Supplementary Information

The *Amphimedon queenslandica* genome and the evolution of animal complexity

Mansi Srivastava[1][2], Oleg Simakov[3]; Jarrod Chapman[4], Bryony Fahey[5], Marie E.A. Gauthier[5][6], Therese Mitros[1], Gemma S. Richards[5][11]; Cecilia Conaco[7], Michael Dacre[8], Uffe Hellsten[4], Claire Larroux[5], Nicholas H. Putnam[9], Mario Stanke[10], Maja Adamska[5][11], Aaron Darling[12], Sandie M. Degnan[5], Todd H. Oakley[13], David C. Plachetzki[13], Yufeng Zhai[8]; Marcin Adamski[5][11], Andrew Calcino[5], Scott F. Cummins[5], David M. Goodstein[4], Christina Harris[5], Daniel J.Jackson[5][14], Sally P. Leys[15], Shengqiang Shu[4], Ben J. Woodcroft[5]; Michel Vervoort[16], Kenneth S. Kosik[7], Gerard Manning[8], Bernard M. Degnan[5] and Daniel S. Rokhsar[1,4]

Table of Contents

S1. Background Information on *Amphimedon queenslandica*

S1.1 More about the animal

Amphimedon queenslandica (Niphatidae, Haplosclerida, Demospongiae, Porifera) was formally described in 2006 from Heron and One Tree Island Reefs, southern Great Barrier Reef, Queensland, Australia and named for the type locality of Queensland.[1] There are approximately 50 valid species in the genus *Amphimedon* worldwide (see http://www.vliz.be/vmdcdata/porifera/), although the relationship of these to each other and *A. queenslandica* remains largely unknown. *A. queenslandica* has been found on a number of reefs in the central and northern regions of the Great Barrier Reef and has high sequence similarity with an *Amphimedon* species collected from Dahab, Egypt, suggesting a even wider distribution (unpublished).

On the southern Great Barrier Reef, *A. queenslandica* has a patchy distribution, living on the shallow intertidal reef flat and crest. There, the grey-blue to green adult grows over decaying coral or other substrata (Figure 1a). The thick encrusting body form rarely exceeds 20 cm in diameter. Brood chambers are up to 1 cm in diameter and contain up to 150 mixed-staged embryos (Figure 1b). Brood chambers can be found in adults any time of the year, although fecundity increases in the summer months October to April.[2] Embryos are brooded until they hatch as parenchymella larvae (Figure 1c). Year-round access to embryos and larvae and the ability to readily rear and stage live embryos, larvae, postlarvae (Figure 1d) and juveniles (Figure 1e) have made *A. queenslandica* a model for the study of demosponge development. In addition, *Amphimedon* is relatively robust compared to many other sponges, and can be transported and handled for months without marked deterioration. Whether they can be transported worldwide is currently being tested (see Degnan *et al*. 2009[3] for detailed description of the ecology, life history and development of *A. queenslandica*).

Amphimedon cells are small (approximately 8 μm diameter), and to date, no gene transfer system has been developed. *In situ* hybridization protocols for embryonic and larval gene expression, however, have been worked out,[4] so that spatiotemporal expression patterns can be readily observed in the developing embryo and larva.

S1.2 Description of Figure 1a-e in the main paper

An adult *Amphimedon queenslandica* (Figure 1a) was photographed *in situ* in about 1 m of water at Shark Bay, Heron Island Reef, Great Barrier Reef. Brood chambers (Figure 1b) on the basal side of the adult were dissected by making 7-10 mm slices through the adult and photographed under a stereomicroscope. Multiple stages of embryogenesis can be observed, including spot and ring stage embryos. A whole mounted fixed larva cleared in benzyl alcohol/benzyl benzoate[4] was photographed under a compound microscope with DIC (Figure 1c). The anterior swimming pole is to the left and the posterior pigment ring to the right. The inner cell mass, middle subepithelial layer and the epithelial layer are evident. A newly settled postlarvae photographed under a stereomicroscope (Figure 1d). In the laboratory, settlement can be induced in competent larvae about 4 hours after emerging from the adult.[5] *A. queenslandica* larvae attach to the substratum by the anterior end, flatten and immediately begin metamorphosis and resorbing the pigment ring. 3

day old juvenile viewed under a stereomicroscope (Figure 1e). Numerous choanocyte chambers are present and the first large osculum is protruding from the apical surface (see Leys and Degnan 2001[2], 2002[6]; Adamska *et al*. 2007[7] for further description of embryogenesis, larval development and metamorphosis).

S2. Genome Sequencing and Assembly

S2.1 Source of materials

Genomic DNA was isolated from ~1,500 embryos and larvae obtained from the brood chambers of a single mother sponge by the SDS/Proteinase K lysis method.[8] The identity and number of fathers of this brood are unknown, although the genotyping of other brood chambers with multiple polymorphic microsatellite loci confirms that there are multiple fathers per brood chamber (S. Degnan, unpublished; Degnan et al. 2008[9]). Prior to lysis, embryos and larvae were extensively washed in filtered seawater to remove associated external contaminants. High molecular weight genomic DNA was sent from The University of Queensland to JGI for sequencing.

To construct *Amphimedon queenslandica* cDNA libraries for expressed sequence tag (EST) analysis, total RNA was isolated from larvae obtained from multiple adults collected from Heron Island Reef, Great Barrier Reef. Larvae were released from adults maintained in large containers of sea water at 28°C. Isolated larvae were transferred to and washed in 0.2-µm-filtered seawater at 22°C. These were then placed in RNALater (Ambion Inc.) and transferred to The University of Queensland and JGI, where RNA was extracted and cDNA libraries were constructed.

S2.2 Shotgun dataset

The *Amphimedon* whole genome shotgun (WGS) data set comprises paired end reads from seven libraries, including ~3 kb and ~7 kb insert plasmid libraries and three ~35 kb insert fosmid libraries (Table S2.2.1). In total, 2.92M reads were sequenced; all trace data is deposited at the NCBI Trace Archive (accession ACUQ00000000). A total of 2.21M (75.7%) passed stringent quality and vector trimming protocols. All reads were trimmed for quality using the JTRIM15 protocol (after masking vector with CrossMatch).[10] The JTRIM algorithm finds the subsequence within a read with the maximum expected alignment score to an idealized reference given the Phred base-quality scores of the read and a specified match/mismatch penalty (+1/-30.6). "Passing" reads with trimmed length of at least 400 bases and a mate pair of at least 400 (trimmed) bases were used in the assembly. The passing sequences are summarized in Table S2.2.1.

S2.3 Genome assembly

The *Amphimedon* genome was assembled using a custom approach developed for polymorphic

genomes. In brief, passing reads are aligned in all pairwise combinations; clusters of overlapping reads are identified as likely derived from the same genomic locus; contigs are formed by local assemblies of clusters, which are subsequently ordered and oriented into scaffolds.

The following methods were used for alignment, clustering, and assembly of reads:

- Alignment: Read-to-read pairwise alignments were calculated using the MALIGN aligner module.[10,11] MALIGN used the co-occurrence of at least 16 distinct 15-mers between pairs of reads to trigger banded, semi-global Needleman-Wunsch alignment. 15-mers occurring more than 80 times in the data set were not allowed to trigger alignments.
- Clustering: Alignments of at least 100bp and 95% identity were used to define the n-ring neighborhood sizes (n=1,2,3,4) for all reads (using the ringer3 perl script[12]) (Figure S2.3.1). Single-linkage clustering was performed using read-read alignments of at least 100 bp and 99% identity (using the make_clusters program,[12]). Read-read alignments were rejected if both reads had a 2-ring neighborhood of more than 60 reads. (This step excludes highly repetitive regions.) 28,308 clusters of at least 5 reads were generated containing 1,551,646 reads.
- Contiging and Scaffolding: Each read-cluster was assembled with phrap (version 0.960731) using parameters -minmatch 35 -minscore 55. Quality scores were not used, as this allows distinct haplotypes to be assembled together (data not shown). The resulting contigs were ordered and oriented (*i.e.*, "scaffolded") using the perl script phrapOut2Scaffolds[12] in which contigs are iteratively merged via a greedy ordering based on the number and consistency of read-pair linkages between them.

The bulk statistics for the assembly of the *Amphimedon* genome are reported in Tables S2.3.1, S2.3.2 and S2.3.3. Half of the genome is captured in 310 scaffolds longer than 120kb or 2,652 contigs longer than 11.2kb. "Captured" gaps comprise 21,996,065 bp (13.2%) of the total scaffold sequence. The mean gap size is 1,500 bp; the median gap size is 650 bp.

A rough estimate of the true genome size can be made as follows. The assembly contains 145 Mb of contigs with an estimated ~10% residual redundancy (estimated below). It accounts for 70% of the shotgun read dataset. Assuming that the unassembled reads corresponding to repetitive and/or heterochromatic regions are shotgun sampled at the same rate as the assembled regions, the genome size is then approximately 0.9 x 145 Mbp/0.7 = 190 Mbp. We note also that there may be significant haplotype-unique sequence in the sponge, as found in other heterozygous genomes (*e.g.*, *Ciona savignyi*[13,14] and *Vitis vinifera*[15]).

S2.4 EST sequencing, clustering, and mapping

Three cDNA libraries were constructed for paired end EST sequencing (Table S2.4.1). Library CABF was prepared at UQ; libraries CAYH and CAYI were prepared at JGI. Sequencing was performed with standard JGI protocols[16,17] on ABI 3730 and GE MegaBACE sequencing instruments. All ESTs that passed quality and vector filters were assembled using a custom EST clustering and assembly pipeline developed at JGI (Brokstein *et al.*, unpublished). A total of 66,375 EST sequences produced 15,333 consensus sequences assembled from 2 or more ESTs,

and 975 singlets (16,308 total). Once aligned to the genome the maximal assemblies were found using the PASA algorithm described in Haas et al. 2003.[18]

The 66,375 processed/filtered ESTs plus an additional 70 full length cDNAs from the Degnan lab and 38 genbank mRNAs were aligned to the genome and assembled using the PASA pipeline. Sequences were trimmed for length, vector and DUST - 19,936 sequences were trimmed and 66,412 sequences remained after seqclean.[19] ESTs were aligned to their best hit in the genome using gmap.[20] These alignments were then evaluated by PASA[18] to ensure valid splice sites, and alignment over 90% coverage, 90% identity. If gmap alignment did not meet validation criteria, then sim4[21] was used. 62,767 (94.5%) ESTs had some alignment to the genome. 52,818 (79.5%) had an alignment that met the validation criteria of 90% identity, 90% coverage, intron length <= 55 kb, and valid splice sites. The validated alignments assembled into 9,699 assemblies, creating 8,478 subclusters (loci). 7,261 of the assemblies are comprised of two or more ESTs. 2,713 assemblies appear to be nominally complete genes, with start and stop codons and at least 150 bp of coding sequence.

S2.5 Evaluation of completeness and correctness of genome assembly

Comparison to ESTs

66,375 *A. queenslandica* ESTs were clustered and assembled into 16,308 contigs via the JGI EST pipeline. Of these, 4,861 contigs were found to have a complete (start codon to stop codon) ORF of at least 450 bp. Of these putatively full-length EST contigs, 3,375 had a hit to a human Refseq gene on the correct strand under blastx (-e 1e-5). 3,354 (99.4%) of these had a blat alignment (under default blat settings) to the assembled scaffolds. 3,275 (97.0%) had an alignment spanning the midpoint of the EST contig. 3,140 (93.0%) of these midpoint-spanning alignments are of at least 95% identity. Thus, we estimate that at least 93% of coding bases are spanned by the assembled scaffolds and that ~99% of genes are at least partially represented in the assembly. Additionally, 390 EST contigs (11.5%) had exactly two midpoint-spanning alignments to the scaffolds of at least 95% identity (consistent with scaffold redundancy estimates below based on assembly self-alignment).

Assembly self-comparison

As an alternate appraisal of the redundancy of the assembled scaffolds, they were aligned to themselves using blastn (-e 1e-100 -F 'm D' -W 24). In scaffolds larger than 10kb, we found 9.1M bases (8%) to be aligned to a single HSP of at least 96% identity from another scaffold.

Comparison with finished fosmids

Fifteen finished fosmid sequences totalling 557.3 kb (GenBank IDs AC167695-AC167709) were compared to the trimmed reads and the assembled contigs and scaffolds to assess the

completeness of the libraries and the assembly. Passing reads were aligned to the fosmid sequences with BLASTN (-e 1e-25 -F 'm D' -W 21). The distribution of the number of reads aligned (over at least 95% of their trimmed length) per fosmid base is shown in Figure 2.5.1. The fosmid AC167706 is essentially uncovered, and was excluded from further analysis as a presumed contaminant. (AC167706 has a notably different GC content than the rest of the genome, does not hit any *A. queenslandica* ESTs, and has no clear candidate genes, but does have a hit to human c17orf27, an expressed RING finger-containing gene). Two other fosmids (AC167695 and AC167699) have sparser than expected whole genome shotgun coverage, but seem likely to be sponge sequences and have average GC content. These represent either biases of the shotgun libraries or regions in which the haplotypes sampled in the fosmid are different enough from the shotgun reads as to evade our conservative alignment criterion, as would occur with modest indel variation, for example. The distribution of coverage depths at a cutoff of 97% identity is shown in Figure 2.5.2, displaying a broad peak centered at ~8-9x (mean 9.3, consistent with a total genome size ~185 Mbp) and including ~4.2% unsampled regions at the specified alignment cutoffs (mostly contributed by fosmids AC167695 and AC167699). Figure 2.5.3 shows coverage of the fosmids by the assembly, demonstrating the high level of completeness of the assembly relative to shotgun coverage.

S2.6 Analysis of overrepresented 15-mers

The trimmed reads were decomposed into overlapping 15-mer sequences and the frequency distribution of the resultant 15-mers was examined to assess depth of sequence sampling. An ~1% contamination of the BAYB library and significant haplotypic polymorphism were detected in the resultant distribution and complicate any direct assessment of depth of coverage as seen in Figure S2.6.1. The 15-mer frequency distribution approximates a power law (ax^b ; b=-2.55; see Figure S2.6.2) for mers occuring more than 20 times in the data set, a finding consistent with moderate repetitive content[10] (J. Chapman, unpublished).

S2.7 Bacterial sequence analysis

Assembled contigs from the full sponge assembly were separated into a high GC and low GC fraction. The high GC fraction contained 2.7 Mbp of assembled sequence in 388 contigs. BLAST searches [22] against the NCBI nonredundant nucleotide and amino acid databases identified approximately 100 contigs with perfect identity to the γ-proteobacterium *Serratia proteomaculans*, also sequenced at the JGI. Primers designed to amplify these reads failed to amplify DNA on the source sponge embryo material. Hence, reads belonging to *Serratia proteomaculans* were considered as contamination and were discarded from further analysis.

Taxonomic assignments for the remaining unassembled reads were made using the MEGAN 2.0 program.[23] Reads were first mapped against the *Amphimedon queenslandica* draft assembly, with any mapped reads being removed from analysis. Remaining reads were then screened against the vector database UniVec and reads with homology to cloning vectors were removed. 217,873 reads remained after all filtration. Those reads were searched against the NCBI

nonredundant amino acid database with BLASTX 2.2.15, implemented in parallel with mpiBLAST.[24]

Due to computational constraints, taxonomic read assignments were performed on a randomly chosen representative subset of unassembled, filtered reads (120,000 out of 217,873 total). Of those, 7,720 (6.4%) were putatively assigned to the bacterial domain of life, and a very small number were assigned to archaea (161, 0.1%). No database hits were found for 38% of reads. Assignment to phyla and other major clades within Bacteria is shown in Figure S2.7.1, alongside the number of sequenced genomes for each clade. The majority of reads map to α- and γ-proteobacteria. We note that our method for read mapping is sensitive to the depth of taxonomic sampling for each group, such that a read might be erroneously recruited to a well-sequenced clade even if it truly belongs in a poorly sequenced clade for which a genome containing a homologous gene has yet to be sampled. For that reason, we plot the total number of finished and unfinished genomes for each group alongside the number of reads assigned, in order to qualitatively assess the relationship between assigned reads and depth of genomic coverage for each clade. In this way, we observe that the fraction of reads assigned to α-proteobacteria exceeds the fraction of all sequenced genomes that belong to α-proteobacteria. Likewise, we find an excess of reads assigned to Planctomycetes relative to the number of genomes sequenced, however the total number of putatively bacterial reads assigned to Planctomycetes is small (2.8%).

Thus, metagenomic analysis of sequence reads is consistent with existence of a dominant proteobacterial symbiont. Further phylogenetic analysis of marker genes is necessary to determine the exact branching point within the α-proteobacteria. We note, however, that MEGAN analysis assigns most bacterial reads to clades no deeper than the Class level (data not shown).

The presence of 10 randomly chosen proteobacterial sequences in DNA from independently-sourced larvae was assessed by PCR. Six of the 10 sequences were amplified, verifying that most of the putative prokaryotic sequences were sourced from organisms associated with *Amphimedon queenslandica* larvae (Figure S2.7.2). Transmission electron microscopy revealed bacteria present in *Amphimedon* larvae with similar ultrastructure to those found in other demosponges.[25]

S2.8 Evidence for CpG methylation

From a random sampling of 50,000 pairs (100k trimmed reads), the quantity [XpY]/[X][Y] was computed over all di-nucleotides. For random sequence, this quantity is unity. The mean value for CpG dinucleotides is significantly lower (0.36) relative to random sequence, and the products of deamination of mCpG to TpG (reverse complement CpA) are higher (1.27). The self-complementary TpA dinucleotide is also depleted (0.87). Other values of [XpY]/[X][Y] range from 0.97-1.09. See Table S2.8.1.

S3. Estimation of Polymorphism Levels

Using a suite of four polymorphic microsatellite loci, individual embryos and larvae housed within single brood chambers from multiple mothers were genotyped. This revealed that different fathers have fertilized eggs within a single brood chamber (data available upon request). In some sponges we have detected over 20 paternal alleles represented in the embryos in a single brood chamber, with the mother consistently contributing one of two alleles (unpublished data). Four predominant alleles per locus were detected in the DNA isolated for the genome project (Supplementary Section S2.1), suggesting 4 dominant haplotypes, although it is likely that this genomic DNA contains additional paternal haplotypes.

To assess SNP levels from shotgun data, all scaffolds of at least 100 kbp were realigned to repeat-masked reads from the BAYA and BAYB libraries using blastn with parameters: -W 16 -q -3 -U -F 'm D' -e 1e-50 . Best in genome alignments were used to call SNPs at positions with depth of at least four reads and at least two reads supporting two distinct alleles. The scaffolds were subdivided into windows of 100 bp and the number of SNPs per window were tallied for windows with mean depth of coverage between 10- and 15-fold and at least 4-fold coverage at every base in the window. This distibution of the number of SNPs per window is shown in Figure S3.1. The observed number n of SNPs per 100 bp is well-fit by a geometric distribution $A\ p(1-p)^n$ with A = 0.912 +/- 0.005 and p = 0.297 +/- 0.003 . The mean number of SNPs per 100 bp is then given by (1-p)/p = 2.37 +/- 0.03 (*i.e.*, one SNP every 42 bp on average). The normalization factor, A, is required due to an excess of zero-SNP windows. This small excess of zero-SNP windows (~ 8.8%) is proposed to represent the fraction of the assembly that can only be aligned to a single haplotype in the reads, either because of extreme divergence from, or complete absence in, the other haplotypes. This is consistent with the mean depth of coverage in windows with no observed SNPs (9.5X) being lower than that observed in windows with at least one SNP (12.1X). This is also partly due to a residual ascertainment bias (higher depth positions are more likely to reveal polymorphism).

Independent sequencing of cDNAs reverse transcribed from pooled mRNA from a range of individuals revealed extensive polymorphisms in many genes (unpublished data). To validate the gene models produced from genome assembly, we PCR amplified regions of a representative set of gene models, comprising both exons and introns, from individual embryos derived from multiple mothers. Nine gene models (*Axin*, Aqu1.225694; *Dishevelled*, Aqu1.226072; *Gsk3*, Aqu1.221634; *Tcf*, Aqu1.229819; *Par-6*, Aqu1.225622; *Igtir1*, Aqu1.221082; *Cadherin1*, Aqu1.212079; *EPRS*, Aqu1.222829; *Notch*, Aqu1.224719) were amplified from 8-12 individuals, cloned and sequenced. At least 10 sequences were obtained per gene per individual. Sequence data were consistent with the gene models at these loci, with two alleles per assembled locus (alignments available upon request, B Degnan). There was no evidence for the observed polymorphisms representing gene duplicates.

S4. Annotation of Protein Coding Genes

Models for protein-coding genes were generated using homology-based methods (Augustus[26], Genomescan[27]) and one *ab initio* method (SNAP[28]). Putative loci were defined for Genomescan

homology modeling using homology and transcription evidence. Regions with homology were found by blastx of the soft-masked scaffolds to the human, *Nematostella*, *Drosophila melanogaster*, and *Caenorhabditis elegans* proteomes using a significance cutoff of 1e-5.[22] Additionally, transcript assemblies from *Amphimedon queenslandica* were aligned to the genome using Blat.[29] Blat output was processed so that all the best hit to the genome, as well as any other hit within 97% coverage of the best hit, were considered matches. Likely pseudogeneous matches were filtered out by disallowing secondary matches with only 1 exon if the best hit has multiple exons. Putative loci were defined by these peptide and EST hits and joined if overlapping. Each region with flanking sequence was submitted with its best template to genomescan. A training set for the *ab initio* modeler SNAP was generated from transcript assemblies mapping to the genome with canonical splice sites which appear to code for complete transcripts.

The model sets from Genomescan, Augustus, and SNAP were then submitted to PASA[18] to find models consistent with *A. queenslandica* ESTs and cDNA. An additional model track was generated from the longest ORF from PASA-assembled ESTs and fl-cDNAs. All models were compared to Uniprot90 with blastp with an e-value cutoff of 1e-5. The best model per locus was then chosen by the following procedure: if there are PASA-validated models (completely consistent with EST evidence) the PASA validated models are chosen for this locus. If there are not PASA-validated models, the best hit to Uniprot90 is chosen. The best hit is chosen via a reciprocal coverage criteria where the coverage score = (2*residues aligned)/(lengthq + lengths), and the score closest to 1 is best. If there is no hit to Uniprot90, the model is kept only if there is EST evidence. This procedure generated 30,327 models at 29,867 loci, resulting in an average of 178 genes per megabase of assembled sequence.

