35	33
RHD_DNA_bind	PF00554	Rel homology domain (RHD)	2	2	3	1
DM	PF00751	DM DNA binding domain	14	8	12	6
Runt	PF00853	Runt domain	1	1	1	1
P53	PF00870	P53 DNA-binding domain	3	3	3	2
T-box	PF00907	T-box	12	10	16	7
ARID	PF01388	ARID/BRIGHT DNA binding domain	8	5	6	7

Basic	PF01586	Myogenic Basic domain	0	0	0	0
AT_hook	PF02178	AT hook motif	3	0	0	0
CUT	PF02376	CUT domain	1	1	2	0
TF_AP-2	PF03299	Transcription factor AP-2	5	2	1	1
TF_Otx	PF03529	Otx1 transcription factor	0	0	0	0
GCM	PF03615	GCM motif protein	1	1	2	0
OAR	PF03826	OAR domain	5	5	4	0
Prox1	PF05044	Homeobox prospero-like protein (PROX1)	0	0	0	0
SIM_C	PF06621	Single-minded protein C-terminus	0	0	0	0
Hairy_orange	PF07527	Hairy Orange	8	7	6	0
P53_tetramer	PF07710	P53 tetramerisation motif	3	1	1	0
bZIP_2	PF07716	Basic region leucine zipper	53*	17	34	20
zf-C2H2	PF00096	Zinc finger, C2H2 type	278*	88	209	106
zf-C4	PF00105	Zinc finger, C4 type (two domains)	13	12	19	8
zf-C2HC	PF01530	Zinc finger, C2HC type	4	4	4	2
SCAN	PF02023	SCAN domain	4	4	0	0

Table S9. Comparison of the number of proteins with signaling molecule-related domains of *Aurelia aurita/sp.1* (*Aau*), *Acropora digitifera* (*Adi*), *Nematostella vectensis* (*Nve*) and *Hydra vulgaris* (*Hvu*). The counts for cnidarians other than *Aurelia* come from the *Acropora* genome paper[33].

domain name	**accession**	**Description**	***Aau***	***Adi***	***Nve***	***Hvu***
MCPsignal	PF00015	Methyl-accepting chemotaxis protein (MCP) signaling domain	10	1	15	7
Wnt	PF00110	wnt family	17	15	29	11
G-alpha	PF00503	G-protein alpha subunit	53	18	34	19
RGS	PF00615	Regulator of G protein signaling domain	16	11	13	13
G-gamma	PF00631	GGL domain	7	3	4	2
HAMP	PF00672	HAMP domain	1	0	6	0
DIX	PF00778	DIX domain	2	2	4	1
STAT_alpha	PF01017	STAT protein, all-alpha domain	5	1	2	0
CheW	PF01584	CheW-like domain	0	0	1	0
Hpt	PF01627	Hpt domain	0	0	1	0
Cbl_N	PF02262	CBL proto-oncogene N-terminal domain	**3**	1	1	1
Dishevelled	PF02377	Dishevelled specific domain	1	0	0	0
Cbl_N2	PF02761	CBL proto-oncogene N-terminus, EF hand-like domain	1	1	0	1
Cbl_N3	PF02762	CBL proto-oncogene N-terminus, SH2-like domain	1	1	0	1
STAT_bind	PF02864	STAT protein, DNA binding domain	0	1	1	1
STAT_int	PF02865	STAT protein, protein interaction domain	1	1	0	1
NPH3	PF03000	NPH3 family	2	0	0	0
Focal_AT	PF03623	Focal adhesion targeting region	9	0	1	2
Olfactory_mark	PF06554	Olfactory marker protein	0	0	0	0
Phe_ZIP	PF08916	Phenylalanine zipper	1	1	0	0
TRADD_N	PF09034	TRADD, N-terminal domain	4	0	0	0

TGF_beta	PF00019	Transforming growth factor beta like domain	11	10	9	11
FGF	PF00167	Fibroblast growth factor	16	13	14	13
PDGF	PF00341	Platelet-derived growth factor (PDGF)	1	0	0	1
TGFb_propeptide	PF00688	TGF-beta propeptide	8	11	6	10
IL2	PF00715	Interleukin	5	2	0	0
IL4	PF00727	Interleukin	1	4	0	0
PTN_MK_C	PF01091	PTN/MK heparin-binding protein family, C-terminal domain	1	0	0	0
GM_CSF	PF01109	Granulocyte-macrophage colony-stimulating factor	0	0	0	0
IL7	PF01415	Interleukin 7/9 family	1	0	0	0
IL5	PF02025	Interleukin	0	5	0	0
IL3	PF02059	Interleukin-3	0	0	0	0
IL12	PF03039	Interleukin-12 alpha subunit	1	0	0	0
Rabaptin	PF03528	Rabaptin	2	1	1	0
PDGF_N	PF04692	Platelet-derived growth factor, N terminal region	1	0	0	0
AMH_N	PF04709	Anti-Mullerian hormone, N terminal region	0	0	0	0
PTN_MK_N	PF05196	PTN/MK heparin-binding protein family, N-terminal domain	0	0	1	0
CSF-1	PF05337	Macrophage colony stimulating factor-1	0	0	0	0
PSK	PF06404	Phytosulfokine precursor protein (PSK)	0	0	0	0
IL11	PF07400	Interleukin	0	11	0	0

Table S10. Classification and gene counts for cnidarian homeodomains.