Summary statistics for the gene models are reported in Table S4.1. 24,743 (83%) of the gene models are supported by BLAST against ESTs, or PFAM domains, or human proteins in the RefSeq database, or other proteins in the SwissProt database (Table S4.2) (these BLASTs were done using standard blastx parameters with an e-value cutoff of 1e-5).

S5. Intron Splice Site Conservation

Based on mutual-best BLAST alignments of protein sequences to human genes, a collection of 1,196 sets of orthologous genes was constructed (as described previously[16]) with representation from human, *Lottia gigantea*, *Nematostella vectensis*, *Trichoplax adhaerens*, *Amphimedon queenslandica*, *Monosiga brevicollis*, *Arabidopsis thaliana*, and *Cryptococcus neoformans*. MUSCLE alignments[30] were run for each of the clusters and intron positions were marked in the alignments.

A splice site was regarded as “highly-reliable” if the 8 flanking alignment positions (in both directions) had at least 3 amino acids with the same biochemical properties (+ or * in Clustal), and no gaps. Additionally, we required that no splice site exist in any species within 4 amino acids from a highly reliable splice site. This last criterion removes seemingly different splice sites that may be due to ambiguous or cryptic splice sites artificially introduced by gene modeling. In total 1,993 highly-reliable conserved intron sites were identified. Out of these, 472 could be traced back to the eukaryotic ancestor; *Monosiga* retained 220, *Amphimedon* retained

348, *Trichoplax* retained 402, *Nematostella* retained 398, and the human genome retained 334. Out of the 1,993 highly-reliable intron positions considered, 928 could be traced back to the common ancestor of metazoans; out of these, *Amphimedon* retained 785, *Trichoplax* retained 826*, Nematostella* retained 815, and the human genome retained 708. Similarly, 1,063 were inferred to be present in the ancestor of eumetazoans; out of these *Trichoplax* retained 967*, Nematostella* retained 940, and the human genome retained 790.

Median intron sizes in *Amphimedon* are reduced relative to *Monosiga* and other metazoans: *Monosiga* has a median intron size of 117 bp, *Amphimedon* has 81 bp, *Trichoplax*, *Nematostella* and human have 105 bp, 377 bp, and 1045 bp, respectively. A similar trend is observed for the intergenic regions: *Monosiga* has a median size of intergenic regions of 832 bp, *Amphimedon* has 824 bp, *Trichoplax*, *Nematostella*, and human have 2,809 bp, 4,109 bp, and 22,345 bp, respectively. These data might indicate significant secondary loss of DNA in certain genomic regions in *Amphimedon*. Fisher's exact analysis shows that introns in MF00258 CAM family adhesion molecule proteins show the highest difference between *Amphimedon* and other metazoans and *Monosiga* (median for *Amphimedon* is 54 bp, whereas for *Monosiga* 112 bp).

S6. Conserved synteny in *Amphimedon*

As in Putnam *et al.* 2007[16], and Srivastava *et al.* 2008[17], orthologous genes were identified between the *Amphimedon* and *Nematostella* genomes. We used Fisher's Exact Test to test the significance of the apparent concentration of orthologous genes between regions of the *Amphimedon* and *Nematostella* genomes, and between *Amphimedon* and the ancestral linkage groups (Figure S6.1). In the dot-plot in Figure S6.1, segments of the *Nematostella* genome (*i.e.* assembled scaffolds) are organized into groups that represent the gene content of a putative chromosome in the cnidarian-bilaterian ancestor.

S7. Phylogenetic Analyses

Relationships of phyla at the base of the animal tree have remained contentious, and one of the benefits of whole-genome sequence data from these phyla is that they would lend more certainty to where these phyla are placed on the tree. In the methods discussed below, we present the most likely relationships using currently available genome sequence from a representative sponge (*Amphimedon*), a placozoan (*Trichoplax*) and two cnidarians (*Nematostella* and *Hydra*). This study is limited to organisms represented by whole-genome sequence to avoid the missing data problem associated with EST studies. We note that our aim is not an exhaustive molecular phylogenetic study of eukaryotes, but rather an initial exploration of the use of such complete gene sets. Since complete genomes are available for only one sponge, only one placozoan, and no representatives of ctenophores, these must be viewed as only provisional tests of phylogenetic hypotheses. It is clear that multiple representatives of the various basal metazoan phyla will be needed to refine these tests.

A major debate in the field focuses on the relationships of different sponge lineages to each other and to other animals. Specifically, the monophyly of all sponges - *i.e.,* the idea that all groups of sponges share a single common ancestor that is not shared by other animals - has been

challenged.[31-34] The homoscleromorph sponge, *Oscarella carmella,* may belong to a lineage that diverged from eumetazoans after the separation of other sponge lineages.[33] Another debate in early animal evolution surrounds the phylum Ctenophora, the comb jellies. We used EST data available for *Oscarella* [35] and *Mnemiopsis leidyi*, a representative ctenophore to evaluate the positions of homoscleromorphs and ctenophores.

S7.1 Generation of datasets of orthologous genes

Phylogenetic analyses to establish the position of *Amphimedon* and *Trichoplax* in the animal tree were conducted on datasets generated using two methods -- the filtered mutual best-hit (fMBH) method and the four taxon kernel (FTK) method. The use of complete genomes minimizes difficulties in determining orthology and incompleteness found in transcriptome datasets.

In the filtered mutual best-hit method, lists of mutual best-hit genes from seventeen proteomes (human, *Strongylocentrotus*, *Branchiostoma*, *Lottia*, *Capitella*, *Helobdella*, *D. melanogaster*, *C. elegans*, *Pristionchus*, *Nematostella*, *Hydra*, *Trichoplax*, *Amphimedon*, *Neurospora*, *Arabidopsis*, *Paramecium*, and *Dictyostelium*) to genes in the *Monosiga* genome were generated. Only hits with e-value lower than 0.001 were retained. To avoid the use of confounding paralogs, the MBH lists were filtered such that an MBH-pair was retained only if the second best hit of either gene in the other genome had a score smaller than half the score of the best hit. Single-linkage-clustering using the *Monosiga* genes was used to generate clusters of orthologous genes. Since any gene in the *Monosiga* genome had only one MBH in each of the other genomes, the resulting clusters contained no more than one gene from each of the eighteen genomes in the analysis.

In the four taxon kernel method, a two-step approach was taken. In the first step, genes that yielded trees that mirror well-established topologies of a subset of species were identified. In the second, orthologs from other species were pledged to these gene clusters.

For the first step, lists of mutual best hit genes were generated for four genomes against each other; these genomes were *Lottia*, *Branchiostoma*, *Nematostella* and *Monosiga*. Only hits with e-value equal to or less than 0.001 were retained. Single-linkage-clustering of these MBH lists was used to generate sets of potentially orthologous genes from these proteomes. Only clusters with one gene from each of the four proteomes were retained for further analysis. Each gene set was aligned using ClustalW[36,37] and trimmed using GBlocks[38] (with default settings b3=8, b4=10). Neighbor-joining (using Phylip)[39] and maximum likelihood (using PhyML)[40] trees were generated for a subset of these gene sets. Out of 12 test cases, the trees from NJ and ML analyses were identical for 11 sets. Thus, neighbor-joining trees were generated for all the gene sets. These trees were rooted using the *Monosiga* gene (choanoflagellates are an outgroup to animals and, in the case of single-copy orthologous genes, choanoflagellate genes can be considered as an outgroup for animal genes). The topologies of these trees were evaluated and only trees that showed the expected ((*Lottia*, *Branchiostoma*), *Nematostella*) relationship of ingroup genes were retained for further analysis. The justification for filtering the genes in this manner is that, by limiting the phylogenetic analyses to genes that yield a topology congruent with previously known relationships, we can avoid the use of genes with altered rates of evolution that don't yield even well-established relationships. However, it is possible that this selection would bias the

gene set toward genes that support bilaterian monophyly by virtue of having evolved rapidly along the bilaterian stem. Of the 2,149 gene sets that had one gene each from *Lottia*, *Branchiostoma*, *Nematostella* and *Monosiga*, only 1,215 showed the expected topology. The 1,215 kernels had one gene for each of the four species.

In the second step of the FTK method, lists of MBH-pairs from fourteen proteomes (human, *Strongylocentrotus, Capitella, Helobdella, Drosophila melanogaster, Caenorhabditis elegans, Pristionchus, Hydra, Trichoplax, Amphimedon, Neurospora, Arabidopsis, Paramecium* and *Dictyostelium*) were used to pledge more genes into the 1,215 kernels. For each of the fourteen additional proteomes, a gene was placed into a kernel only if it was a mutual best hit to each of the four genes originally in the cluster. No more than one gene from a proteome could be pledged into a cluster, as by definition it was the best hit in that genome for the kernel genes. Similar to the fMBH-method, this method also yielded clusters of orthologous genes that contained no more than one gene from each of the eighteen genomes in the analysis.

In both approaches, the orthologous clusters could be used to generate data sets with different amounts of missing data. For example, requiring that gene sets had sequences from all eighteen proteomes resulted in a small number of genes, but allowing gene sets to miss sequences from up to one, two, or three species yielded more gene sets. Six different datasets were generated (Table S7.1). Each set of orthologous genes was aligned using ClustalW and trimmed to retain high-confidence homologous positions using GBlocks (with default settings b3=8, b4=10). Gene sets belonging to any given dataset were then concatenated for total-evidence phylogenetic analyses as described below.

S7.2 Likelihood analyses and hypothesis testing

The results of maximum likelihood inference of phylogeny on all the datasets described above are summarized in Table S7.2. Three datasets for each orthologous gene set method (fMBH and FTK) were generated by allowing different amounts of missing data. The alignments were analyzed in PhyML using the Whelan and Goldman (WAG) model of amino acid evolution.[41] The proportion of invariable sites was estimated using the data and four substitution rate categories were allowed with the γ distribution parameter estimated from the dataset. All six datasets (small, medium and large of the fMBH and FTK methods) were also analyzed by removing the two nematode (*C. elegans* and *Pristionchus*) branches.

To determine if the topology with the maximum likelihood in any of these analyses was significantly better than alternative topologies, a site-wise log-likelihood score matrix was generated for testing competing topologies using TREEPUZZLE 5.0.[42] The alternative topologies cover a range of hypotheses including the placement of placozoans as the earliest animal branch or as sister to cnidarians (for details see Tables S7.6 and S7.7). The WAG substitution model was used with a mixed model of rate heterogeneity (four rate categories with the proportion of invariant sites and the γ distribution parameter α estimated from the data). The p-values for the one-sided Kishino Hasegawa (KH) test[43] and the Estimated Likelihood Weight (ELW) test[44] were determined using TREEPUZZLE 5.0.[42] The p-values for the Shimodaira Hasegawa (SH) test, [45] the weighted KH test, the weighted SH test and the Approximately

Unbiased (AU) test[46] were determined using CONSEL.[46] These results are summarized in Tables S7.8 – S7.11. Given the inaccuracies of the KH test[47] and the conservative nature of the SH tests,[44] we rely on the weighted KH test, the ELW test and the AU test to evaluate the alternative topologies. A p-value below 0.05 allows rejection of the null hypothesis that the topology under consideration is derived from the same distribution as the topology with the highest likelihood.

The smallest dataset from the fMBH method recovers nematodes (*C. elegans* and *Pristionchus*) as a sister group to all other animals (Table S7.2, Figure S7.4). This is a well-known long-branch attraction phenomenon.[48-50] In this tree, *Trichoplax* appears to be a sister group to cnidarians (*Nematostella* and *Hydra*), but removing nematodes from the alignment places placozoans as a eumetazoan branch, sister to cnidarians and bilaterians. Neither dataset is powerful enough to reject alternative topologies, which is also the case for the smallest dataset in the FTK method (though nematodes appear correctly as ecdysozoans in this method). In the latter case, *Trichoplax* appears as the earliest animal branch, sister to all other animal groups (Table S7.2, Figure S7.1).

Allowing up to one species to be missing from each gene set (in the FTK method) and allowing up to two species to be missing from each gene set (in the fMBH method) gives mid-sized datasets of about 22,000 amino acid positions each (Table S7.1). Long branch attraction of nematodes persists in the fMBH dataset, but upon removal of nematodes, *Trichoplax* is recovered as the earliest eumetazoan branch and *Amphimedon* as the earliest metazoan branch with high bootstrap support, and alternative topologies are rejected (Table S7.2, Figure S7.5). The same relationships are recapitulated in the FTK dataset, regardless of the inclusion of nematodes in the matrix (Table S7.2, Figure S7.2).

The largest datasets in both the fMBH and FTK methods produce the topology shown in Figure 1f, with *Amphimedon* as the earliest animal branch and then *Trichoplax* emerging before the divergence of cnidarians and bilaterians (Figure S7.3, S7.6). In the fMBH method, the long-branch attraction of nematodes is corrected with the addition of more data to the medium-sized matrix. The FTK dataset is discriminating, and is able to reject all other topologies, as is the fMBH dataset without nematodes (the FTK dataset including nematodes is unable to reject the alternative hypothesis of placozoans as sister to cnidarians, possibly due to confounding signals from the nematode sequences) (Table S7.2).

One hundred bootstrap replicates were generated for each of the twelve datasets analyzed in the likelihood framework using PhyML. The six datasets that contain nematodes disagree regarding the placement of *Trichoplax* relative to sponges, cnidarians and bilaterians. However, there is very poor bootstrap support (<50) for the various positions (*Trichoplax* sister to cnidarians, *Trichoplax* as the earliest animal branch, *Trichoplax* as sister to cnidarians and bilaterians). Effectively, these analyses are indifferent to the position of *Trichoplax* and the relationship of *Amphimedon*, *Trichoplax* and cnidarians should be shown as a polytomy. The only position with high bootstrap support (71) is *Trichoplax* as sister to cnidarians and bilaterians in the large fMBH dataset. Five of the six datasets without nematodes, however, converge and place *Amphimedon* as the earliest animal branch and *Trichoplax* as sister to cnidarians and bilaterians with high bootstrap support (Figure 1f, Table S7.2, Table S7.3, Figures S7.1-S7.6).

The increasing support for the topology in Figure 1f with increasing amounts of data in the matrix, as well as with the removal of nematodes, suggests that the alternative placements for *Trichoplax* may be artifacts of a small amount of data and the confounding effect of long branches (nematodes). The large fMBH and large FTK datasets were analyzed further with alternative inference methods as described below.

S7.3 Bayesian analyses

Because of the known over-estimation of posterior clade probabilities in Bayesian methods[51], likelihood analyses were used to evaluate the different datasets described above. To confirm that the maximum likelihood topology obtained from the largest fMBH and FTK datasets is also supported by other methods, Bayesian inference was used as an alternative method to obtain support values. The 229 gene FTK and 242 gene fMBH datasets were analyzed using MrBayes v3.1.2[52,53] using a mixed amino acid model prior and a variable rate prior. Two chains were run for both datasets (one tree was sampled per 100 generations; 10,000 trees were discarded as burnin) and after 470,000 MCMC generations, the runs had converged. Both datasets yielded the same tree, with the same topology as in Figure 1f, with *Amphimedon* as the earliest branching animal lineage and *Trichoplax* as sister to cnidarians and bilaterians (all branches on these trees have a posterior probability of 1) (see also Figure S7.7). All nodes have posterior probabilities of 100. The 100% credibility tree set has only one tree for the FTK dataset, and the fMBH dataset explored a second tree switching the positions of *Dictyostelium* and *Arabidopsis* with a probability of 0.001.

S7.4 Use of site-heterogeneous models

The Bayesian and likelihood analyses described above use empirically-derived models of amino acid evolution (specifically, the Whelan and Goldman WAG model, which is known to work for nuclear protein sequences).[41] These models do not model heterogeneity in rates across sites or time, but rather hold equilibrium frequencies of amino acids constant. Recently, methods have become available that allow for more complex models of amino acid evolution. Aamodel, a Bayesian inference program, allows for a model similar to the GTR+γ model often used for nucleotide sequence alignments.[54] CAT, a model that allows for different substitution processes across sites was proposed by Lartillot and Philippe 2004[55] and can be used in the Bayesian inference program PhyloBayes.[55,56] We used both of these methods to infer animal phylogeny using the 229 gene FTK dataset and the 242 gene fMBH dataset.

Aamodel was run using the following parameters: -a=poisson, tempering parameter -r=190.0, γ rate categories -g=4, for 1,000,000 generations. (-a=poisson simply sets the parameters of the Dirichlet prior distribution on the 190 exchangeability parameters of the GTR model; the rate parameters are allowed to vary). Both the large FTK and the large fMBH datasets give the same 50% majority rule consensus tree, which has the same topology as the tree in Figure 1f (Figures S7.8, S7.9). The fMBH dataset has 100% posterior probability for all nodes, but the FTK dataset supports *Trichoplax*+cnidarians+bilaterians to the exclusion of *Amphimedon* with a posterior probability of 95% (Table S7.3). This is consistent with the likelihood analyses, in which the

fMBH dataset has higher bootstrap support for this clade and is able to reject all alternative topologies, whereas the FTK dataset is unable to reject some alternative positions of *Trichoplax* (consistent with the appearance of three other trees in the 100% credibility tree set in the Aamodel analysis). Thus, it appears that the additional computational burden of modeling complex amino acid evolution did not give results very different from the methods using the empirically-derived WAG model.

PhyloBayes 2.3 was used to infer phylogeny using default settings (CAT-Poisson - the equilibrium frequency profiles of substitution processes modeled as simple Poisson processes; γ distributed rates with four categories) (Table S7.12). Two chains were run for each dataset for >1,000,000 generations, with one tree sampled every 100 generations (the first 100 trees were discarded as burnin). For both datasets, the two runs showed good congruence/mixing (maxdiff < 0.1) and were used to generate a consensus tree. The 50% majority rule consensus tree from the 229 gene FTK dataset separates early animal lineages (*Amphimedon*, *Trichoplax* and cnidarians) into a monophyletic group that is sister to bilaterians (Figure S7.11). However, the posterior probability for this clade is 0.68, and the node should be collapsed given that in Bayesian analyses only clades with posterior probability of 0.95 or higher are considered well-supported. This would leave the earliest animal node as a polytomy with three branches - cnidarians, bilaterians and a clade containing *Amphimedon* and *Trichoplax*. This polytomy is consistent with the better-resolved tree from the 242 gene fMBH dataset, in which cnidarians and bilaterians are supported as sister groups with a posterior probability of 1 (the partition of cnidarian and bilaterian taxa to the exclusion of other species does appear in the FTK dataset analysis, albeit with a posterior probability of 0.21) (Figure S7.10). The fMBH-based tree also recovers *Amphimedon*+*Trichoplax* as a clade with high credibility (0.98).

Though PhyloBayes 2.3 has a better model, CAT+GTR, which allows for a mixture of general time reversible (GTR) substitution processes with different equilibrium frequency profiles for nucleotide sequences, it does not implement such a model for amino acids. This method would be comparable to the one implemented in aamodel, as described above. It is difficult to reconcile the different results from these two methods that account for rate heterogeneity in amino acid sequence evolution. The 100% credibility tree set in the aamodel analysis of the FTK data does contain the FTK PhyloBayes tree (but with posterior probability 0.04) and the fMBH PhyloBayes tree (but with posterior probability 0.009). One can only speculate about which method (CAT+Poisson or GTR) models the real underlying substitution processes across the large set of eukaryotic genes used in these analyses. Lartillot and Philippe report that the GTR model tends to fit the data better for large amino acid datasets (see PhyloBayes 3.2 instruction manual). There is great debate among researchers on the benefits of modeling rate heterogeneity[55,56] and some warn against using models with many parameters.[57] Further studies are needed to evaluate these new methods. For the purposes of this study, we therefore propose the position of *Trichoplax* as a sister to cnidarians and bilaterians, with *Amphimedon* as the earliest branching animal lineage as the most likely topology (since it is well-supported by likelihood and Bayesian inference methods, in the latter both with or without complex models for amino acid evolution).

S7.5 Evaluating the positions of homoscleromorphs and ctenophores

To address the question of the monophyly of sponges[31-34], we used EST data available for the homoscleromorph sponge, *Oscarella carmella*, which allowed us to generate 5,117 open reading frames.[35] To address the position of ctenophores in the animal tree, we used EST data available for *Mnemiopsis leidyi*, which allowed us to generate 2,278 open reading frames.

Oscarella genes could be pledged into 170 of the 1,215 clusters from the FTK method, as they were mutual best hits of all four kernel genes (one each from *Lottia*, *Branchiostoma*, *Nematostella* and *Monosiga*) in those clusters. Of the 170, 64 were in the 229 clusters (missing sequences from no more than two species) of the large FTK dataset that resulted in the tree in Figure 1f (Table S7.1). Similarly, *Mnemiopsis* genes could be pledged into 124 of the 1,215 clusters, and 46 of these 124 clusters were a subset of the 229 clusters of the large FTK dataset (Table S7.1).

In the fMBH method, 3,412 sets of genes were obtained by single linkage clustering of lists of filtered MBH genes as described above. Of these, 426 could be assigned an *Oscarella* gene and 404 could be assigned a *Mnemiopsis* gene (based on filtered MBH lists of *Oscarella* and *Mnemiopsis* to *Monosiga* proteins). Of the 426 *Oscarella*-containing clusters, 53 were a subset of 242 clusters selected to generate the largest matrix for the fMBH method; similarly, 48 of the 404 clusters that received *Mnemiopsis* genes were a subset of the 242 selected genes (Table S7.1).

The 229 (from the FTK method) and 242 (from the fMBH method) clusters, now with *Oscarella* and *Mnemiopsis* sequences added where possible, were used to infer the maximum likelihood tree (Table S7.1). The FTK dataset reproduces the same relationships of animals as it did previously in the absence of *Oscarella* and *Mnemiopsis*, with *Oscarella* now placed as a sister lineage to *Amphimedon* (with a very poor bootstrap support of 31) and *Mnemiopsis* as the earliest diverging animal lineage (sponges, placozoans and other animals form a clade to the exclusion of *Mnemiopsis* in 80 of 100 replicates) (Figure S7.12). The fMBH dataset is sensitive to the addition of *Oscarella* and *Mnemiopsis* sequences, in that the relationships of other animals are altered (*Oscarella*+*Trichoplax*, both in turn being sister to cnidarians; *Mnemiopsis* appears as sister to *Amphimedon*, both in turn being sister to all other animals), and the poor bootstrap support leaves all early animal relationships unresolved (Figure S7.13).

When known long branches (*Caenorhabditis*, *Pristionchus*, and *Paramecium*) were removed from these analyses, the support for *Mnemiopsis* as the earliest animal branch increased to 93, and support for *Oscarella* as sister to *Amphimedon* increased to 46 in the FTK analysis (Figure S7.12). In the fMBH analysis, *Mnemiopsis* was recovered as sister to all other animals in 69 out of 100 replicates, but the position of *Oscarella* remained poorly resolved (Figure S7.13).

The loss of resolution of relationships upon the addition of *Mnemiopsis* and *Oscarella* sequences is potentially due to the large amount of missing data from these species. Thus, we evaluated alternative positions for *Oscarella* and *Mnemiopsis* using the data that are available from these EST projects, and fixed relationships between other animals based on the tree in Figure 1f.

For testing alternative hypotheses for the placement of *Oscarella*, the following datasets were considered:

1. 229 genes of the FTK method that were used to establish the topology in Figure 1f, with *Oscarella* genes assigned where possible (to 64 gene sets), and *Mnemiopsis* sequences removed from the concatenated alignment.
2. 242 genes of the fMBH method that were used to establish the topology in Figure 1f, with *Oscarella* genes assigned where possible (to 53 gene sets), and *Mnemiopsis* sequences removed from the concatenated alignment.
3. Dataset in 1) without nematodes
4. Dataset in 2) without nematodes

For testing alternative hypotheses for the placement of *Mnemiopsis*, the following datasets were considered:

1. 229 genes of the FTK method that were used to establish the topology in Figure 1f, with *Mnemiopsis* genes assigned where possible (to 46 gene sets), and *Oscarella* sequences removed from the concatenated alignment.
2. 242 genes of the fMBH method that were used to establish the topology in Figure 1f, with *Mnemiopsis* genes assigned where possible (to 48 gene sets), and *Oscarella* sequences removed from the concatenated alignment.
3. Dataset in 1) without nematodes
4. Dataset in 2) without nematodes

Most alternate topologies for relationships among all other animals could be rejected for both the 229 FTK and the 242 fMBH genes (Table S7.2). Alternative placements for *Oscarella* and *Mnemiopsis* were thus evaluated by adding data from these two taxa to the gene sets where possible, and constraining the relationships of all other animals. (Tables S7.13, S7.14). The genes in each of these new datasets were concatenated to yield a single matrix, and the sequence from *Paramecium* was removed to avoid confounding signals from this long branch. The results of these tests, done using TreePuzzle and CONSEL, are summarized in Table S7.4 (Tables S7.15, S7.16). Of the topologies tested, the 229 genes of the FTK method gave the most likely position of *Oscarella* as sister to *Amphimedon*; however, this placement is not significantly more likely than topologies with alternative positions for *Oscarella* - only the sister-relationship of *Oscarella* with cnidarians and of *Oscarella* with bilaterians, and the placement of *Oscarella* as the branch after *Trichoplax* but before cnidarians can be rejected. The FTK method does not allow us to discriminate between two alternate positions of *Mnemiopsis* (earliest branching metazoan lineage, or sister to Amphimedon). The fMBH method datasets are less discriminating, possibly because of less stringent assigning of *Oscarella* and *Mnemiopsis* genes to orthologous gene clusters.

The removal of long branches from these analyses did not change the evaluation of different topologies by the FTK analysis, although a few more alternate placements for *Oscarella* and *Mnemiopsis* could be rejected in the fMBH analysis (but these were also rejected by the FTK analysis) (Table S7.4, Tables S7.15, S7.16). This may suggest that the fMBH analysis is more prone to artifacts from long-branch effects.

S7.6 Summary of phylogenetic analyses

Phylogenetic analyses on concatenated alignments of single-copy nuclear genes from eighteen genomes (comprising thirteen animal species and five other eukaryotes) place demosponges (as represented by *Amphimedon*) as the earliest branching animal lineage, with placozoans as a sister group to cnidarians and bilaterians, consistent with previous studies[17] (Figure 1f). This topology is the most likely, all its nodes have a posterior probability of 1 in Bayesian analyses, and it gains greater support with increasing numbers of characters in the amino acid matrix. The largest datasets from two different methods allow us to formally reject the possibility that the placozoan lineage diverged prior to the divergence of sponges, cnidarians and bilaterians (Table S7.2, Tables S7.6, S7.8), a position that is supported by other studies using concatenated morphological, nuclear and mitochondrial nucleic acid and amino acid data.[58]

The addition of orthologous EST sequences from the homoscleromorph sponge, *Oscarella carmella*, and the comb jelly, *Mnemiopsis leidyi*, to the eighteen-taxon matrix suggests that ctenophores may be the earliest branching animal lineage and that *Oscarella* may be sister to *Amphimedon*. Although these positions are recovered as the most likely, they are not significantly better than other potential placements for *Oscarella* and *Mnemiopsis* (Table S7.4). The position that we find to be most likely for *Mnemiopsis* has previously been proposed by others,[59] but in our analyses ctenophores are almost as likely to be sister to sponges. The recovery of sponges as a monophyletic group adds to the debate generated by recent analyses that place homoscleromorphs as a lineage separate from other sponges, and sister to eumetazoans.[32,33]

S7.7 Remaining issues in deep animal phylogeny

The placement, in our analyses, of *Trichoplax* as sister to cnidarians and bilaterians, with sponges as the earliest animal branch, disagrees with results from mitochondrial trees[60-62] However, mitochondrial analyses are complicated by the long branch lengths (*i.e.*, unusually high levels of amino acid divergence) found in bilaterian mitochondrial peptides relative to their basal metazoan orthologs.[62,63] Figure 1f shows that peptides encoded by the nuclear genome show no notable differences in amino acid substitution levels between basal metazoans and bilaterians, suggesting that our proposed phylogeny based on nuclear genes is less susceptible to long-branch attraction artifacts.

Recently, another study using several EST and whole-genome datasets to attempt to resolve animal relationships also placed placozoans as sister to cnidarians and bilaterians, albeit with weak support.[64] However, a different study that takes a “total evidence” approach - combining nuclear and mitochondrial protein coding genes, ribosomal RNA genes, and morphological characters - disagrees with our results and with the Philippe *et al* tree.[58] The topology recovered in this "total evidence" study places all early branching animal lineages in a monophyletic group that is sister to bilaterians; *i.e.,* the metazoan ancestor, in this scenario, gave rise to bilaterians in one lineage, and to cnidarians, ctenophores, placozoans and sponges in the other. Within the “early animal” clade, placozoans diverge first, then sponges, and ctenophores and cnidarians are sister taxa. This study attempts to address the possible swamping of the tree by mitochondrial

data, and though the authors argue that a minority of nodes in the tree are supported only by mitochondrial data, they fail to point out that it is the key nodes in their topology that are the ones indeed supported by mitochondrial data. Also, most of the methods and weighting schemes in this study give very poor support to the critical node where early animals diverge from bilaterians. Given this lack of robust support, and the ambiguous coding of morphological characters, it is difficult to evaluate the validity of this study.[65] The datasets used in our analyses consistently reject topologies with an "early animal" clade.

Though the putative position of *Mnemiopsis* as the earliest branching taxon in our analyses appears in agreement with the Dunn *et al.* (2008) dataset, more data from this species and other ctenophores will be critical in establishing just how robust is this position. It is possible that *Mnemiopsis* falls out as the earliest branch in our and others' analyses because of long-branch attraction, and that better taxon sampling is needed to determine the correct placement of ctenophores. The Philippe *et al.* (2009) study that uses the CAT model in fact recovers the more traditional "Coelenterata" clade, with ctenophores and cnidarians as sisters. These different scenarios offer very different interpretations of events in animal evolution. If comb jellies are the earliest branching animal phylum, then the features they share with bilaterians (such as neurons and mesoderm) were characteristics of the metazoan ancestor that have been lost in sponges (possibly more than once), placozoans and cnidarians. If ctenophores form a monophyletic group with cnidarians, then absence of nerves and muscles in sponges and placozoans is primary.

The inability of datasets in this study to reject alternative placements for *Oscarella* suggests that, at this time, we do not have enough data to resolve its relationships to other animals. As with ctenophores, broader taxon sampling within sponges for large molecular datasets will be crucial in resolving the issue of sponge paraphyly. If different sponge lineages branched off at different times from the lineage that gave rise to eumetazoans, then it is likely that the metazoan ancestor bore resemblance to modern sponges (given that it is unlikely that sponge morphology and lifestyle evolved convergently more than once). If sponges are monophyletic, then the morphology of the metazoan ancestor is more difficult to ascertain. The recent dataset from Philippe *et al.* (2009) contains the largest sampling of sponges yet (all sponge groups are represented with large amounts of newly-generated EST data) and supports that idea that all modern sponges descended from one common ancestor that is not shared by other animals.

<u>Limitations of these methods with regards to eukaryotic group relationships:</u> Since our main focus was on early metazoan relationships, we only used a small selection of diverse non-holozoan outgroups, intended to polarize the root of Holozoa. To address the deep relationships among these (and other) diverse eukaryotic groups would require far more complete taxon sampling than attempted here. (In contrast, for deep metazoan relationships we included all available genomes from "basal" metazoan phyla.) The tree shown in Figure 1 places *Arabidopsis* as a sister clade to opisthokonts, which is contradictory to the results of recent phylogenomic analyses aimed at resolving deep relationships among eukaryotic lineages. Some have suggested that Alveolates (a group containing ciliates, *e.g., Paramecium*) and plants are sister-groups.[66,67] However, other phylogenomic studies[68] though studies with broader taxon sampling[69] do not recover high support for this new hypothesis. Furthermore, the placement of the root for the eukaryotic tree is highly debated,[70] rendering sister-group relationships between the major eukaryotic groups uncertain. However, these studies recover amoebozoans as a sister

clade to opisthokonts, and given that the eukaryotic tree is unlikely to be rooted between amoebozoans and opisthokonts, the phylogeny we propose here does not recover expected relationships (*i.e.*, we do no recover support for *Dictyostelium*, the single amoebozoan in our analyses, as sister to opisthokonts). Further analysis with more complete taxonomic coverage is clearly required to address deep eukaryotic relationships.

S7.8 Estimating divergence in time and percent change

Based on the tree in Figure 1f, we find that 28% (0.0717/0.2559 substitutions per considered site) of the amino acid substitutions in the human lineage since the holozoan ancestor occurred on the metazoan stem.

Based on reports placing the earliest unequivocal bilaterian fossils at 555 mya,[71] we fixed the bilaterian radiation in our trees at 555 mya, which is in the estimated range for this radiation using other fossil dates for calibration.[72] Four trees generated from four different datasets as described above (Table S7.8.1) were used to estimate divergence times using the r8s software.[73] The estimates for the last common holozoan ancestor range from 923-990 mya, comparable to the estimates from fossil calibrations in other studies.[74]

S8. Analysis of the Gene Complement of Sponge

S8.1 Identification of Amphimedon genes

Putative orthologs of genes involved in various processes in bilaterians were identified by reciprocal BLAST of human, mouse, or *Drosophila* genes against the *Amphimedon* gene models (blastp) or the assembly (tblastn) (in the latter case, gene models predicted at the best-hitting loci were tested for orthology). PFAM[75] domain composition, assignment of PANTHER HMMs[76,77] and phylogenetic trees were used to determine orthology. Trees were built using the neighbor method in Phylip[39] with the distance matrix generated using protdist and one hundred bootstrap replicates (unless noted otherwise). Appropriate outgroup sequences were used when available. By studying the distribution of orthologs across species representing various eukaryotic clades, all gene families were annotated based on which stem in the eukaryotic tree they most likely first appeared. For example, a family such as the Wnt family, which has no recognizable homologs outside of animals, can be annotated as a novel metazoan gene, which most likely originated on the metazoan stem.

S8.2 Cell cycle and growth

Cell cycle regulators

The hallmarks of multicellular life are social controls on cell division, growth and death to achieve balance; coordinated cell-cell and cell-matrix adhesion to produce organismal

morphology; specification of differentiated cell types to achieve division of labour and processes for distinguishing self from non-self to maintain individuality. While most of these features are also found outside of animals in other multicellular clades (e.g., plants, fungi, volvocale alga, etc.), the corresponding molecular functions are typically executed by analogous rather than homologous genes and proteins.

Orthologs of genes involved in the mammalian cell cycle were identified in *Nematostella*, *Trichoplax*, *Amphimedon*, *Monosiga*, and other outgroups (*Paramecium*, *Dictyostelium*, *Saccharomyces*, *Arabidopsis*) using a variety of methods (Table S8.2.1). The bilaterian members of these families were analyzed for domain composition and orthologs were identified based on essential domain compositions. In some cases, clear orthologs could be identified using reciprocal best BLAST methods. In the case of large families, such as cyclins and cyclin-dependent kinases (CDKs), phylogenetic trees were made using the neighbor joining method in Phylip with one hundred replicates (Figures S8.2.1-S8.2.6).

Cell cycle control is an ancient process that allows organisms (single- and multi-celled) to respond to stress (e.g., lack of nutrients, DNA damage) and many of the molecules for cell cycle progression are conserved among eukaryotes.[78-80] However, the cell cycle of extant vertebrates has been the result of novel proteins and duplicates of ancient eukaryotic cell cycle genes that appeared at different times over the course of evolution (Figure 3a). For example, Cyclin E proteins appear to be unique to animals (there is a clear ortholog in *Amphimedon*); though a divergent potential homolog is present in *Monosiga*, it does not fall into a monophyletic group with animal Cyclin E in phylogenetic trees (Figure S8.2.1). Most subfamilies of cyclin dependent kinases are ancient, some are unique to choanoflagellates and animals (Cdk10, CCRK, PCTAIRE), some appear unique to animals (Cdk2 and PFTAIRE) and one is present only in eumetazoans (Cdk4/6) (Figure S8.2.2). The cell cycle transcriptional regulator Rb is an ancient protein found in animals as well as plants, but the protooncogene Myc and the tumor suppressor p53 have recently been reported as evolving more recently in holozoans.[81] The Myc (a bHLH-ZIP transcription factor) homolog in the unicellular *Monosiga* only retains homology in the bHLH and zipper regions and lacks certain N-terminal amino acids that are highly conserved in animal Myc proteins, including the 'DCMW' motif, mutations in which abrogate Myc function in vertebrates (Figure S8.2.3).[82]

The E2F/DP group of transcription factors are ancient, with single genes from all eukaryotic species represented in the DP subfamily (Figure S8.2.4). The major families of mitotic kinases (polo-like, aurora and NIMA-related) are ancient, however certain Plk and Nek subfamilies diverged more recently - many Nek subfamilies appeared in the common ancestor of choanoflagellates and animals, and one family (Nek11) is novel to animals (Figure S8.2.5). Negative regulators of the cell cycle have evolved independently along different multicellular lineages[83]. The CDKN1 (Cip/Kip) family of CDK inhibitors appears unique to eumetazoans (bonafide orthologues of p21, p27 and p57 were found in *Nematostella*, *Hydra* and *Trichoplax*, but are missing in the current assembly of the *Amphimedon* genome). The CDKN2/INK4 family (p15, p16, p18, p19) is a chordate innovation (Table S8.2.1). The Myt1 subfamily of the Wee1 family of CDK-inhibiting kinases is novel to metazoans.

From the unicellular perspective cell proliferation results directly in reproductive success, but in a multicellular context inappropriate proliferation can be detrimental to the organism. The use of the CDKN1 family appears to be a uniquely animal way of controlling the cell cycle through which signals from neighbouring cells can negatively regulate the cell cycle. The novel transcriptional regulator Myc is itself regulated by various signaling pathways that metazoan cells use to communicate with each other. It is possible that cell cycle regulators novel to animals are recruited to bring the cell cycle under social control (Myc and p53 could have been pre-adaptations in the holozoan ancestor that were recruited for social control in animals/eumetazoans).

Akt signaling

The growth of multicellular animals is a consequence of both cell growth and cell proliferation. Cell growth is an outcome of the synthesis of proteins and other molecules that compose the cell. While cell division and cell growth are coupled in single celled organisms such as yeast, external and developmental signals can also regulate the extent to which cell growth results in cell proliferation. Various pathways that regulate growth in response to extracellular signals have been identified.[84-86]

The tuberous sclerosis proteins Tsc1 and Tsc2 function together as a GAP (GTPase activating protein) for Rheb. Inactivation of Tsc1 or Tsc2, or increased expression of Rheb, result in increased activation of the Target of Rapamycin (Tor) kinase in the Tor-Raptor complex (TORC1) (Figure 3b). This most likely occurs through a direct interaction of Rheb and Tor. Activation of TORC1 leads to the phosphorylation of the ribosomal protein S6 by the S6 kinase, resulting in an increase in translation of TOPs mRNAs (which encode ribosomal components and translation initiation factors). Ultimately, this results in mass accumulation. S6 kinase may inactivate eIF-4E BP, which inhibits the translation initiati on factor eIF-4E. Once phosphorylated, eIF-4E BP releases eIF-4E, thus resulting in translation of a variety of cellular proteins. All of these proteins are ancient eukaryotic proteins except for Tsc1, which appears to have originated on the stem leading to the fungal-holozoan radiation.

In mammalian cells insulin-dependent signaling leads the activation of TORC1, hence resulting in growth. Akt, another kinase downstream of the insulin receptor, inactivates eIF-4E BP. Of the main cytosolic effectors of insulin signaling, most (PI3K, PTEN, Akt, S6K, PDK-1 and Tor) are ancient eukaryotic proteins, whereas the insulin receptor substrate (IRS-1) is novel to animals. The insulin receptor itself is a eumetazoan invention (though the receptor tyrosine kinase (RTK) family is a holozoan invention) (Table 8.7.1), associated proteins Gab1 and Gab2 novel to animals. Given the absence of the insulin receptor in *Amphimedon*, it remains to be tested if an RTK molecule signals through the PI3 kinase pathway to modulate growth.

RTKs can also signal via the Ras pathway to stimulate growth - Ras stimulates Erk1/2 (an ancient eukaryotic protein, see Table S8.2.2), which then phosphorylates Mnk1/2 kinases (a metazoan-specific subfamily of MAPKAPK, see Table S8.7.2), which has been shown to phorphorylate eIF-4E, resulting in an increase of cap-dependent translation initiation (*i.e.* greater

protein synthesis). Ras can also activate PI3K (Figure 3b), and therefore activates components downstream of insulin signaling.

The Myc oncogene is thought to activate growth via transcriptional activation of a number of genes involved in protein synthesis and metabolism. As described in the previous section, the Myc subfamily of bHLH genes may have originated in holozoan stem, but *bona fide* Myc orthologs are only found in animals (Figure S8.2.3).

Cytokine signaling mediated by Janus kinase (JAK) and Signal transducer and activator of transcription (STAT) has been implicated in growth.[87] Ligand binding induces the multimerization of the cytokine receptor which results in phosphorylation of JAK. The activated JAKs subsequently phosphorylate additional targets, including both the receptors and the major substrates, STATs. STAT phosphorylation results in dimerization of STATs, which then enter the nucleus. The biological consequences of JAK/STAT pathway activation are complicated by interactions with other signaling pathways.[88-90] The best characterized interactions of the JAK/STAT pathway are with the RTK/Ras pathway. Activated JAKs can phosphorylate tyrosines on their associated receptors that can serve as docking sites for SH2-containing adapter proteins from other signaling pathways. These include SHP-2 and Shc, which recruit the GRB2 adapter and stimulate the Ras cascade. The same mechanism stimulates other cascades, such as the recruitment and JAK phosphorylation of insulin receptor substrate (IRS), which results in the activation of the phosphoinositide 3-kinase (PI3K) pathway [for more on PI3K signaling, see Foster *et al*. 2003[91]]. JAK appears to be unique to bilaterians, suggesting that the control of growth by this pathway may be a recent innovation relative to the metazoan radiation (Table S8.2.2). Consistent with the appearance of JAK in the bilaterian stem, cytokine receptors that signal through JAK/STAT also appear to be a bilaterian novelty (only one highly divergent protein known from invertebrates is known to function as a cytokine receptor upstream of JAK/STAT.[92]

Though G1 cyclin/CDK complexes are traditionally thought to promote cell cycle entry in response to growth factors, some studies suggest that Cyclin D can promote growth. Overexpression of Cyclin D and Cdk4/6 increases the growth of post-mitotic cells in *Drosophila*. In contrast, Cyclin E promotes S-phase entry but does not promote growth. It is difficult to generalize the role of Cyclin D in growth regulation to other organisms, however, this hypothetical role in the metazoan ancestor can be tested by studying the functions of Cyclin D in sponges.
Thus, it appears that ancient growth pathways acquired novel regulators that allow for complex interactions of growth and proliferation in animals.

Warts-Hippo pathway

Warts/Hippo/Mats are ancient eukaryotic proteins and their functional cassette is preserved in fungi, where they operate in the mitotic exit and septation initiation networks to promote cytokinesis.[93] Though the module is preserved, the net outcomes are different - in *Drosophila* the pathway limits cell proliferation/growth, but in yeast it enables cell division. All these proteins are present in *Arabidopsis* and known to not be involved in cytokinesis, but potentially in cell

fate specification.[94] Hippo (Hpo) (Mst in mammals) autophosphorylates and then phosphorylates Salvador (WW45 in mammals), Warts (Lats in mammals) and Mats (Mob in mammals)[95] (Figure S8.2.7). Salvador, which facilitates the phosphorylation of Warts by Hpo, appears to be a novel animal protein. Warts autophorphorylates, and then phosphorylates the transcription factor Yorkie (Yap in mammals). Yap, like salvador has two WW domains and appears to be novel to animals (though a divergent protein with one WW domain that appears equally related to Salvador and Yorkie has been identified in *Monosiga*). In *Drosophila*, the FERM-domain proteins Expanded (a *Drosophila* tumor suppressor) and Merlin (known to be a mammalian tumor suppressor) were found to be upstream regulators of the Warts/Hpo/Mats cassette. While Expanded appears to be a protein unique to *Drosophila*, Merlin is a metazoan novelty. Recently, the Fat type cadherin (a tumor suppressor and a holozoan novelty) has been linked to the Warts/Hippo pathway. The unconventional myosin Dachs, which is inhibited by Fat and inhibits the activity of Warts, is likely to belong to a novel myosin subfamily in animals. Discs overgrown (Dco), related to casein kinase 1 delta and epsilon families, promotes Fat signaling. Choanoflagellates, sponges, placozoans and cnidarians all have a putative CSK1e/d orthologue.