Class	Family (+ *Drosophila* homolog)	*Acropora*	*Exaiptasia*	*Nematostella*	*Aurelia*	*Hydra*
ANTP (HOX)						
	Evx (eve)	1	0	0	1	0
	Gbx (unpg)	1	1	1	0	0
	Gsx (ind)	1	1	1	1	1
	Hox1 (lab)	1	1	1	1	1
	Hox-like	3	4	5	3	1
	Hox6-8	0	0	0	1	0
	Hox9-13/15 (Abd-B)	1	2	2	5	4
	Meox (btn)	3	4	4	0	1
	Mnx (exex)	1	1	1	0	0
	Ro (ro)	0	1	1	0	0
	Total	12	15	16	12	8
ANTP (OTHER)						
	Abox	1	2	4	0	0
	Barx/Bsx (bsh)	3	4	4	1	1
	Bari/Barhl (CG11085)	9	5	6	3	0
	Dbx/Hlx	2	3	3	1	0
	Dlx (Dll)	1	1	1	3	2
	EMX (Es,ems)	2	2	2	1	0
	Hhex (CG7056)	2	1	1	1	1
	Lbx/Ventx (lbe,lbl)	1	1	1	0	1
	Msx (Dr)	1	1	1	1	1
	Mxlx (CG1696)	2	2	2	1	0
	Nkx1 (slou)	1	1	1	3	1
	Nkx2 (scro,vnd)/Nkx4 (tin)	6	7	7	2	1
	Nkx3 (bap)	1	1	1	2	0
	Nkx5 (Hmx)	1	1	1	1	0
	Nkx6 (Hgtx)	1	1	1	1	0
	Nkx7 (Nk7.1)	1	1	1	1	0
	Vax	1	2	2	1	0
	Nedx (CG13424)	2	1	2	0	0

	Noto (CG18599)	7	5	7	1	1
	Unclassified	1	3	9	2	0
	Total	46	45	57	26	9
CERS						
	Cers/Lag (Lag1)	1	1	1	1	1
CUT						
	Onecut (ct)	1	0	1	0	0
	Cmp (dve)	1	2	1	0	0
	Total	2	2	2	0	0
HNF						
	Hnf1/2	0	1	1	0	0
LIM						
	ISL	1	1	1	1	0
	Lhx1/5 (Lim1)	1	1	1	1	1
	Lhx2/9 (ap)	1	1	1	1	1
	Lhx6/8 (Awh)	1	1	1	1	1
	Lmx (CG4328,CG32105)	1	1	1	1	1
	Unclassified	0	1	1	0	0
	Total	5	6	6	5	4
POU						
	POU1	1	1	1	1	1
	POU3 (vvl)	2	2	2	1	0
	POU4 (acj6)	1	1	1	1	1
	POU6 (pdm3)	1	1	1	1	1
	Total	5	5	5	4	3
PRD						
	Alx	1	2	1	1	0
	Arx (al,php13)	1	1	1	1	1
	DMBX	7	5	8	1	1
	Dux	4	3	8	0	0
	GSC (Gsc)	1	1	1	1	1
	Hbn (hbn)	1	1	1	2	1
	Nobox	1	1	1	1	1
	Otp (otp)	1	1	1	1	2
	Otx (oc)	6	4	3	7	3
	Pax3/7 (gsb,prd)	2	1	2	0	0
	Pax4/6 (ey,toy,toe,eyg)	1	2	2	1	1
	Pitx (Ptx1)	1	1	1	1	1
	Prop (CG32532)	1	1	1	0	0

	Rax (Rx)	1	1	1	1	0
	Repo (repo)	1	1	1	1	1
	Shox (CG34367)	1	1	1	0	1
	Uncx (OdsH,unc-4)	3	3	2	1	1
	Vsx (Vsx1,Vsx2,tup)	2	2	2	4	2
	Unclassified	7	4	4	2	0
	Total	43	36	42	26	17
TALE						
	Atale	1	1	1	1	0
	Irx (mirr,ara,caup)	1	1	1	1	1
	Meis (hth)	1	1	1	0	0
	Pbx (exd)	1	1	1	0	1
	Pknox	0	1	1	1	0
	Tgif (achi,vis)	1	1	1	0	0
	Unclassified	0	1	0	0	0
	Total	5	7	6	3	2
SINE						
	Six1/2 (so)	1	1	1	1	0
	Six3/6 (Optix)	1	1	1	0	0
	Six4/5 (Six4)	1	1	2	1	1
	Unlcassified	1	1	1	1	0
	Total	4	4	5	3	1
UNCLASSIFIED						
	Total	4	6	6	10	2

Table S11. Fossil calibrations used for molecular clock analysis.

Clade	**Taxon 1**	**Taxon 2**	**Offset**	**Fossil**	**Ref.**
Anthozoa	*Acropora*	*Nematostella*	420	kilbuchophyllids	[34]
Arthropoda	*Drosophila*	*Limulus*	514	*Yicaris*	[35]
Protostoma	*Lottia*	*Drosophila*	550.25	*Kimberella*	[35]
Chordata	*Homo*	*Branchiostoma*	518	*Haikouella*	[36]
Cnidaria	*Aurelia*	*Nematostella*	529	*Olivooides*	[35]
Medusozoa	*Aurelia*	*Hydra*	505	fossil medusas	[37]
Spiralia	*Capitella*	*Lottia*	532	*Aldanella*	[35]