Thus, it appears that the ancient Warts/Hippo/Mats cassette may have been co-opted as a tumor suppressor cassette[96] in animals by coming under the control of proteins novel to the animal lineage (Table S8.2.3, Figure S8.2.7).

S8.3 Programmed cell death

Programmed cell death is executed by the caspases, a metazoan-specific family of cysteine aspartyl proteases, which are activated either by the Intrinsic or the Extrinsic pathway. The *Amphimedon* genome encodes for three putative initiator caspases that possess the characteristic pro-domains [two proteins with a caspase recruitment domain (CARD) and one with two death effector domain (DEDs)], and an expanded repertoire of putative effector capases. With the exception of three *Amphimedon* sequences (including a putative DED-containing protein) that clade within the caspase 8/10 subtypes, all other candidate sponge caspases could not be further assigned to the subtypes defined for vertebrates (Table S8.3.1; Figure S8.3.1).

The intrinsic pathway drives programmed cell death by initiating the permeabilization of the outer mitochondrial membrane and is tightly regulated by the Bcl2 oncogene family of pro- and anti-apoptotic factors. Among the pro-apoptotic family members, Bak arose in the metazoan lineage, while Bax and Bok are eumetazoan-specific. As for the anti-apoptotic protein family members, *Amphimedon*, other sponges[97,98] and *Nematostella* encode putative divergent Bcl2/Bcl-X-like proteins (Table S8.3.1; Figure S8.3.2). In bilaterians, these molecules trigger cell survival by interacting with the pro-apoptotic proteins. Since lineage-specific expansions have been reported for both the Bcl2 and the caspase family in other animal groups,[99,100] it is possible that relaxed constraints on the evolution of these gene families have allowed different invertebrate lineages to exploit alternative apoptotic networks that are not found in the bilaterians, perhaps reflecting life-cycle traits that necessitate a higher cell turnover (*e.g.* metamorphosis and regeneration) or simply the co-option of these protein families in a non-apoptotic functional context. Perhaps as a consequence, the protein families that regulate both

the anti-apoptotic Bcl2 and the capsase families in bilaterians, namely the BH3-only proteins (Bid, Bim, and NOXA) and the BIR domain-containing proteins (cIAP1 and cIAP2), are not found in *Amphimedon* and other early-branching metazoan phyla (Table S8.3.1), suggesting that the first metazoans either lacked these additional layers of control or used different genes altogether. Mitochondrial permeabilization releases various proteins including the ancient AIF (apotosis-inducing factor) that contributes to caspase-independent apoptosis, metazoan-specific APAF1 (apoptotic protease activating factor 1), and eumetazoan *sensu strictu (s.s.)*-specific caspase-activated DNase (CAD) and its regulator ICAD.

In the extrinsic pathway, external signals that lead to apoptosis are typically detected by death domain-containing transmembrane receptors belonging to the tumor necrosis factor receptor (TNFR) family. These receptors rely on their death domain for interactions with downstream adaptors. The *Amphimedon* genome encodes a nerve growth factor receptor (NGFR) p75-like protein, though it lacks the crucial death domain that is seen in *Nematostella* and other animals.[99] Classic death TNFRs (*i.e.* Fas, DR4, DR5 and TNFR1) appeared late in the vertebrate lineage. [99,101] Since the intrinsic cascade is composed of components that are found in both metazoan and non-metazoan groups, it is likely to have been the original mechanism for inducing apoptosis.

S8.4 Germline specification

In sexually-reproducing organisms, gametes are the cells that transmit genetic material across generations. The precursors of all cells that can become gametes are the germ cells, collectively known as the germ-line. Germ cells are unique among metazoan cell types in that they must become highly specialized to produce either sperm or egg while simultaneously retaining their potential to give rise to all types of differentiated cells, including extra-embryonic tissues, in the adult organism.[102] In most bilaterian animals, the germ-line originates as primordial germ cells (PGCs), a population of undifferentiated stem cells that are capable of undergoing meiosis and that will give rise exclusively to germ cells.[103] This single founder population of germ cells is segregated from diploid somatic cells at a single point in time during embryogenesis, and thereafter is not significantly amplified, replaced or renewed throughout the entire life of the animal.[103]

Molecular markers have become a widely accepted way to identify germ cells upon their first appearance during embryonic development and thus to establish their embryonic origin. Conserved members of the Vasa and Nanos gene families have been shown to be important for the specification and differentiation of germ cells from PGCs in diverse bilaterians (reviewed in Extavour and Akam 2003[103]), although the mechanisms may vary.[104] In nonbilaterian animals studied to date, vasa, nanos and PL10 seem to be associated with specifying the germ line in the cnidarian *Nematostella vectensis*[105], but perhaps not so in the ctenophore *Mnemiopsis leidyi* (Pang & Martindale, pers. comm.). It is therefore unclear at present whether the germ line is homologous in all metazoans, or whether alternative genes may be involved in specifying the germ cells in the basal metazoan phyla.

Poriferans reproduce both asexually and sexually, the latter usually as simultaneous hermaphrodites with internal fertilization.[106] Oocytes and spermatocytes undergo gametogenesis in the mesohyl to form eggs and sperm, respectively[106]; gametes derive from various subpopulations of pluripotent mesechymal cells.[107] Sponges lack gonads, and gametes instead occur either in simple clusters (sperm, sometimes eggs) or individually (eggs, usually), but in both cases widely distributed throughout the mesohyl of the adult, although within diffusion distance of a canal or chamber. In contrast to the single developmental origin of germ cells in bilaterian animals, sponges are thought to generate germ cells continuously throughout their adult reproductive life.

The *Amphimedon* genome contains several genes implicated in primary germ cell development in eumetazoans (Table S8.4.1). Consistent with an earlier report on the freshwater sponge *Ephydatia fluviatilis*[108], these include a single vasa and a single PL10 gene, in addition to 18 other genes belonging to the DEAD-box helicase family. We also find a single zinf finger family nanos gene, and 3 piwi genes (see Grimson *et al.* 2008[109] for details on the piwis). The presence of these core bilaterian germline genes suggests that sponges might use similar genetic tools to bilaterians to segregrate the germ line, although we currently have no expression or cell lineage data to validate this. The only existing evidence in non-bilaterians comes from the cnidarian *Nematostella vectensis*[105], in which vasa, nanos and PL10 seem to be associated with specifying the germ line late in embryonic development. One notable difference is that the *N. vectensis* genome contains 2 vasa, 2 nanos and 1 PL10 gene, compared to just one of each in *Amphimedon queenslandica*; interestingly, the genome of the placozoan *Trichoplax adhaerens* contains only a single PL10 and a single nanos gene. Both the *A. queenslandica* and the *N. vectensis* genomes encode 3 piwis, but none are detectable in *Trichoplax adhaerens*. In *A. queenslandica*, we also find two mago nashi, three tudor-related, a single pumilio and a single a par-1 gene, all of which are present throughout the eukaryotes and are known in other eumetazoans to play an essential role in germ cell determination often via interaction with either vasa or nanos.

Given the lack of both mesoderm and a gonad in sponges, it is particularly interesting that we are unable to identify a DM DNA-binding-domain-containing *DMRT1* gene (vertebrate ortholog of *Doublesex* and *Mab-3*). This gene has been implicated in many eumetazoan taxa as playing a highly conserved role in development of the mesoderm-derived somatic gonad.

S8.5 Signaling pathways

Wnt Signaling Pathway

Detailed description of the origins of the Wnt pathway can be found in the paper in preparation by Adamska *et al*. (Adamska, unpublished) Some highlights are described below and in Table S8.5.1.

The Wnt pathway is a critical factor in determining polarity in eumetazoan *s.s.* development (*e.g.* Wikramanayake *et al*., 2003[110]; Broun *et al*., 2005[111]; Lee *et al*., 2007[112]; Momose *et al*., 2008[113]), and the polar expression of Wnt ligands in *Amphimedon queenslandica* embryos may

indicate a more ancient, pan-metazoan ancestry of the Wnt pathway's role in axial patterning.[7] *Amphimedon* has 3 Wnt family genes, but these cannot be confidently assigned to the defined bilaterian orthology groups, nor do they appear to represent a lineage-specific expansion. Of note, a dramatic expansion of the Wnt family has occurred after poriferan divergence, with eumetazoans *s.s.* possessing 12-15 Wnt genes.[114]

The reception of Wnt ligands is carried out by Frizzled (Fzd) receptors, in complex with low density lipoprotein receptor related proteins (LRP5/6s). *Amphimedon* has 2 Fzd genes and while Fzd-related genes have also been described from amoebozoans,[115] no other Fzd-like genes are present in any other non-metazoan organisms making the ancestry of this family unclear. LRP5/6s are single-pass multidomain transmembrane proteins with no discernable homologs outside the Metazoa. Upon ligand binding, Dishevelled (Dsh) interacts with Fzd whilst Axin is bound by LRP5/6. These interactions cause the dissolution of a downstream cytosolic protein complex, the so-called destruction complex. Comprised of Axin, Adenomatous Polypotis Coli (APC) and GSK, the complex phosphorylates, and subsequently degrades, cytosolic β-catenin in the absence of a Wnt signal. GSK is a pan-eukaryotic kinase, but APC and Axin are not found outside the Metazoa. Non-bilaterian Axins and APCs are lacking specific protein-protein binding motifs that are required for the correct formation of the destruction complex in their bilaterian counterparts (*e.g.*, the sponge and cnidarian APC and Axin proteins appear to be missing β-catenin binding domains). However it is unclear whether this truly reflects a lack of interaction between these proteins. For example, whilst APC lacks recognisable Axin binding domains, APC binding domains are detected in Axin - suggesting that the molecules can interact at some level.

In the absence of nuclear β-catenin, TCF/LEF proteins form a transcriptional repression complex by recruiting the co-repressor Groucho and Histone deacetylases. When Wnt signaling stimulates the nuclear accumulation of β-catenin, Groucho is displaced and a transcriptional activation complex of β-catenin, Tcf/Lef and the Histone acetylase CBP is formed instead. TCF/LEF and Groucho, like the majority of Wnt pathway components, also likely arose on the metazoan stem (Figure S8.5.1). A *β-catenin* related gene (*Aardvark*) is present in amoebozoans, but Aardvark is more akin to certain plant proteins that also share armadillo repeats, in both sequence similarity and domain composition.[115,116]

TGF-β Signaling Pathway

Transforming growth factor-β (TGF-β) signaling (reviewed by Massagué 2000[117]) is restricted to the Metazoa; neither ligand nor receptor molecules are found outside the animal kingdom (Table S8.5.2). TGF-β pathway receptors are serine threonine kinases (STKRs) of two types – Type I and Type II. Both types are present in *Amphimedon*, but these receptors cannot be further assigned to eumetazoan subfamilies within these groupings (Figure S8.5.2) (see kinome analysis for further details on STKRs). Two major clades of ligands are recognized, the TGF-β *sensu strictu*/TGF-β related (*e.g.* Activins, Leftys, GDF8s), and BMP related (*e.g.* BMPs, Nodals).[118] In phylogenetic analyses, one *Amphimedon* gene lies outside these clades, along with other divergent ligands such as GDF9/15 (Figure S8.5.3). Five *Amphimedon* genes group together, suggesting an independent expansion event, and fall within another divergent clade,

the *DVRs*. A further two *Amphimedon* genes are nested within the TGF-β related clade. Their placement as sister to the TGF-β *s.s* subclass was found to be consistent across a number of phylogenetic analyses (data not shown). To date, TGF-β *s.s* ligands have only been identified in deuterostomes, with no members found in genome screens of *C. elegans*, *Drosophila* or *Nematostella,*[118] so this placement of *Amphimedon* ligands warrants further investigation.

Transmission of the TGF-β signal from the membrane to the nucleus occurs via Smad family proteins – another metazoan invention. *Monosiga* has a Smad-like MH2 domain, but this is coupled with a C2H2 zinc finger as opposed to the metazoan Smads which comprise of a MH1 and MH2 domain only. In contrast to the lack of phylogenetic resolution among *Amphimedon* TGF-β receptor and ligands, *Amphimedon* Smads can be assigned to recognized eumetazoan subclasses (Figure S8.5.4). Type I receptors recruit and phosphorylate receptor regulated Smads (R-Smads, Smad 1/5, Smad 2/3) that form multisubunit complexes with common partner Smads (Co-Smads, Smad4) before entering the nucleus to affect a response. Both R-Smads and Co-Smads are found in *Amphimedon*. Inhibitory Smads (I-Smads, Smad6/7) interfere with either the phosphorylation of R-Smads, or the formation of the R-smad/Co-Smad complexes. I-Smads have not been located outside the Eumetazoa, suggesting that the regulatory activity of I-Smads did not evolve until after the divergence of Porifera and Placozoa.

In the nucleus, Smad complexes recruit a number of proteins including Fos/ATF3 and Jun - metazoan subfamilies of bZIP transcription factors (Figure S8.5.5)- and Myc and Max, which belong to the bHLH superfamily of transcription factors. While a *Myc*-like gene is present in *Monosiga*, it does not possess a classic Myc domain, in contrast, Max is definitively present in *Monosiga* suggesting that the Max subfamily originated in the holozoan stem lineage (Figure S8.5.6). Smads also recruit the co-activators CBFβ and CBP (which arose in the Metazoa), and the co-repressor, Ski/Sno. Ski/Sno is present in all eumetazoans; in *Amphimedon* while the most similar gene does possess a Smad binding domain, it lacks the DNA binding domain of classic Ski/Sno proteins.

Extracellular inhibition of TGF-β signaling can occur via eumetazoan specific 'ligand trapping' proteins such as Follistatin and members of the CAN (Cerebrus/DAN) family. Chordin is another inhibitory molecule that also acts by sequestering TGF-β ligands extracellularly; it is restricted to cnidarians and bilaterians. The E3 ubiquitin ligase Smurf is another pathway regulator that acts by targeting R-Smads and receptors for degradation - Smurfs are an animal specific subfamily of E3 ligases (Figure S8.5.7).

To summarize, the primary components of the TGF-β pathway (ligands, receptors, Smads) have emerged in the metazoan stem lineage, prior to poriferan and placozoan divergence, with no discernable precursors present in choanoflagellates, indicating that TGF-β signaling is an ancient metazoan synapomorphy. However, clear differences exist in the potential activity of the TGF-β pathway in the Porifera when compared with Eumetazoa, as the addition of multiple regulatory elements - I-Smads, ligand traps and SARA - has occurred after the divergence of sponges.

Hedgehog Signaling Pathway

The canonical Hedgehog signaling pathway is activated by the binding of a secreted ligand from the Hedgehog (Hh) family to the multi-transmembrane receptor Patched (Ptch) which in turn releases a second transmembrane protein, Smoothened (Smo), from Patched repression (reviewed by Ma 2008[119]*).* No evidence of Hh ligands outside the Eumetazoa *s.s* has been found (Table S8.5.3). While the *Monosiga* and *Amphimedon* genomes do possess Hh N-terminal signaling domains, these are located within the large membrane-bound Hedgling proteins instead of linked to an autocatalytic intein domain as in Eumetazoa *s.s.*[120] It has been proposed that Hedgling represents an alternate ligand for the Hh pathway, implying that Hh signaling originated as a short-range cell-cell mechanism with the addition of the diffusible ligand occurring later, in the proto-eumetazoan stem. However the *Amphimedon* genome does not encode a Ptch receptor. Although there are two Ptch-like proteins in *Monosiga*, the most similar gene model in *Amphimedon* is a member of the Ptch-related Niemann Pick-C family of sterol-sensing receptors (Figure S8.5.8). *Amphimedon* also lacks the transmembrane protein Dispatched (Disp) that transports the Hh ligands across the membranes of signaling cells, yet again, there is a Disp-like molecule in *Monosiga* (Figure S8.5.8). *In lieu* of Ptch, a candidate receptor for Hedgling is the Ihog/CDON family of IgCAM proteins. This family binds the Hh ligand in vertebrates and flies, and gene models with similar domain configurations to Ihogs/CDONs are present in the *Amphimedon* genome, although not in *Monosiga*.

As Smo is not found outside Eumetazoa *s.s*. (Figure S8.5.9), the initiation of Hh signaling through Hh-Ptch-Smo interactions is a eumetazoan *s.s.* invention. A further invention is the Hedgehog interference protein (Hhip), which can also bind Hh ligands and in doing so regulates the availability of Hh to the Ptch receptor. Signal transduction downstream of Smo is not wholly conserved between flies and vertebrates . While both systems make use of Sufu, CK1, GSK and PKA (all of which are pan-metazoan proteins), whether Fused (pan-eukaryotic) and Kif7/Cos2 (a metazoan-specific class of kinesins) (Figure S8.5.10) are also common components remains unresolved. The outcome of canonical Hh signaling is the regulation of Gli/Ci transcription factors that are common to all metazoans and represent an animal-specific subfamily of zinc finger binding proteins (Figure S8.5.11). The pan-eukaryotic kinases CK1 and GSK function in the phosphorylation and processing of Gli/Ci.

Despite the cytosolic components of the Hh pathway being common to all Metazoa, the absence of Hh and Smo from poriferans and placozoans suggests that canonical Hh signaling is a eumetazoan *s.s.* synapomorphy. A subject that requires further investigation is the presence of *Ptch*- and *Disp*-like genes in *Monosiga*.

Notch Signaling Pathway

The Notch signaling pathway is unusual in that both the ligand and receptor molecules are membrane-bound, meaning that the signal can only be propagated between directly neighboring cells (reviewed by Bray 2006[121]). The ligands (Delta and Jagged/Serrate) and receptor (Notch) are multidomain proteins, and whilst the majority of these domains (EGF, ANK, NL) are not restricted to the Metazoa, the combination of the domains is thought to be metazoan-specific

(Table S8.5.4). Delta ligands are present in the *Amphimedon* genome, but Jagged/Serrate ligands (which have an additional VWC domain and expanded EGF region) appear later, in the Eumetazoa *s.s*. Notch receptors in bilaterians contain two additional domains, Nod and Nodp, which are variously present in some non-bilaterians (both in *Nematostella*, Nod only in *Hydra* and *Trichoplax*), but absent from *Amphimedon*. Of interest, *Monosiga* also possesses a gene model containing the same domain configuration as *Amphimedon* Notch, albeit with a greatly reduced number of EGF repeats, perhaps hinting at the origins of this molecule.

Prior to reaching the membrane, both the receptor and ligand molecules undergo glycosylation by the holozoan specific *o*-fucosyltransferase, and the metazoan-specific Fringe proteins (which are related to the pan-eukaryotic β3GLT glycosyltransferase superfamily) (Figure S8.5.12). The degree of glycosylation affects the ligand/receptor binding abilities and is a key regulatory aspect of the pathway and *Amphimedon* possesses six *Fringe* genes.

Signal transduction in the Notch pathway occurs via regulated intramembrane proteolysis. The first cleavage of Notch is by Furin and results in the formation of a heterodimeric receptor structure that is an absolute requirement for signaling in vertebrates, but has not been shown to be essential for signaling in *Drosophila*. At the outer membrane, Notch is then further cleaved by members of the ADAM amily of proteins (ADAM10/17), releasing the extracellular region of the receptor. Subsequently the Notch intracellular domain (NICD) is released by the action of the transmembrane γ-secretase complex. These proteins that process Notch, thereby facilitating the signaling event, are common throughout the Metazoa, many of them with far more ancient origins (*e.g.* Pan-Eukaryota : Presenilin, Nicastrin). On reaching the nucleus, the NICD forms a transcriptional complex with the CSL DNA binding protein (metazoan-specific) that transforms CSL from a transcriptional repressor to activator. This implicates a number of other nuclear cofactors including the pan-metazoan HDAC, CBFβ and CBP nuclear proteins. The cofactor Mastermind is an integral part of this complex in bilaterians, but while present in cnidarians, no *mastermind* gene has been identified in *Amphimedon* or *Trichoplax*.

On the whole, across all metazoans, the core members of the Notch pathway can be identified, indicating that this signaling system arose on the protometazoan stem. While many of the cytosolic components are pre-metazoan, the ligand and receptor molecules are animal-specific.

Growth Factor, GPCR and Ras signaling

Growth factors activate cellular proliferation and/or differentiation during development and throughout the adult lifespan of animals. Growth factor signaling is propagated via the reception of growth factors by tyrosine kinase receptors on the surface of receiving cells (*e.g.*, Epidermal Growth Factor Receptor, EGFR; Fibroblast Growth Factor Receptor, FGFR; Placental Derived Growth Factor Receptor, PDGFR – see kinome section for analysis of receptor origins). Aberrant growth factor signaling has major effect on organism viability due to their mediation (both positive and negative) of cell proliferation. These activities are a result of the induction, by growth factors, of nuclear localized proto-oncogenes such as *Fos*, *Myc* and *Jun*.

The *Amphimedon* genome contains a gene model with strong similarity to the *EGF* ligands of humans and mice, indicating that this ligand arose prior to poriferan divergence (Table S8.5.5). No similar gene is found in *Monosiga*, suggesting that *EGF* was a protometazoan innovation. In contrast, *FGF*, *PDGF* and *TGFα* are not present in the sponge or *Trichoplax* genomes, suggesting that they evolved later in eumetazoan history. FGF ligands are first recognized in cnidarians, where they have undergone major diversification: fifteen FGFs have been identified in *Nematostella.*[122] PDGFs are also present in Cnidaria, but there is only a single member in the *Nematostella* and *Hydra* genomes indicating that this ligand has not undergone the same level of expansion in Cnidaria as has been the case for the FGFs (Table S8.5.5).

An analysis of the origins of components downstream of growth factors and GPCRs is described in Table S8.5.5.

S8.6 Developmental transcription factors

A detailed discussion of several developmental transcription factor families in *Amphimedon* can be found in Larroux et al 2008.[123] A list of *Amphimedon* orthologs in ETS, HMG, bZIP and C2H2 families is shown in Table S8.6.1. Origins of developmental transcription factors involved in neurogenesis are shown in Table S8.9.1.

Three Gli C2H2 zinc finger transcription factors (TF) were detected in the *Amphimedon* genome (two Gli2/3-like genes and one Glis1/3-like gene) but we found no Snail or Zic C2H2 zinc finger TFs. A sponge gene was characterised containing two GATA zinc fingers, a BIM domain, and a DUF1518 domain. 17 BZIP, 13 non-Sox HMG, and 9 ETS TF genes were identified in the *Amphimedon* genome and appear to belong to various subfamilies. Of these three classes, ETS appears to be metazoan-specific with only two ETS genes and one ERM gene clearly belonging to these eumetazoan subfamilies.

Of the transcription factor classes with important roles in metazoan development, NK, paired-like, Pax, POU, LIM-HD, and Six homeobox genes as well as ETS, mef2, Sox, and nuclear hormone receptor genes are present in the sponge genome but not in *Monosiga* or other non-animal eukaryotes to date. These classes hence appear to have arisen in the lineage leading up to the metazoan LCA. In contrast, Fox, BZIP, non-Sox HMG, TALE, and bHLH transcription factors are more ancient with the former two present in *Monosiga* and fungi and the latter three also in plants.[81,123-125] Although they are absent from *Monosiga* and fungi, T-box genes are likely to be a holozoan invention as they are present in *Amphimedon* [123] and a non-choanoflagellate holozoan protist.[126] The developmental zinf finger GATA and Gli genes are present in the sponge genome but not Snail and Zic. As Gli, Snail, and Zic C2H2 zinc finger TFs are present in *Nematostella* but not found in *Monosiga*, it appears that Gli is a metazoan innovation while Snail and Zic are eumetazoan innovations. GATA are ancient eukaryotic genes.

S8.7 Kinases

All kinases in the *Amphimedon* genome were identified and classified using previously published methods.[127] Many kinases were fragmentary; most of these mapped to the end of contigs, or next to internal gaps, though some mapped well within contigs and may reflect assembly limitations. 23 fragmentary predictions from short contigs with >95% AA identity to longer predictions were removed as possible second haplotypes or assembly errors. Gene models with homology to other kinases but without a kinase domain were omitted. All kinases with their assigned classes are shown in a separate supplemental spreadsheet (Supplemental_table_S8.7.1.xls).

The 705 *Amphimedon* kinases include representatives of more than 70% of all human kinase classes (compared with 59% in choanoflagellate and 83% in sea anemone, see Tables S8.7.1 and S8.7.2). Curiously, the kinomes of advanced bilaterian models *Drosophila* and *C. elegans*, have only 77% and 70% respectively of human classes, with extensive gene loss outpacing invention of new kinases, further highlighting the signaling complexity in early metazoans (Fig S8.7.1). Several kinase losses are coordinated within pathways. For instance, the MEKK2-MEK5-Erk5 variant of the MAPK cascade is fully present in Amphimedon and Nematostella, but all three members are lost in flies and nematodes. Several other pathways (Fig. 3) show gradual addition of kinase and other components over evolution, indicating that these are highly modular, and can be functional in many different combinations. Of 196 defined kinase classes, 47 kinase classes are found in *Amphimedon* and eumetazoa, and another 17 are eumetazoan-specific, and 12 kinase classes are lost from *Amphimedon* (Table S8.7.2). 150 classes have 1-2 members, with over half of the kinome (360 genes) being found within 11 expanded classes of the TK and TKL groups, including the Met (58 genes), Eph (64) receptor tyrosine kinases and a sponge-specific subfamily of Src kinases (Src-Aque1, 26 members).

Most kinases have domain combinations similar to their metazoan counterparts, but there are several exceptions. In particular, many receptor tyrosine kinases have unusual extracellular domains, probably responding to extrinsic rather than intrinsic signaling. One Eph-like and one Met-like receptor lack extracellular domains entirely and appear to be membrane anchored by PH domains instead. Unusual intracellular domain combinations include death domains in 3 members of the extended Src family.

Table S8.7.1 (uploaded as a supplemental Excel spreadsheet): The *Amphimedon* kinome. A list of all known kinases in *Amphimedon* and their kinase and Aqu1 IDs. Full sequences and additional data available at http://kinase.com/amphimedon/. These are derived from standard gene models, sometimes edited, and from de-novo kinase gene prediction. Aqu1 gene model IDs are listed where they overlap with kinase models; cases where no Aqu1 model includes the kinase catalytic domain are noted.

S8.8 Cell-cell and cell-matrix adhesion and formation of polarized epithelia

Animal cell-cell adhesion molecules (CAMs) are often large multidomain proteins with highly variable domain architecture (particularly with variation occurring in the numbers of tandemly repeated domains). We focused largely on cell adhesion superfamilies which are characterized

by the presence of a particular type of extracellular domain. We searched representative eukaryote, opisthokont, holozoan and animal genomes for the occurrence of putative proteins containing leucine rich repeat (LRR), cadherin, immunoglobulin-like (Ig) and Ig plus fibronectin type III (FN3) domains in association with predicted signal peptides and transmembrane helices. Putative LRR-containing transmembrane proteins were found in plant, *Dictyostelium*, *Monosiga* and animal genomes indicating a probable ancient eukaryotic origin. Cadherins and IgCAMs were found in *Monosiga* and animal genomes but not in fungal or non-opisthokont genomes. A single *Monosiga* protein was found to contain both Ig and FN3 domains, however the domain architecture consisting of an N-terminal stretch of Ig domains followed by a stretch of FN3 domains was found to be specific to metazoans, with many such proteins present in the *Amphimedon* genome. We also searched for orthologs of Neurexin I/II/III like proteins, a family of proteins that function in synaptic adhesion,[128] and found these to be restricted to eumetazoans *s.s.*.

For the extracellular matrix (ECM) proteins we focused more closely on orthology groups characterized by particular combinations of domains. All ECM protein families were found to be metazoan-specific with the exception of collagen triple helix repeat encoding proteins which were also found in the *Monosiga* genome. Fibrillar collagen and thrombospondin-like proteins appear to be present in *Amphimedon,* whereas agrin and netrin appear to be specific to the Eumetazoa. All analyzed ECM binding transmembrane proteins were found to be metazoan-specific, with good homologs of integrin α, integrin β and dystroglycan family proteins present in the sponge genome.

The results of these analyses are summarized in Table S8.8.1.

Tissues with an epithelial grade of organization are generally believed to be a eumetazoan innovation. There are three defining characteristics of epithelial tissues: (a) aligned apical-basal polarity of component cells, (b) adhesion (cell-cell) via belt-form junctions (adhesive or occluding), and (c) adhesion (cell-ECM) to an underlying basal lamina. Studies in bilaterian model organisms have allowed for the identification of many of the genes involved in giving rise to these characteristics, particularly in the epithelia of *Drosophila* and of vertebrates. We used sequence similarity and domain searches to identify orthologs of these genes in representative animal, choanoflagellate and fungal genomes (Table S8.8.2).

S8.9 Neuronal genes in Amphimedon

Transcription factor gene families

Although metazoan transcription factor (TF) classes antedate metazoan cladogenesis, many specific bilaterian families arose through a duplication and divergence events early in eumetazoan evolution[123-125,129,130] When assessing the presence of TF genes and families in *Amphimedon* that are associated with the specification, determination and patterning of neurons in bilaterians and cnidarians,[131] it is clear that many regulatory gene families are eumetazoan-specific gene families, although *Amphimedon* does possess some orthologues of

genes that control neurogenesis (*e.g.*, PaxB, Lhx, SoxB, Msx, Mef2, and group A bHLH neurogenic factors; Table S8.9.1).

Synaptic genes

Genes that encode proteins associated with both the post-synaptic density and pre-synaptic element are well-represented in Amphimedon (Tables S8.9.2 and S8.9.3). These include the post-synaptic scaffolds DLG, SHANK, HOMER, GRIP, GRASP, SCRIB, and MAGI, as well as associated signaling molecules, such as Cript, SPAR, GKAP, CIT, NOS, KALRN, and SYNGAP. Scaffolding molecules that coordinate the localization of synaptic vesicles, calcium channels, and signaling machinery to the pre-synaptic compartment, such as LIN10, LIN7, UNC13, RIMBP, PTPRF, and PPF1A are also present (Figure S8.9.1). The Amphimedon genome encodes genes for synaptic vesicle proteins that allow vesicle exocytosis to be regulated in response to calcium influx. These genes include Synaptophysin (binds to VAMP to regulate its availability for SNARE complex formation), SV2 (calcium uptake), and Synaptotagmin (calcium sensor). UNC13, also present, regulates SNARE complex formation by binding to Syntaxin. Trans-synaptic adhesion genes such as cadherin, beta-catenin, and cortactin are present.

While most of the core pre- and post-synaptic genes are present in Amphimedon, some key genes are conspicuously missing. For example, there are no members of the ionotropic glutamate receptor family,[132] although neuronal type metabotropic glutamate receptors, as well as homologs of dopamine and serotonin receptors are present. Amphimedon also lacks a homolog of RIMS, a central presynaptic scaffold involved in priming synaptic vesicles for fusion. While Amphimedon does possess a homolog of the ephrin receptor, a protein functioning in axon guidance, the ephrin ligand is not present. Several other axon guidance proteins (e.g., slit, netrin, unc-5, and robo) also appear absent from the genome.

Neuropeptide and neurohormone processing and secretion

Proteins involved in the key steps of neurohormone and neuropetide production are predominantly encoded by genes belonging to eight main families [133]. These proteins include (i) peptidases that cleave the immature neuropeptide/hormone precursors to produce distinct functional subunits, (ii) enzymes that are required for specific modifications (such as C-terminal amidation, N-terminal acetylation and pyrolation) of these subunits, and (iii) molecules required for their Ca2+-dependent release. The *Amphimedon* genome encodes proteins that belong to these eight gene families (Table S8.9.4).

We found in *Amphimedon* 10 proprotein convertases (PC, also known as proprotein convertase subtilisin/kexin type, PCSK), which are key proteins in the processing of several proteins to give biologically active neuropeptides or peptide hormones.[134] Two major types of PCs (PC1/3 and PC2) are expressed exclusively in neuroendocrine tissues in bilaterians. Five PC2-like proteins but no PC1/3 are found in *Amphimedon* (Figure S8.9.2). Arginyl aminopeptidase B (AP-B) are other proteins involved in the cleavage of immature neuropeptide/hormone precursors in

vertebrates.[134] *AP-B* belongs to a gene family that also contains, as close relatives, two other subfamilies in vertebrates, *arginyl aminopeptidase O (AP-O),* and *leukotriene A4 hydrolase* (*LTA4H*). Interestingly, *AP-B* genes are only found in deuterostomes and molluscs while in other species of eukaryotes either *AP-O* and *LTA4H* or only *LTA4H* are present. The *Amphimedon* genome contains a *LTA4H* gene but no *AP-B* and *AP-O* genes (Figure S8.9.3). The c*arboxypeptidase* (*CP*) gene family encodes proteases, one of which, CP-E, is involved in the cleavage of immature neuropeptide/hormone precursors.[134] Other proteins of the family, CP-D in particular, may also be involved in the same process.[135] The *Amphimedon* genome encodes a single *CP* gene of the *CP-D* type (Table S8.9.4). *CP-E* is also absent from the genomes of *Trichoplax*, *Nematostella*, and *Hydra*, suggesting that the presence of *CP-E* is specific to bilaterians. A fourth family of proteases, cysteine cathepsins are also involved in the proteolytic processing of neuropeptides and peptide hormones.[134] The *cathepsin* genes form a large family with several subfamilies among which cathepsin-L is involved in neurosecretion.[136] There are 25 cathepsin genes in *Amphimedon*, of which 3 are closely-related to eumetazoan cathepsin-L genes (Figure S8.9.4).

C-terminal amidation is required for the biological activity of many animal peptides and has been shown to successively involve two enzyme activities, peptidylglycine α-hydroxylating monooxygenase (PHM) and peptidyl-α-hydroxyglycine α-amidating lyase (PAL) activities.[137] In some species, such as vertebrates, the two enzymes are co-synthesized as adjacent domains of a bifunctional protein, peptidylglycine α-amidating monooxygenase (PAM) that, following specific cleavage events, or as a consequence of alternative splicing, produces monofunctional PHM and PAL enzymes. In other species, such as *Drosophila*, the two enzymes are encoded by distinct genes.[138] In still other species, such as *Caenorhabditis*, both monofunctional *PHM* and *PAL,* and bifunctional *PAM* genes exist.[133] In both *Amphimedon* and *Trichoplax*, a single *PAM* gene can be found, but neither *PAL* nor *PHM* genes (Figure S8.9.5). Another post-translational modification of neuropeptides is N-terminal pyrolation which is mediated by a glutaminyl-peptide cyclotransferase (GC).[139] A single *GC* gene is found in the genome of *Amphimedon* (Figure S8.9.6).

Secretion of neuropeptides is a Ca2+-dependent process that involves several proteins, including the calcium activated protein for secretion (caps) and protein tyrosine phosphatase receptor type N (ptprn) (also named islet cell autoantigen, ia2), both of which are markers for neurosecretory cells in bilaterians[140-142]. The *Amphimedon* genome encodes a single representative for each of these metazoan gene families (Figures S8.9.7 and S8.9.8).

Taken together, these data suggest that *Amphimedon* has the molecular machinery required to produce secreted biologically active peptides. However, sequence similarity searches failed to detect homologs of the known neuropeptides and peptide hormones of bilaterians, suggesting that the *Amphimedon* peptides are different from those found in bilaterians.

G-protein coupled receptors

The GRAFS classification system[143] was used to place all *Amphimedon* G-protein coupled receptor (GPCR) gene models into recognised families. GPCRs were compiled from the draft assembly and parsed into Rhodospin, Frizzled/Taste, Glutamate and Adhesion/Secretin families based on identity with sequences in NCBI and domain architecture (shown in the uploaded spreadsheet Supplemental_table_S8.9.5.xls). Many of the Rhodopsin GPCR genes are organised in head-to-tail tandem arrays ranging in size from 2 to 18 (Figure 8.9.9; Supplemental_table_S8.9.5.xls). A number of small Glutamate and Adhesion/Secretin gene clusters, ranging in size from 2 to 4 genes, were observed. Phylogenetic analysis of Rhodopsin GPCRs reveals that a majority of *Amphimedon* genes form a species-specific clade, although there are a small number of genes with affinity to other defined clades, including LGR/Hormone receptor, opsin/prostanoid receptor, gonadotropin releasing hormone/neurotensin/somatostatin receptor and mas oncogene subclasses (Figure S8.9.10).

S8.10 Allorecognition and innate immunity

While some genes related to metazoan immune receptors are present in early eukaryotic lineages, others are restricted to metazoans (Table S8.10.1).[144,145] For instance, *Amphimedon* encodes putative proteins that display the characteristic tripartite domain structure found in NLR proteins, but with an N-terminal death domain. This gene family contributes to a wide variety of functions in vertebrates, including immune recognition and apoptosis.[146] Multiple scavenger-receptor cysteine-rich (SRCR) proteins can also play a role in immunity.[147] The origin of these proteins is ancient,[148] but some of the domain configurations they display are unique to animals (*e.g.,* association with the complement control protein and the fibronectin domains in *Amphimedon* and other sponges[149]).

The Toll-like receptors (TLRs) and Interleukin1 receptors (IL1Rs) are other crucial immune receptors. A protein related to the TLR has been reported in the sponge *Suberites domuncula* but is atypically short and lacks the diagnostic LRRs.[150] While no true TLRs are present in *Amphimedon* either, the demosponge possesses two putative receptors with an intracellular TLR-like Toll/Interleukin1 receptor/resistance (TIR) domain and IL1R-like Igs, suggesting that an ancestral form to the receptor superfamily evolved before metazoan cladogenesis and independent duplication and divergence led to the diversified TIR-containing receptors present in sponge and cnidarian lineages (Table S8.10.1).

S9. Novelty Analysis

S9.1 Clustering of orthologous animal genes

To identify orthologous sets of genes a phylogeny-informed clustering method was applied.[16] Briefly, the clustering implements a graph approach where, in the first step, mutual

best hit pairs between the species are identified. In the second step, paralogs that have shorter edges to the proteins from the same species than to the outgroup are added. To generate a more accurate all-against-all homology relation, position specific scoring matrices (PSSM) as implemented by PSI-BLAST were used. This has proven to be particularly useful for fast-evolving sequences from *Drosophila*, or divergent domains from basal metazoans. Very similar paralog sequences might introduce bias into the PSSM. Therefore, this redundancy has to be removed before running PSI-BLAST. This was accomplished in each species by pre-clustering sequences that share high similarity (using standard BLASTP scores) and are connected by edges shorter than those leading to the other species.

S9.2 Type I, II and III novelties

Gene families that potentially appeared at a given node in the animal tree can be obtained by requiring absence of the orthologs in the outgroup and their presence in both ingroups (*e.g.*, metazoan novelties are defined as being absent in non-animals+*Monosiga* and present in *Amphimedon* and one of the eumetazoans) . Gene family clusters from the PSI-BLAST clustering (S9.1) were used. Type I novelties are defined as gene families that share no significant (1E-10) PSI-BLAST homology to any of the outgroup sequences. Type II novel gene families have a new domain absent in the outgroup and are subdivided into 2a (domain absent in all outgroups) and 2b (domain absent in outgroups, except *Monosiga*). Type III possess a new domain combination (architecture) and are subdivided into 3a (architecture absent in all outgroups) and 3b (architecture present in *Monosiga*). Several gene families that are classified as novel by the clustering do not fall into any of the groups, i.e., have an outgroup PSI-BLAST hit, and no novel domain architecture. These cases can be classified as novel gene families that acquired new function by accelerated divergence and not by gain or loss of a domain. Gene clusters that appear to be novelties by these measure at various nodes on the animal tree are listed in the uploaded spreadsheet Supplemental_table_S9.2.1.xls.

S9.3 PFAM domain analysis

Novel PFAM domains and architectures were obtained in the same way as for the gene families (separately uploaded spreadsheet Supplemental_table_S9.3.1.xls). Detailed PFAM domain analyses of death domain and laminin proteins presented in Figure 2 of the main paper are discussed below.

PFAM domains that originated in the metazoan ancestor show similar functional distribution as families from the protein clustering (see uploaded file Supplemental_table_S9.3.1.xls; for comparison see also uploaded file Supplemental_table_S9.2.1.xls and the main text of the paper). There are 231 novel PFAM domains and 747 different architectures that originated at the metazoan stem, followed by just 105 novel domains and 481 architectures at the eumetazoan stem.

S9.3.1 Domain evolution of death domain proteins

The death-fold domains exemplify the role of domain novelty and shuffling in metazoan evolution, particularly within the apoptotic signaling machinery. These modules permit distinct proteins to form complexes via homotypic interactions. Most members of the death-fold domain family (i.e. death domain, DED and CARD) appear to be animal novelties whereas the pyrin domain (PYD) arose later in vertebrates (viral PYD-containing proteins that allow pathogens to evade host defence are likely to be the result of a recent horizontal gene transfer). While it has recently been suggested that the origin of the death-fold domains could be even more ancient, this proposition is only substantiated by low confidence scores obtained from automated protein database searches that are likely to be false positives.[151]

The integration of the death-fold modules in various adaptors, such as the metazoan-specific FADD or the vertebrate-specific TRADD, permits signal transduction to occur between receptors (*e.g.* TNFRs) and effectors (*e.g.* caspases) and was probably an important step in the emergence of the apoptotic pathway. While homologous death-fold domain-containing proteins are present in metazoans, reshuffling has occurred repeatedly within the metazoan group, suggesting it had a crucial role in the subsequent generation of novel regulatory networks in specific lineages. For instance, while most animals are equipped with a FADD adaptor that consists of a DED associated with a death domain, the DED has been substituted to a CARD in *T. adhaerens* (see Table S8.3.1). The NLRs (which link innate immunity to cell-death signaling via caspase-dependent and -independent pathways) seem to rely on a single death-fold domain type for their protein-protein interactions in *Amphimedon* and *Nematostella*. While this is also the case in the bilaterians *Capitella* and *Strongylocentrotus,* the NLR family exploits the death domain, CARD or DED in the cephalochordate *Branchiostoma floridae*, and CARD or PYD in vertebrates (Fig 2a; Table S8.3.1). Most animals, with the exception of *T. adhaerens*, encode initiator caspases with either a CARD or two DEDs as pro-domains, but some bilaterians *(e.g. Danio rerio)* also encodes a PYD-containing caspase (Table S8.3.1). Finally, some lineages also possess an expanded repertoire of domain combinations for a given protein, that is otherwise structurally conserved in other metazoans (*e.g.*, APAF1 in *Nematostella* and *Strongylocentrotus*)[99,100] (Fig 2a; Table S8.3.1).

S9.3.2 Domain evolution of laminin proteins

Laminins are multidomain extracellular matrix proteins that polymerize to provide a major structural component of the bilaterian basal lamina (also referred to as basement membrane).[152] The individual units for polymerization are laminin heterotrimers, each formed through the combination of an α, β, and γ chain subunit. Although easily distinguishable from one another on the basis of domain composition and architecture, bilaterian α (α 1/2 and α 3/5), β and γ chains share a common overall organization, with a conserved Laminin N-terminal domain (LamNT), a series of repeating Laminin-type EGF (LamEGF) domains and a rod-like coiled coil region. These commonalities suggest that laminin genes evolved through duplication from a single precursor, with domain addition and domain shuffling giving rise to the distinct chain types. Laminin proteins and their component domains are not found in fungi or

plants, indicating that these events are likely to have occurred in the lineage leading to the Metazoa, concomitant with the elaboration of the extracellular matrix.[81]

A comparison of laminin genes from the genomes of *Monosiga*, *Amphimedon*, *Trichoplax* and *Nematostella* with those from model bilaterians (*Drosophila*, *C. elegans* and mammals), reveals key differences in the overall gene complement for each individual lineage, as well as differences in the domain architectures of individual genes. *Monosiga* possesses a single laminin-like gene with no globular domains (IVA, IVB or α3/5) interspersed within the LamEGF repeats, and a large region of coding sequence C-terminal to the putative coiled coil region (data not shown). *Amphimedon* has a gene similar to the *Monosiga* gene in its lack of globular domains, but with no additional coding sequence C-terminal to the coiled coil. Mammals are also known to possess a gene like this (β 3 - not shown), but considering that similar genes are not found in other bilaterians this laminin chain may represent the result of domain modifications occurring subsequent to the whole genome duplications at the base of the vertebrate lineage. In addition to the gene described above, *Amphimedon* possesses a single laminin gene with an architecture resembling the bilaterian γ chain. However unlike bilaterian laminin γ, this gene contains a short sequence which is similar in location and amino acid composition to the laminin β knob motif found only in the coiled coil region of bilaterian laminin β proteins (data not shown). The *Amphimedon* genome also contains an α 3/5-like gene that appears to be missing a IVA domain and contains only short, degenerate LamNT and α3/5 domains.
Lastly, *Amphimedon* possesses a gene with a combination of IVB and IVA globular domains that are not found in any of the laminin chains described in *Drosophila*, *C. elegans* and mammals. Interestingly, similar genes are detected in *Nematostella*, the sea urchin, *S. purpuratus*, and the annelid, *Capitella sp.I*, suggesting that this form represents an ancestral chain type that has been lost independently in several bilaterian lineages. *Nematostella* and *Trichoplax* both appear to possess a nearly complete complement of the four bilaterian laminin subunit types (α 1/2, α 3/5, β and γ) but each contains a different α subtype, providing another example of apparent whole gene loss.

Despite the differences between *Amphimedon* and bilaterian basal lamina forming laminins, it seems likely that *Amphimedon* laminin chains can form a similar heterotrimer structure (S9.3.1). This hypothetical molecule is likely to have the capacity to interact with cell-surface adhesion proteins through the Laminin G modules of the α 3/5-like chain. The significance of the lack of a well-conserved Laminin N-terminal domain in the same chain is not clear, but it may affect the ability of the molecule to polymerize into a network.[153]

Fig 2 Expanded legend

2A. A summary of domain architectures for putative death-fold domain containing proteins related to NLR, caspase and APAF1 found in *A. queenslandica*, *T. adhaerens*, *N. vectensis* and bilaterian genomes. The examples of bilaterian proteins provided do not necessarily occur across all the bilaterian lineages surveyed. We only report NLR proteins for which complete model predictions could be detected. However, the presence of fragmented *NRL*-like models on short genome contigs or on the edge of contigs suggests that additional domain configurations are

likely to be found in some lineages. For instance, the NLR model proposed for *N. vectensis* is a composite of two JGI protein prediction models. In *N. vectensis*, the APAF1 prediction models that display different numbers of CARDs are isoforms of the same protein.

2B. A summary of domain architectures for putative laminin related proteins found in *M. brevicollis*, *A. queenslandica*, *T. adhaerens*, *N. vectensis* and bilaterian genomes. Only genes containing typical laminin domains as well as a putative coiled coil region are included (excludes netrin, perlecan and usherin-like proteins). Mammalian laminins α 4, β 3 and γ 2, which are assumed to represent vertebrate-specific architecture modifications not found in other bilaterian laminins, are not depicted. For similar reasons, three unique or partial *S. purpuratus* laminin α genes are also not depicted. For multiple proteins sharing the same architecture, the order of domains remains constant while the number varies only for LamEGF domains (and LamG for *T. adhaerens* laminin α 1/2 which contains only three LamG domains). Domain diagrams are not drawn to scale but do reflect the locations of domains on the primary sequence. The laminin β-knob motif is not depicted. A degenerate LamG-like domain occurring at the beginning of the coiled coil region for *N. vectensis* laminin α 3/5 is also not depicted. Putative coiled coil regions were assigned by checking for sequence similarity with the corresponding regions of bilaterian laminin proteins. The following proteins are depicted – *D.melanogaster* laminin α 1/2 (wing blister), *N. vectensis* laminin α 3/5, *N. vectensis* laminin β, *A. queenslandica* laminin γ-like, *A. queenslandica* laminin β/γ-like1, *A. queenslandica* laminin α 3/5-like and *A. queenslandica* laminin-like. The length in amino acids for each of these proteins is displayed at the bottom of the figure.

S9.4 Molecular function enrichment of novelties

Panther annotations were obtained for members of novel gene families and the gene families inferred to be present in the most recent common ancestor of metazoans using Panther annotation pipeline.[76] Each term was mapped to a molecular function (MF) as provided by the Panther database. To allow for a higher resolution, all high-level (more general) molecular function categories were mapped to the next, more specific, category one level lower in the hierarchy. For example, the high level "MF00036 Transcription factor" category was subdivided into more specific categories (*e.g.*, "MF00038 Homeobox transcription factor") when such subcategorization was available. To maintain a comparable level of functional annotation, all categories below this level were mapped back to it. All mappings included removing redundant annotations. A table with each row for a distinct MF and columns containing counts of gene families in the novel set versus ancestor set was produced. Fisher's exact test as implemented in R[154] was run to test for enrichment or depletion of each MF category (contingency table contained counts for novel versus ancestor gene family sets and total count of MFs for novel versus ancestor sets). Multiple testing correction was done with the Bonferonni method (see uploaded spreadsheet Supplemental_table_S9.4.1.xls).

Metazoan novelties are particularly enriched in certain categories of transcription factors (*e.g.*, homeobox, bHLH, zinc finger), adhesion molecules (cadherin and CAMs), signaling (RTK, GPCR), as well as serine proteases and reverse transcriptases, the latter presumably because of

residual transposable elements in the gene set (see uploaded file Supplemental_table_S9.4.1.xls).

S10. Gene Family Expansion Analysis

S10.1 Analysis of expansion in eukaryotic families

In identifying sources of genomic novelty that may have lead to new functions in early animal evolution, it is important to consider not only genes or domains that are new by virtue of having no recognizable homologs in outgroup species, but to also consider genes that are novel by virtue of gene duplication, yet have homologs in outgroup species. Gene duplication followed by subfunctionalization is a recognized process for the evolution of new gene functions. We sought to identify gene families that have expanded at different nodes in animal evolution and this may have been significant in giving new functions to different animal lineages.

The first step was the development of an algorithm for reconciling gene trees with the species tree given a particular rooting. To overcome the difficulty in identifying the correct rooting of hundreds of gene trees, phylogenetic trees were simulated with different parameters of duplication and loss to determine the rooting most likely to be correct by rooting the simulated trees over all possible roots and determining gene duplication histories for each scenario. These simulations suggested that the rooting with the minimum count comes closest to the real count of duplications, and thus metazoan gene families were reconciled with the species tree using the rooting that gave the minimum counts for all the species tree nodes. These methods are described in detail below.

<u>Reconciliation of gene trees with the species tree</u>

A Perl script (subfamily_count.pl) was written to take as input a gene tree (with the names of the genes containing information on which species they are from) and a species tree, both in Nexus format. The script uses BioPerl modules Bio::NEXUS and Bio::Phylo to manipulate and traverse through the trees. The flow of the algorithm is as follows:

For a given rooting of the gene tree:

1. All gene tree nodes (internal and tip nodes) are mapped to species tree nodes based on which species the genes descending from a particular gene tree node belong to. For example, a gene tree node containing a sponge, a cnidarian, and a human gene will be mapped to the ancestral metazoan node on the species tree.
2. All gene tree nodes are classified as duplications or speciations based on which species the genes descending from a particular gene tree node belong to. For example, a node that gives rise to a sponge gene on one side, and a human and a cnidarian gene on the other side will be classified as a speciation as the branching pattern of the genes mirrors that of the species tree. However, if the gene tree node gives rise to a sponge and a cnidarian gene on one lineage, and a human gene on the other, it will be classified as a duplication

because the relationship of the species implies that there had been a duplication prior to the divergence of the three species and over time one duplicate was lost in the human lineage and the other in the sponge and cnidarian lineages.

3. For each species tree node, each gene tree node is evaluated for whether it can or cannot be counted as a subfamily that was present at the given species tree node. Thus, the number of subfamilies for each species tree node is inferred given this rooting of the genetree.

The criteria for counting the number of subfamilies at a given internal species tree node are:

1. If there are no gene tree nodes that map to this species tree node (or its ancestral nodes), this gene family was not present in this species tree node, hence the count of subfamilies for this node is **zero**.
2. If the only gene tree nodes that map to this species tree node (or its ancestral nodes) are speciations, then only **one** subfamily of these gene family was present in the ancestral organism that this species tree node represents.
3. If there are duplications on the gene tree that map to this species tree node (or its ancestral nodes), then a gene tree node is counted as a subfamily for this species tree node if:
    a. if the gene tree node has parent node that is a duplication (but not a polytomy) that maps to the species tree node (or its ancestral nodes) and none of the daughter nodes of the gene tree node are duplications that also map to the species tree node (or its ancestral nodes).
    b. if the gene tree node is a polytomy, then its daughter nodes are resolved in a way that minimizes the number of implied duplications. If the polytomy maps to the species tree node (or its ancestral nodes), each of its daughter nodes that appear to have arisen from a duplication that maps to the species tree node (or its ancestral node) are counted as subfamilies mapping to this species tree node (as long as none of the daughter nodes of the gene tree node are duplications that also map to the species tree node (or its ancestral nodes).

To put it simply, a duplication in a species tree ancestor leads to the creation of subfamilies, that are inherited by the descendant species of this ancestor. So, to count the number of subfamilies at any given species tree ancestor, we count all gene tree nodes that are daughters of a duplication that happened in this species tree ancestor, or in the ancestors of this ancestor. For example, the eumetazoan ancestor has all the subfamilies that were created by duplications in the metazoan ancestor, and also those that are new duplications in the eumetazoan ancestor itself.

<u>Simulation of gene family evolution</u>

A gene tree representing relationships of a family of genes from different species can be reconciled with the species tree to understand the patterns of duplication (expansion) and loss along different lineages on the species tree. However, the correct inference of expansion or loss relies on knowing the correct rooting of the tree - different rootings of the same tree will imply different histories of gene family evolution. For example, consider a tree with four genes (A, B, C, D) from two species (S1 and S2). Let A and B be genes in S1 and let C and D be genes in S2.

Let the topology of the tree for these four genes be ((A,C),(B,D)). If the true root of the tree is at the mid-point, then the history of the family suggests that the common ancestor of S1 and S2 already had two genes (i.e. the family had duplicated prior to the divergence of S1 and S2), that gave rise to A and B in S1 and C and D in S2 upon speciation. However, if the true root of the tree were at D, it would imply that the common ancestor of S1 and S2 had three genes (i.e. the family had expanded before the divergence of S1 and S2), one of which gave rise to D in S2 (but was lost in S1), one that gave rise to B in S1 (but was lost in S2), and one that gave rise to A in S1 and C in S2.

To determine the rooting that is most likely correct, simulated trees of different sizes were generated using different levels of duplication, loss of resolution and node loss rates were used. These simulated gene trees were then rooted at all possible nodes and subfamily counts obtained for all internal nodes of the species tree. The counts from the algorithm described above were then compared with the known counts (since the trees were generated and their histories recorded).

To generated gene trees, a Perl script (make_tree_prune_deresolve.pl) was written that uses the Bio::Phylo and Bio::NEXUS modules to make trees taking as input a species tree in Nexus format, and values for duplication rate (-r), rate of node loss or pruning (-p) and rate for losing resolution of a node or "deresolution" (-d). The algorithm for this method is as follows:

1. For each species tree node, the number of duplications is assigned using a poisson distribution - a random number is picked from a poisson distribution with mean lambda where lambda = (branch length * duplication rate r). The branch length of all species tree nodes is set at 1. "r" can take any positive integer value.
2. The species tree is traversed root to tip breadth-first and a gene tree is created according to the duplications generated by the random poisson process above.
3. All gene tree nodes are then considered for loss using a Bernoulli trial process using the probability specified with the -p option. The node and all its descendants are removed from the tree if the random number generated is less that p. "p" can take a value between 0 and 1.
4. Remaining gene tree nodes are then considered for loss of resolution using a Bernoulli trial process using the probability specified with the -p option. The node is removed and its children placed as daughters of its parent node if the random number generated is less that d. "d" can take a value between 0 and 1.

Different combinations of duplication, pruning and deresolution rates were used with r ranging from 0 to 10, and p and d ranging from 0 to 0.15 and a species tree with five species was used ((58:1.0,(34:1.0,(19:1.0,(11:1.0,24:1.0):1.0):1.0):1.0):1.0;). For all combinations, trees were generated in replicates of 50. All generated trees were then reconciled with the species tree for all possible rootings of the gene tree using the algorithm described above (count_multitree_multiroot_subfamilies3.pl). For each gene tree, the minimum, the maximum, the median and the mean counts over all possible rootings as well as the count for the assumed root of the gene tree were recorded for each species tree node. The differences between these counts and the real counts (known from the process of generating the tree) were compared. The minimum count came closest to the real count for gene trees created with a range of duplication, pruning and deresolution histories (Figure S10.1.1). Thus, it was determined that the rooting with

that gives the minimum count for real gene trees is not only the most conservative estimate, but it may also be a reasonable estimate for the correct root.

Analysis of expansion in clustered eukaryotic families

The 113,220 eukaryotic gene family clusters generated as described above (Section S9.1), were filtered to retain clusters with 20 to 200 genes from at least 5 animal species. This filtering was done to reduce the computational burden and to target clusters that are likely to carry expansions at nodes leading to animal evolution. The resulted 924 clusters were aligned using ClustalW and the alignments filtered using GBlocks (with options b3=15, b4=2, b5=a, e=.gb, p=t, -g). Only those alignments that resulted in more than 20 amino acids after GBlocks were retained for further analysis. These measures resulted in 725 gene families or clusters.

Neighbor-joining (NJ) and neighbor-joining with bootstrap (NJboot) trees were generated for the 725 gene families using Phylip. Each tree generated was then reconciled with the species tree ((Pt,Dd,At,(Nc,(Mb,(Aq,(Ta,((Nv,Hm),((Ce,Dm),(Hs,Sp)))))))));) to obtain the minimum counts of subfamilies for the species tree nodes (using the script get_multiroot_subfamilies.pl). A gene family was considered as expanded at a species tree stem if the the value (log(d)-log(a)) was greater than zero (d = subfamily count of the species tree node under consideration, a = subfamily count of the ancestor of the species tree node under consideration. For example, a gene family has expanded along the protometazoan stem if the value of log(m)-log(h) is greater than zero (m = subfamilies inferred for the metazoan ancestor, h = subfamilies inferred for the holozoan ancestor).

By this measure, 452 families appear as expanded in the protometazoan stem using the NJ trees, while 203 appear expanded at this stem by using the NJboot method (187 of the 725 gene families appear to have expanded at the protometazoan stem using both NJ and NJboot trees) (Figure S10.1.2). Though there is discrepancy in the numbers of subfamilies considered as expanded at various species tree nodes, there is a strong rank correlation (rho = 1, p-value = 0.001) between the numbers of families considered expanded at different nodes by the two methods. Thus, of the families considered, the largest number expanded at the protometazoan stem.

The 725 clusters selected for this analysis do not appear to be enriched for any functional GO categories relative to the set of eukaryotic clusters. The families expanded at the protometazoan stem by both the NJ and NJboot measures are not enriched for any GO terms relative to the 725 clusters.

S10.2 Linkage of expanded gene families

To address the question of whether the new duplicates created at different nodes in early animal evolution were the result of tandem gene duplications or segmental duplications, paralog pairs from all *Amphimedon*, *Trichoplax*, *Nematostella*, and human generated at different animal tree nodes were assessed for linkage (presence on the same scaffold/chromosome in these genomes).

Paralog pairs were determined using the NJ trees of the 725 gene families described above. The significance of the number of paralog pairs found to be linked was determined by generating 10,000 random datasets of the same size as the number of paralog pairs under consideration where random gene pairs in these genomes were tested for linkage. The p-value was defined as the number of datasets that showed at least the same count of linked paralog pairs as the real dataset divided by 10,000.

A significant fraction of paralog pairs generated at the protometazoan stem (up to 30%, as found in *Trichoplax*, $p<0.0001$) remain linked, indicating that (1) many gene family expansions originally occurred as tandem or proximal duplications, and (2) these genomically local duplications have remained linked over time (Table S10.2.1).

S11. Correlation of complexity with molecular functions

S11.1 Enrichment of molecular functions in complexity groups

Enrichment and depletions of molecular functions were tested with the same protocol as in S9.4. Total number of genes in each genome per Panther molecular function category were considered. We were interested in comparing different groups of morphological complexity: basal metazoans (represented by the non-bilaterian animals *Nematostella*, *Hydra*, *Trichoplax*, *Amphimedon*, with or without *Monosiga*); invertebrate bilaterians (*Drosophila*, *Caenorhabditis*, *Stongylocentrotus*); vertebrates (as represented in Fig. 5 by human, the best-annotated vertebrate); and non-animal outgroups (*Neurospora*, *Arabidopsis*, *Dictyostelium*, *Paramecium*, with or without *Monosiga*) (sheets one through four in Supplemental_table_S11.1.1.xls and Supplemental_table_S11.1.2.xls). We combined genomes into a few broad categories to reduce the impact of genome-specific expansions or depletions. Fisher's exact tests were run on several pairs of combinations for these groups.

By comparing pairs of different complexity groups we identifed molecular function categories that correlate with the differences in complexity. See uploaded spreadsheets Supplemental_table_S11.1.1.xls (enrichments) and Supplemental_table_S11.1.2.xls (depletions) which show a comparison between a pair of complexity groups on each sheet. For example, immunoglobulin receptor family members, immunoglobulins, MHC antigens, and cytokine receptors are enriched in human relative to invertebrate bilaterians. Relative to the "basal" metazoans, nuclear hormone receptors, homeobox, bHLH, and zinc finger transcription factors are enriched in other bilaterians group. In basal metazoans, relative to non-animals, we observed expansions of GPCRs, reverse transcriptases, the three groups of transcription factors, cell adhesion and cytoskeletal proteins. (Reverse transcriptase enrichment may be due to incomplete filtering of gene models for reverse transcriptases between different genomes.)

To visualise the functional expansions and depletions that potentially contributed to the increase in morphological complexity graphically, we took molecular function categories that were enriched/depleted in any of the four complexity groups. We limited the analysis to molecular function categories with p-value of equal to or lower than 1E-10 in Fisher's exact test for both

enrichment and depletion analyses. The molecular function categories were organized by decreasing level of significance of enrichment in vertebrates (relative to invertebrate bilaterians), then in invertebrates (relative to basal metazoans), then in metazoans (relative to non-animals), and then in basal metazoans+*Monosiga* (relative to the other outgroups). This order was used to generate a heatmap that shows counts of genes belonging to a particular Panther molecular function category. Counts were normalized to the total number of Panther annotated genes in each organism and each row was then normalized to the sum of squares (Figure S11.1.1). A subset of this heatmap that shows only the significant enrichments in vertebrates, invertebrate bilaterians, and basal metazoans is shown in Figure 5 of the main paper.

S11.2 Principal components analysis

To identify correlates of morphological complexity, molecular function counts (for all available PANTHER categories) were obtained for the four complexity groups. For this analysis we included additional species, specifically: non-animals (*Neurospora*, *Arabidopsis*, *Dictyostelium*, *Paramecium*, *Monosiga, S. cerevisiae*); basal metazoans (*Amphimedon*, *Nematostella, Hydra, Trichoplax*), invertebrate bilaterians (*Drosophila*, *Anopheles*, *Tribolium*, C*aenorhabditis*, *Strongylocentrotus*, *Ciona*, *Branchiostoma*), and vertebrates (human, chicken, frog, zebrafish, mouse, rat). The average gene count in each functional category was computed for each complexity group. Principal components were identified using prcomp function in R (Supplemental_table_S11.2.1.xls). Projection of the individual species onto the first two principal components is shown in Fig. S11.2.1. 51.4% of the variance is explained by the first axis and the subsequent axes explain 26.9%, 21.7%, 1E-14% of the variance respectively. We note that some molecular function categories contribute to both PC1 and PC2 (e.g., G protein coupled receptors, zinc finger transcription factors). Such function categories (1) discriminate between animal/non-animal, and also (2) appear to increase from basal metazoan to invertebrate to vertebrate.

Tables and Figures

Table S2.2.1: Libraries used in whole genome shotgun sequencing

library id	insert size (kb)	insert standard deviation (kbp)	passing paired reads	total sequence (Mbp)	mean JTRIM15 read length (bp)	assembled passing reads (percent of total)	clone coverage (Mbp)
BAYA	3.11	0.52	981,430	758	770	685,028 (70%)	1,526
BAYB	6.74	0.64	1,053,640	824	780	737,993 (70%)	3,551
BAYC	35.4	3.60	46,252	33.4	720	32,558 (70%)	819
BAYS	36.5	3.48	17,986	12.3	680	12,462 (69%)	328
BHGO	35.3	3.91	41,910	29.0	690	29,355 (70%)	740
BHUN	6.78	0.63	7,968	6.6	820	5,813 (73%)	27
BUUG	6.48	0.60	59,520	44.1	740	38,764 (65%)	193
Total			**2,208,706**	**1,707***	**773**	**1,541,973 (70%)**	**7,184***

*Assuming a genome size of ~190 Mbp , 1,707 Mb total coverage is ~9x, and 7,184 Mbp = total clone coverage ~38x (10x in fosmid-end pairs).

Table S2.3.1: Contig summary. The contig N50 number is indicated in bold lettering.

Number of Contigs	Size (base pairs)
28,143 contigs totaling	145,144,367 bp
15,397 contigs larger than 2 kb total	126,516,508 bp
5,952 contigs larger than 5 kb total	97,937,790 bp
2,652 contigs larger than 11.2kb total	74,015,413 bp
1,391 contigs larger than 20 kb total	55,360,054 bp
288 contigs larger than 50 kb total	21,823,814 bp
39 contigs larger than 100 kb total	5,184,926 bp

Table S2.3.2: Scaffold summary. The scaffold N50 number is indicated in bold lettering.

Number of Scaffolds	Size (base pairs)
13,522 scaffolds totaling	167,140,432 bp
6,480 scaffolds larger than 2 kb totaling	157,076,020 bp
3,492 scaffolds larger than 5 kb totaling	148,430,608 bp
2,002 scaffolds larger than 10kb totaling	137,115,837 bp
1,134 scaffolds larger than 20kb totaling	124,907,167 bp
647 scaffolds larger than 50kb totaling	109,921,829 bp
378 scaffolds larger than 100kb totaling	91,073,374 bp
310 scaffolds larger than 120kb totaling	83,571,006 bp
156 scaffolds larger than 200kb totaling	59,659,486 bp
27 scaffolds larger than 500kb totaling	20,975,995 bp
5 scaffolds larger than 1Mb totaling	6,492,255 bp

Table S2.3.3: Gap summary

Size	Number of gaps
>= 10 bp	14,621
>= 100 bp	11,143
>= 1 kbp	5,732
>= 5 kbp	479
>= 10 kbp	253

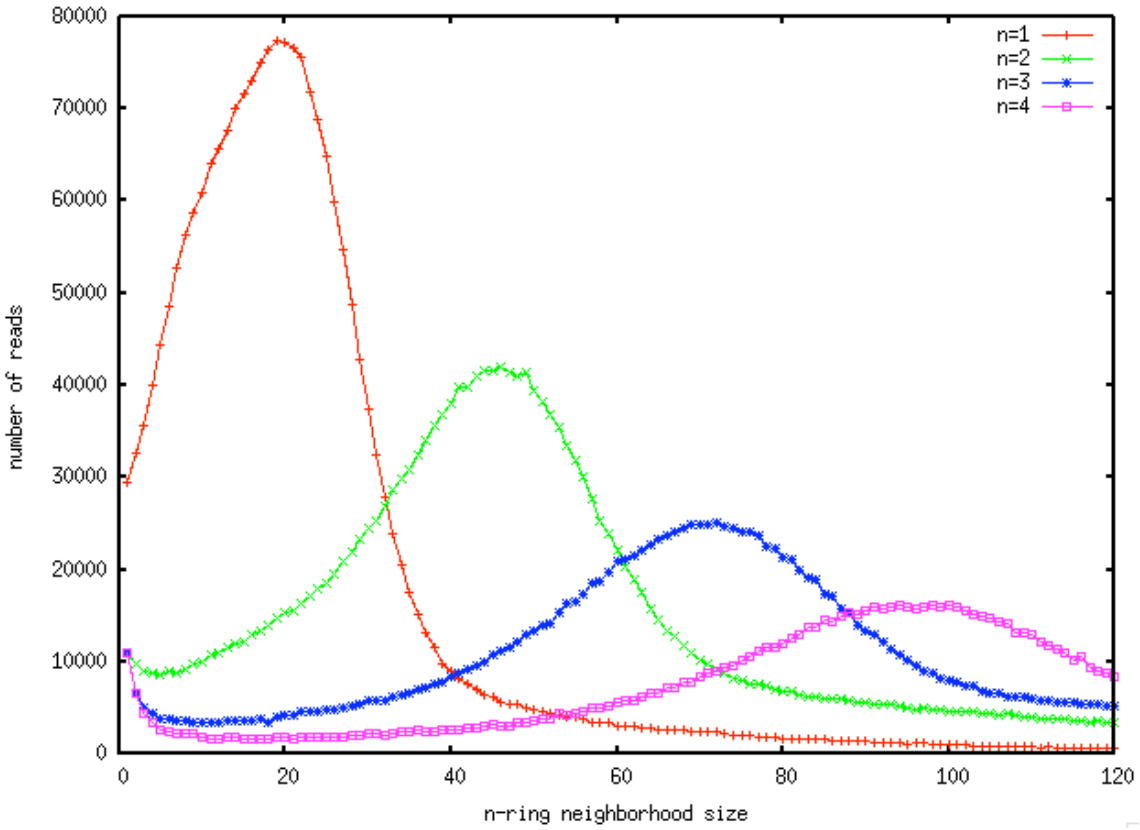
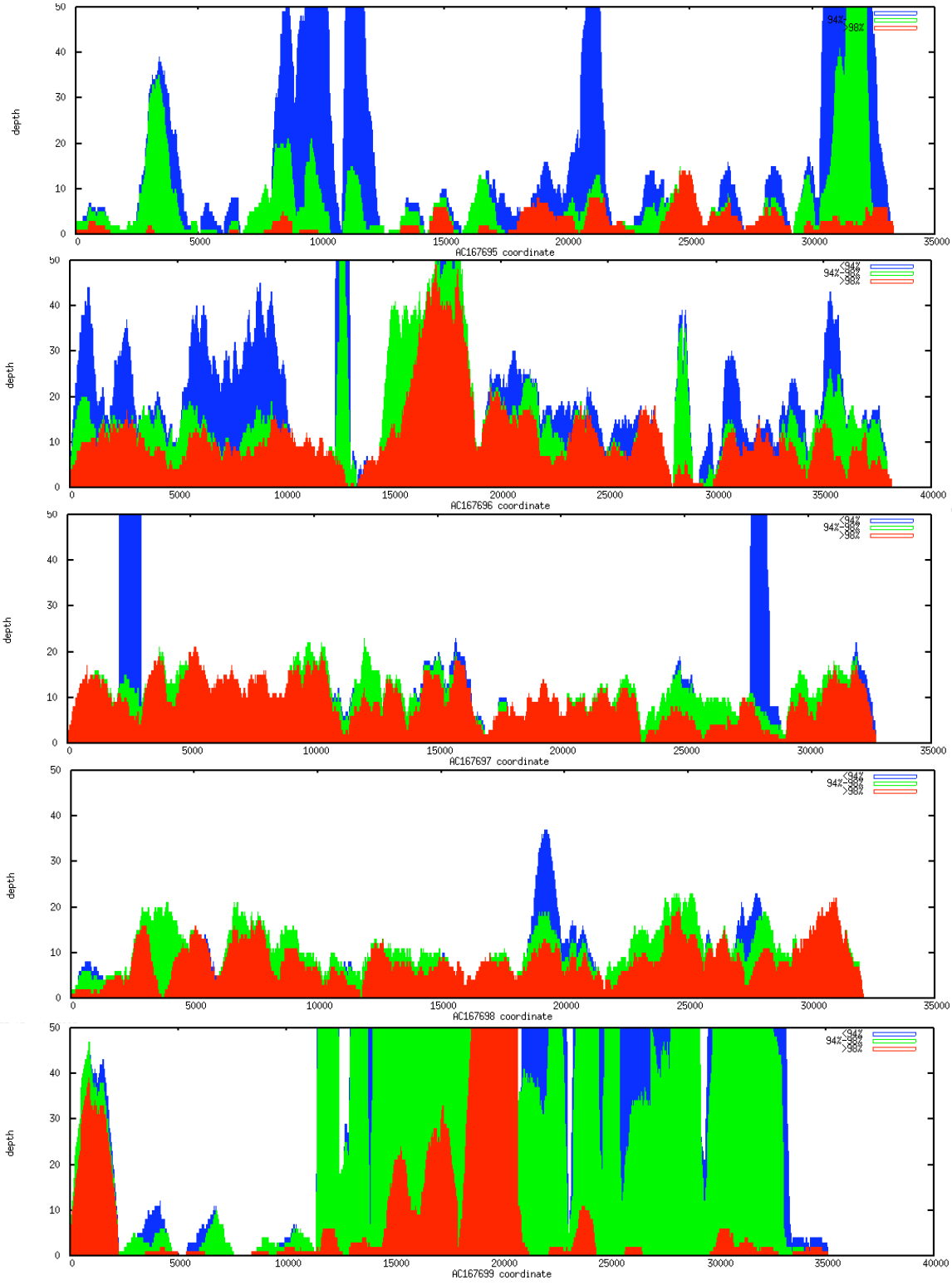
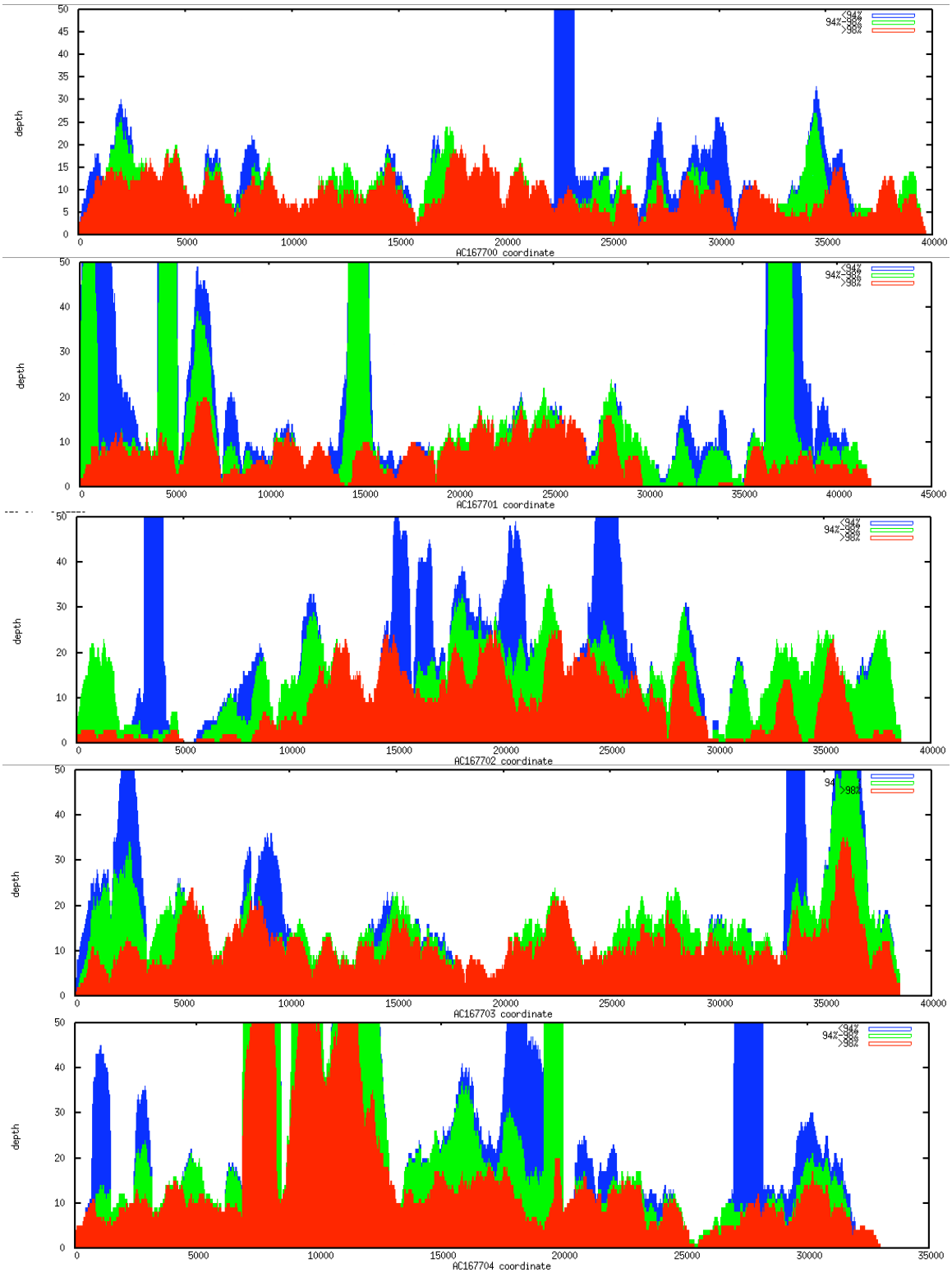


Figure S2.3.1: n-ring neighborhood size distributions

Table S2.4.1: Libraries used in EST sequencing

library ID	Source	Number of ESTs sequenced	Number passing quality and vector filters
CABF	released competent larvae	51,552	40,475
CAYH	competent larvae (one day old) (Larger inserts)	11,520	8,064
CAYI	competent larvae (one day old) (Smaller inserts)	19,968	17,836
Combined		83,040	66,375


depth
AC167695 coordinate
AC167696 coordinate
AC167697 coordinate
AC167698 coordinate
AC167699 coordinate
94%
>98%


depth
AC167700 coordinate
AC167701 coordinate
AC167702 coordinate
AC167703 coordinate
AC167704 coordinate
<94%
94%-98%
>98%

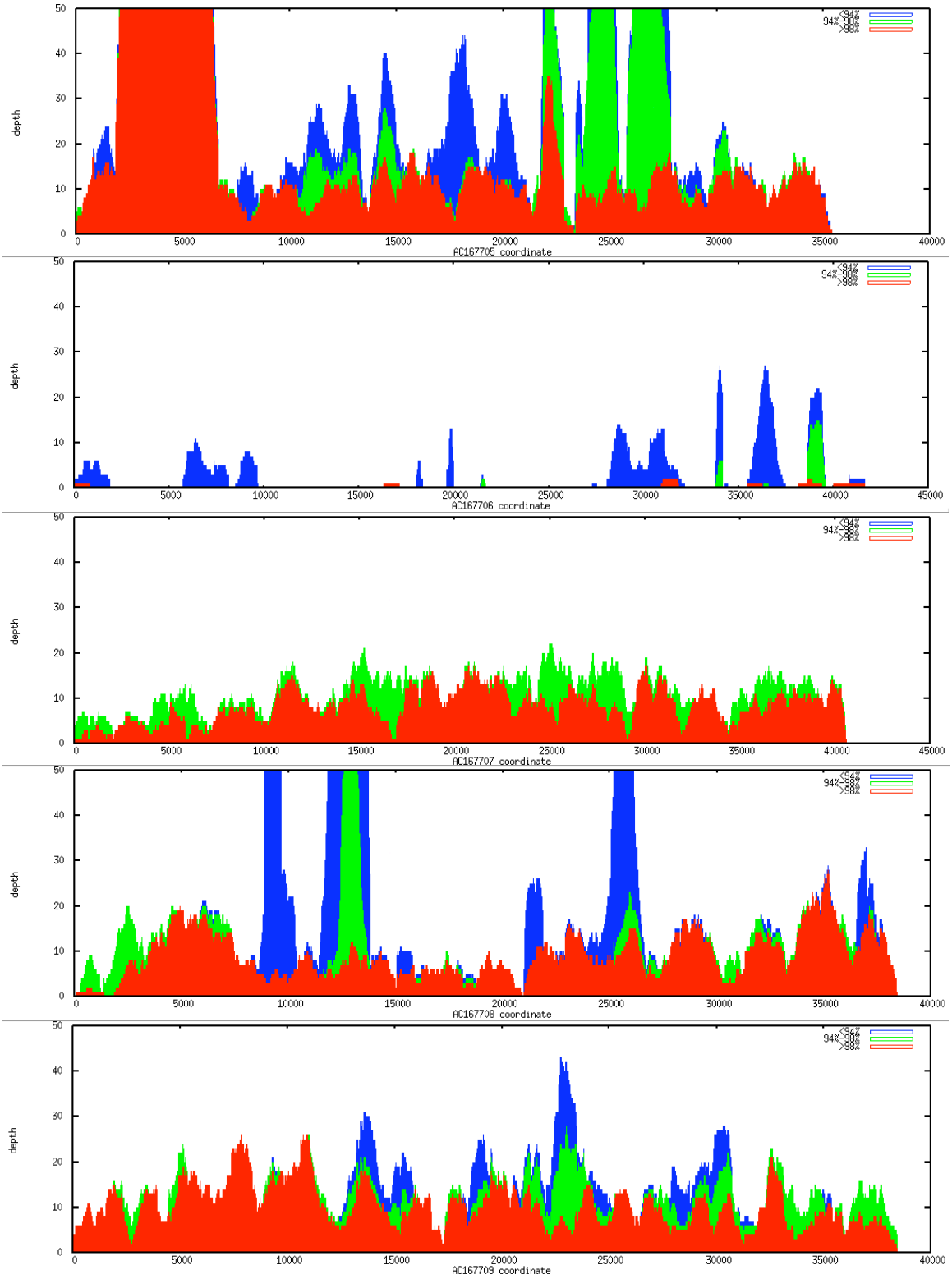


Figure 2.5.1. Coverage of fosmid sequences by shotgun reads. Whole genome shotgun reads were aligned to the fosmids by BLAST as described. Percent identity of read (over >95% of length) vs. fosmid is shown by color code: >98% identical (red), 94-98% identical (green), or <94% identical (blue).

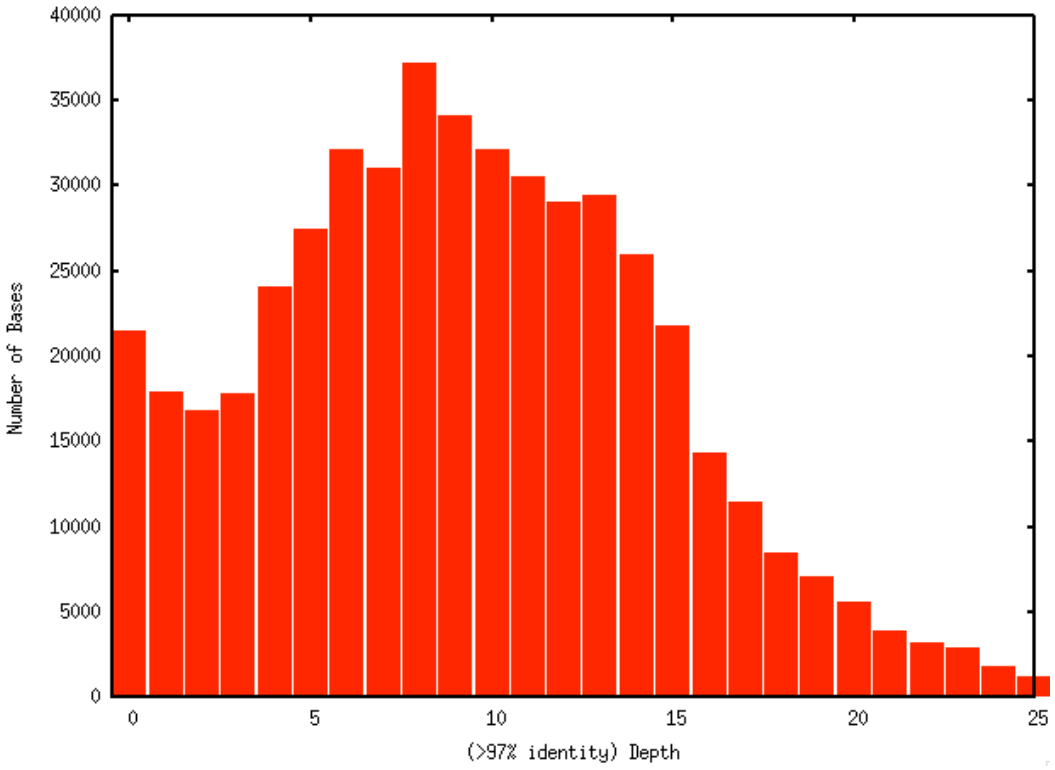


Figure. 2.5.2. Depth distribution vs. fosmids. Histogram of depth of coverage across fosmids (omitting AC167706) using only alignments that span at least 95% of the trimmed read length and at least 97% sequence identity.

Figure 2.5.3. Coverage of fosmid sequences by assembled scaffolds. Whole genome shotgun assembly aligned to fosmids.

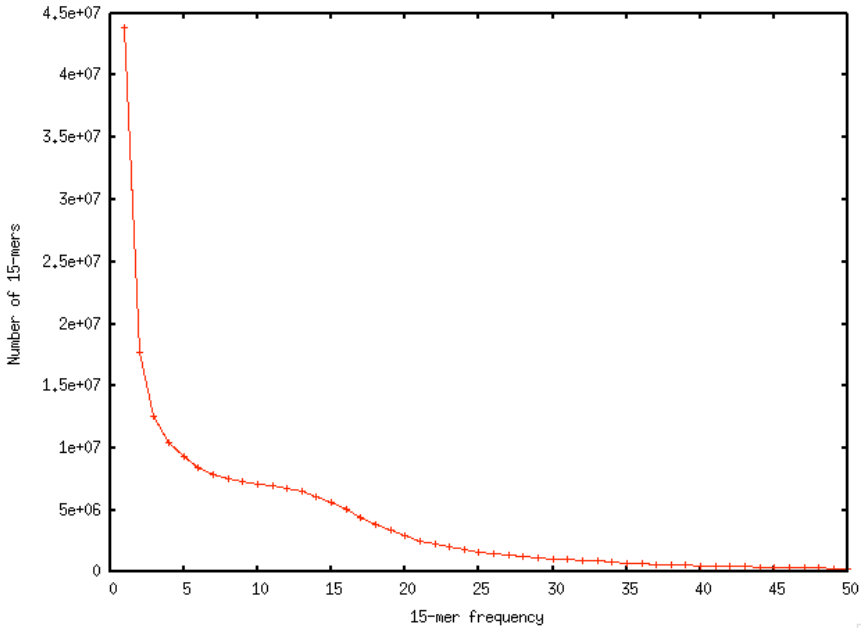


Figure S2.6.1: 15-mer frequency distribution

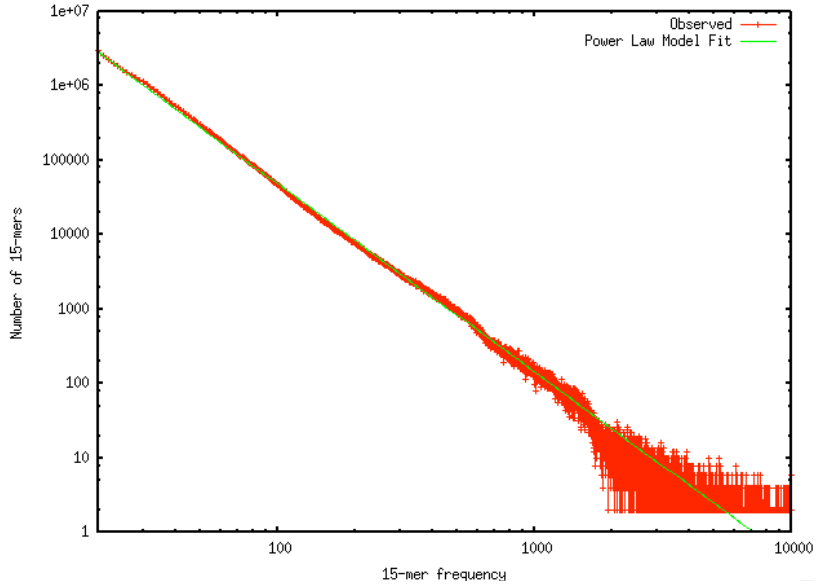
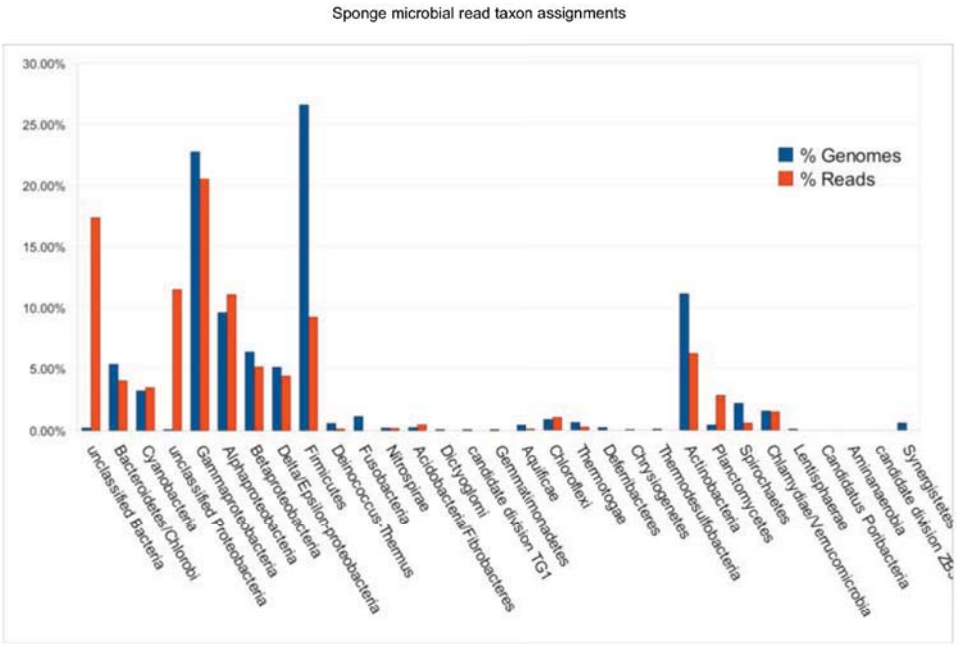


Figure S2.6.2: 15-mer frequency distribution for mers occurring more than 20 times in the dataset



Page 1

Figure S2.7.1: Frequency of sponge metagenome taxonomy assignments and depth of genomic sequencing for each clade. Shotgun sequence reads were assigned to taxonomic groups as described in the text. The fraction of all 7720 putatively bacterial reads assigned to a particular clade is shown in red. The depth of isolated genome sequencing for each clade is shown in blue as the fraction of genomes sequenced for that clade out of all bacterial genomes, taken from Genomes Online Database[155]. We find an abundance of reads putatively assigned to α- and γ-Proteobacteria.

A

1 >gnl|ti|858267137 name:BAYA14918.x1 mate:858267521
CCCCCNNNNNNNNNNGGGGGNNATAAGAGGCAGGCAGCTGCATTGCTGCAGGTCGATCTAGAGGATCCC
CGCCCTTCAGTTCATCGTCTTCTACGCAGGCCTACAAACCGTGCCCACCGAGGTCATGGAAGCGGCGATG
ATCGACGGCGCATCCAAGTGGGAGCGGATGTGGCACGTGGTGGTGCCCCACCTGATGCCATTGGTGATCT
TCGTCACCTTGATCCTGCTGATGGACAACTTTCGGGTCTTCGAGCCGATCGTCGGTTTTTCCGCCGAAGC
CCACGCCCGATCGCTCTCGTGGATCATCTTCAACGACCTGCGCGAAAGCGGCTCACCCCTCTACGGCTCC
GCCGGCGCTACGTCGATGATGACCATCCTCGGCGTGGCGGTGCTGCTGACGCCGGTGTTGATCCGCACTT
GGCGCGACTTCAACCGCAAGGCGCATTGAAAGATGCTCGGCACCCCGACCCGGCGTTTTCGATGGACTCC
CCTGACCATCATCTCGATGGCGTTGGTGGCTCTGTGGCTGATCATCGCCGCATTTCCTTTCCTGTGGACC
CTGTGGGGTTCTTTCAAAGTGCAAGGGGATTTCTTCTCCAAGGCCGACTGGATGAACGCCATCCACGGCG
TGCACACCATCCGAGAAACCGGCGGCGCATTCACCGATGACGGCTACTTCGGCGCTTGGGTGCAAGAGGA
GTTTTGGAGGAACGTCGTCAACACCACCATCGTGGTCTTCTTCACCGTCGTCATCTCGCTGACCATCGGC
ACCTTGGGGGGCTATGCCTTGGCACGTTCGGTTCATCGCTATGCCTTCTGGATCCTGATGGCCGCGCTGG
TATTTCGCGCCATGCCCCATATCACGCTGGTATCGGGCTATCTGCTGCCCTTTTTCGAATGGACTATCTG
GGGCATCCTGCCGACCACGATCATCGTGCTCGTGGCCATCAANCAACCCTTCACCTTGTGGATGCTTGAC
TCGTTTTTCCTGAACATTCCCCACGATATGGACNGN
gb|CP000830.1| Dinoroseobacter shibae DFL 12, complete genome 308 2e-140 1
gb|CP000739.1| Sinorhizobium medicae WSM419 plasmid pSMED01, ... 212 1e-128 1
gb|CP000362.1| Roseobacter denitrificans OCh 114, complete ge... 126 4e-52 1

2 >gnl|ti|858280025 name:BAYA20911.x1 mate:858277723
GCGGTAGAGGCAGGCAGCTGCATGCTGCAGGTCGACTCTAGAGGATCCCCATTGCTTTATCTTGGCATCG
TACTGATGGGCGATCTCGCGATCGGGCTGATCCGCAAGCAGGAGTTCCATCACCATCCGGCGCGAGAAGT
CGGCGCGCTCGCTCGCGGTGCGCACATCGAGACCTGAAGTCACGAGGCGCTGACAGGAGGCGGCGAGGGT
GCGCTCGCCCTCGATCTCGACCATGCAGGCGCGGCAGTTGCCGTCGGGGCGATAGCCGGGCTCGTCGGCA
TGGCAAAGATGAGGGATATCGGTGCCGAGCCGCTGCGCGACCTGCCAGATGGTCTCGTCGGGGCGGGCCT
CGACCGGCCGACCATCGAGGGTAAAACGGATCGTTTCGTTGATCGAGGGCATCTGCGGTTTTCCCGGTTC
TTTCGATTCTTAATCGGTTCTTGCAAGGTGTCCGTCATTCACGGATTCGTGATTTTCCGGCGTTTGCGAT
CCGGAGATCCGCCATCTGCATCCGGCATTCCAGCGGCAATCTCGCCTGCGATCTCAGCCTGCGTTCTCGG
GGGCGACCCCACCGAGCGAGGCTCGTTGCCGCAGCTAGCGCGAAAGCCCGGCCTCGTCTCGGAAATATCG
CAGCACGGCATTCAGCGGATTGGCCGCCGCCTGCCCGAGACCGCAAATCGAAGCGTCGCTCATCACCTGA
CCGAGTTCGCGCAGCAGCGCTTCGTCCCAATCGTCTTTGTCCATCAAACGCACCGNCTTGGCCGTTCCCA
CGCGGCAAGGGGTGCACTGCCCGCAACTCTCGTCTTCGAAAAAGCGCATCAGATTGACGGCCACGGNCTT
CATATCGTCCCGATCCGAGAGCACCGACACCGCATGGGAGCCGACGAGACAGCCATGCTTTTCCAGCGAG
CCGAAATCCAGCGGCAGATCCGCCATCGACGCCGGCCGGATGCCGGCCGAGGCCCGNCGGGCAAATAGCC
GTGCAGGCGATGACCCTCGAGCATGCCGCCGGCGAATTCGTCGACGAGTTCGCGCAAGGTGATCCCGGCG
GGGCCAACTTGACTCGGGGGATCGACCGGCCCGGAACCGATTAGTTCTCAAGGCCTTGCCGCCGTTACGC
CCGTGCTTGCGAACCCCTTGCGGCCCTTTTCGGCGATTTCGGGAACCCGTTACAGGGCCCACCTTTTGCG
gb|CP001350.1| Methylobacterium nodulans ORS 2060 plasmid pMN... 197 3e-99 1
gb|AF489516.1| Methylobacterium extorquens tungsten-containin... 191 1e-97 1
gb|CP001196.1| Oligotropha carboxidovorans OM5 strain OM5, co... 200 7e-96 1

3 >gnl|ti|858292540 name:BAYA28075.x1 mate:858292923
AAAAAAAAAAAAAAGAAAAAGAAAAAAAAAAAAAAAAAAAGCAAAANNNNATCTTTTTTTTCAGACGGCA
GGCCAGCTGCATGCCTGCAGGTCGATCTAGAGGATCCCGCATGTACTGCATTGCTCGAAGCGACTGTCCG
GCCGCTACGAAACGATCAATTGGTGGACCAAGGGCGACAAGTACACTTGGAACTTGGATCCGATTCGCGT
TCCGGCGAAATACCCGGGAAAGCGACACTTCAAAGGCCCCCGAGCAGGTCAGTTATCCGGCAACCCCAAT
GGGAAAAACCCGTCCGATGTCTGGGAATTTCCCAATGTGAAGAATAATCACCCGGAAAAGACCTCCCATC
CATGCCAATTTCCGGTCGAACTCGTCGAGCGACTGGTCCTTTCCATGACGAACGAAGGCGATTCGGTCTT
TGACCCTTATATCGGCGTCGGCTCGACTGCGATCGCAGCCCTGATGCATGGGCGCTCGGCTTATGGATGC
GATATCGAGCGCAAATATATCAATATCGCCTTGGATCGCATCGGCCTATTACAACTAGGTGAGTTGACGA
CACGGCCCATGGGCAAACCGATCTACGATCCAACCCGACCGCGCGGGGGACACGAGTGTAAACCGCCTCA
CGATATTCCCATCTCAACGGCGAGCAATACCTGAAAGTACATAAACCCGATCGGCTTGGAGATATTGAAA
GCGTCATCCAAAAAAGTCGATGCGCTTGCCTGCAAAACAAAAATATCGGATGAAAAACGAATGAACGGGA
AACGCTTCTTTTCACCTCTTGATATGAATCGGGCATTCAAGAATGACTTCGAAGCGCTGGATTGGAGAGC
GCGACGTCGTACCATTTGGGTTACTTCCGATCAGGCACTCCTACGTCGTATCTACCATTGCGAACCGGAT
GAGCAAAAATCGGCCATCGAGGAGGCCGGTCACATACCCATTCCATCGCGCAATGAGACCGACTTTGTGA
AGGATCGGGTTGCGATCGAAGTTCCAGTCGGACAATATGCCTTCGTGGCCACAGATCTATTCGTAAAGCA
CTTAAATTTCTACGGATCGGATATACTCCATGCCGTAATAAACTTTTGCCCATGAAACCATGGAGCAAGA
ATGTCATCAGGAGTCCCTAATACGAACGTGATTGCTGCATGTAACCAACAAGGCGNNNAG
gb|CP000356.1| Sphingopyxis alaskensis RB2256, complete genome 322 3e-123 1
gb|AY197779.1| Geobacillus stearothermophilus BstYI methyltra... 155 8e-107 1
gb|CP000490.1| Paracoccus denitrificans PD1222 chromosome 2, ... 300 2e-78 1

4 >gnl|ti|858302793 name:BAYA29047.y1 mate:858302410
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAACCCGGTTTTTTTTTGCT

ACTAGCTGGNTCCNNTACCAAAGAGATGTCCGTACGNAGAGACATGCCACTNGGGAAGGCGTACAGCCAA
CTTCTCAGTGGTATCTTGCTCTTGTGGAAAACGGTGCCATTTTTCACCGGAATGAATTTTCCGGCAACTC
TTGCACTGATAGTGCCCTTCTCGATGCTTCCAAATACCGTGGCGTCGGCTTTCGGGCTTCGGACAGGTAG
GTATCCCATCATCCCAAAGCCACTTTTCGATGAAGGCTCGAGCGGATGCCCCGCCTGAAAAACCGCGGCA
CGCTAAAGATGACAGACTTTTCTTTCAGGTCTTTTGCGGCATCGTCTCCGATTGGCTGATTTGAACAATC
GGCAGATTGAAATCCATTTCGAGTTGACTTGTATATATTTCTCAAAATAATCGTGCAGACCATATCGCAG
AAAAAAGAATTCGAAAATATTTGAGTCTTGAACGAGAACAGGGACATTAAAACTATATCTTTCACCTGCA
GCTACGGTGTGTATCCTCTCTGATAAAAAGAAGTCTTATGTATCATGTATCTGGTCTGCATAAAAGATCG
AGCCGTCAGATTCTATTGATAGAACTTGAGGAATGTGCTTTCTTATGTCATCGGTGATAGAAAAGTATCC
GATGGTCGATCTCTGTAATCTGCGGAGAATTTTTTCCTCGATATCCTAATCGGGGTGGTACCGAGGCGAT
CGAGGTTTGTATTCGCCCGCTCGATTTCGAAGTCATGGGGTCGGAATCGGCCACTCAAATCATATCGGGA
AAAATGGTGGTCGGTGTTTTAGTGTCTTCAATTGTTTGGATCCCTTCGACGACGTCTCGCTTGGACTGAC
TTGTTAGCGAGACTGTGAGTGCGCGCATCGTCAAGGGTAGGGGTATCGGGATTTCAACTTGGGGTTTCAA
AGGGATATCGGGCCCATCTCCTGCCATTGTCCGGCGGCGGATGATTCCAGGATCGCTTCGAGCAGGACTT
GGAATTCGTAGCCTTGGCGAATCCGAAGGGTCGGGCACGTCTTCGGCGATGGCCGTGGCCAAGGTTGCGA
ACTTCGATGGCTTTGAAGTCGTTGAAGCCGAATTGATGGCCGGGCGGCAAGCAGAAAACAACGAAGGGTG
GATGCTTCGGACCGTCAGGGNT
gb|CP000738.1| Sinorhizobium medicae WSM419, complete genome 59.3 4e-09 1
emb|AL591688.1| Sinorhizobium meliloti 1021 complete chromosome 55.6 2e-08 1
gb|CP000758.1| Ochrobactrum anthropi ATCC 49188 chromosome 1,... 52.4 8e-08 1

5 >gnl|ti|858312037 name:BAYA36779.y1 mate:858315098
AAGAGAAAAAAAAAGAAGAAAGGAAAAGAAAAAAAGAGGGAAAAAAAAAAGCNNNNNNCCGCGCCGCGCC
GAACTACCTGCTCCGAAGAGGATGGCAAGACACAAGCTGCTGCTGCCGTGGCGCGGCAAGGCGCTGCTGC
TGCACGCCGTCGACGCCCTGCTCGAAGCCAAGCAGCAGGGCCTGGTCGCCAAGGTGACGGTGGTGACCGG
CCATGAGCGCGCCAAGGTCGCGCGGCTGCTGCGCGGCCGCGACGTGACCGCCGTCCACAATCCGGCCCAC
GCCACCGGCATGGCCTCCTCGCTGAAGGCCGGGGTAGCAAGCCTCGGCAGCGACGCCGGCGGCGCGCTGG
TCTGCCTCGGCGACATGCCGGGCCTGGACCCTGCCCTGCTGCAGAAGATGGCGGCGGCCTTCGGCGCCGA
CGACGGCAAGGACATCGTGATCCCGTCCTGCCAGGGCAAGTACGGCCACCCGAAGATCTTCGGCGCCAGC
TATTTCGACGAGATCCTGCGCCTTGAAGGCGACGTCGGCGCCCAGGCGGTCATAGGCAGGAACAAGCAAA
GCGTGGCGCTGGTGCCGGCGGGCAAGGAGGTTCTGTTCGACATCGACACGCCGGCGCAGCGCAAAGAATG
ATGCTGAAAAATCCAGCATAAACAAAGAACTGCCCTTTATTCTTTCTCAAGCTTGAAGGCTACGTCGCCG
GCTTGGTTCGCGTCCACGTAGATCTGGTAGCCCGCCCGCCTGGCGTTCAGCGCTAACTGGTACAAGCCGA
TGCCCATCCCCGCGCTCGACGGCAGCGGGGCGCTGAACATCCGCTTGTGGATCTCCTCGGGGATCTCGCT
GCCGTTGTCCCGCACGGTCAGCGCCGGCCCCTCCGCCAGCTCGACCTGGACCTGGATCTCGCCCTCGGNC
CCCTGCTTGCGCAGGGCGTTGTGCGTCAGGTTCTCCAGCGCTCCGTCGAAGACCTCCGNCGGGACGTCGG
GGCCGCCGTCGCCCAGCTCGACGCCGNNNC
gb|CP001096.1| Rhodopseudomonas palustris TIE-1, complete genome 152 7e-34 1
emb|BX572605.1| Rhodopseudomonas palustris CGA009 complete ge... 151 9e-34 1
gb|CP000283.1| Rhodopseudomonas palustris BisB5, complete genome 147 2e-32 1

6 >gnl|ti|858318460 name:BAYA35926.y1 mate:858317311
ACCNNNNNNNNNNNGGNGGGTTTNGTTGAAATAGCTTGTACCTTCGGAACATGGTGCGATTCGTTTCCG
AAAGAAGCACGAAGTCACGCGAAATCGTTATAGGCAATGAGGAAACTGAACAAGGCGGTGGTGATCACCC
CGTGGCCACATCACCGGGATGATGACATGACGAAAGGCCTGGAATTGGGTGCAGCCGTCGACTTTGGCGC
TCTCGTCGAGTTCTTTGGGGATGTTTTGGAAAAAACTGTGCAGCATCCAGAGGGTGAAGGGCTGGTTGAT
GGCGACCAGCACCACGATGGTCGTCGGCAGGATGCCCCAAAGATTCCACTCGAAAAAGGGCAGCATATAG
CCGGAGACCAGAGTGATCGGGGGCATGGCACGGAAGATCAGGGCCAGAATCAGCAAGAAAAAACCGTAGC
GGGCGCGGGCCCGGGAGAGGGCGTAGCCGCCCAGAGTGCCGATGGTCAAGGAGATGCAGACCACGAAAAA
ACAGACGATGGCGGTGTTGATGGCCGGGCGCCAAAAGCCTTCCGAAATCCAAGCCCCGTAATAACCGGCG
CCGGTGAAAGGGCGGCCGTGGGTCTCGGTGGTATTGCGCCCGAGCAGAGCGTTGGTCCAATCCGCCAAGG
AGAAAAAATCGAGTTCCACTTTGAACGAACCCCAAAAGGTCCATAAAAAGGGGAAGGCGGCCACGATCAG
CCATGCGATCAGAAAGAGGCTGCTTAGCAAACGCAGCGCCGGCGGCATTTTTCGGATCATCGATACGAGC
CCCGGTATTCCCGCCAAGTGCGCACCAAGACCGGCGAGAGCAGGATCGCCACTCCGATGATGGTCAGCAC
CGAGGTCGTGGCCGCCGCCGAGAGCTGCCTCGTTTCGCCGGCCAGATCGTTGACGATGATCCAGCTCAAT
GAAGTGGCATGGGCCGAGGCGTTGAAGCCGACGATGGGCTCGAAGACGCGGAAGTTATCCATCAGCTGCA
TCACGGCCACGAACGTGACCAAGGCGCTCCGATGCGGATCCCGACATAGCGCACCTGCTGCCAACGGGTG
GCGCCGTCGATGCGCGCAAGCTCGAGTTGATCCGAGGGCCAGGGTTGCCATCCCGGGGTAGAAGACGAAC
GAAGGCCAAAGGGGCCGGATGCCAAACCCCGTGGATTATCACACAGGCCAGTCAGCTCGATCGAGGCCTT
GACCAAAGGTN
gb|CP000830.1| Dinoroseobacter shibae DFL 12, complete genome 324 1e-122 1
gb|CP000739.1| Sinorhizobium medicae WSM419 plasmid pSMED01, ... 309 2e-115 1
gb|CP000264.1| Jannaschia sp. CCS1, complete genome 111 9e-49 1

7 >gnl|ti|858319749 name:BAYA34143.y1 mate:858319365
ACCATCGCATTCAAGCGGAGTGTCGAGGAGGAATCCGTTTCGACTCGTCCCTCGGGGGGCGAGAAAAAAC
CCGGCGAGATCAAGCGCCGGCGAGCGAAGGGTTTGTCGATCGCTCCGCCCATCGAAACCGAAAATCCCCG
ATACTTCGATGGAAAGCGAATGGATCGGGCTTGGATTGCAGGTCTCGAGTCGAATGCTTCGATGCGATCG
GATAGGCTCTTCGGGTCTTGACTTCGCTTTCATCGGAAAAGACAATCGTCCGCAATCCTCGTCATCTGTT
GAATCTGTTGAAAGCGATTGCTGCACCACTCGATCATGACAGGTATCGCCGGTATCGACAAGCCGCTCGG

ATATCGATAGCGCTCTATAAGACAACATTATTATCGGGATTATCGGGCGCGCCGCCCTTGTCGTATCAGG
TGCCCACTCCAAGCGGGAGGCCAAGGCATCGATCGCATCTCGCGCTACCCGGCGTCTATCGTGCTCATTC
GCCAGACACTTTCCTCCGTCGCCTCGCTGCTGCTCTCCTACGGGCTGTTGCTGGTCGCCAACGGTTTGTT
CGGAACCCTGTTCGGCTTGCGCGCCAAACTCGAGGGTTTCCCGACCCTGCTGGTGGGGTTGATCGTCAGC
GCCTATTTCGTCGGGATGTTCGCAGGCGGAATCTGGGCGGTGCAGGTGGTTGCCAGGGGCGGGCATATTC
GGGCTTTTGCCGCCTTCGCCTCGCTGATGTCGGTGACCGNCCTCGGGATCGTTCTCGTGATCGATCCCTT
GCTTTGGATGGTGATGCGCTTCGTCGGCGGTTTTTGCCTGGCGGGCATGATCATGGTCACCGAGAGTTGG
CTCAACGAGCGGACCGTCAACGCCTCTCGGGGGCAGGTGCTGTCTTTTTAC
gb|CP000453.1| Alkalilimnicola ehrlichei MLHE-1, complete genome 119 2e-28 1
gb|CP000356.1| Sphingopyxis alaskensis RB2256, complete genome 129 1e-27 1
gb|CP000830.1| Dinoroseobacter shibae DFL 12, complete genome 129 4e-27 1

8 >gnl|ti|858325212 name:BAYA45383.x1 mate:858328655
CAAGGCAGCTTGATGCTGCAGGTCAGACTCTAGAGGATCCCCCCTTCTGGTCTCGACCTTCACCTATACG
GTCTACGGCAACCACGCCCGCCCCGAATGGTCGAGCGATGCCGACTGGCGCAAGCGGTGGAAGCGGCAAT
CCTCGGCTTGGGGGGCGAATGCGCATAACCCCGGAGACCATCGGGAGTACGGGCTTTCCACCTATAACAC
CCATAGCGATGGCAGCGGAATAGCGATCGCTTCCTGGCATCGCCCGATGCTGAACCTGCGCATCGGCTAT
ATCACCTATCCCGATCCCGAAATTCGCGGCTCGGGGATGCGTCATTTCCCCGCCGATACCCACCTGATCG
CGTGGTTGGAGGCCAAGGGCATCGCCTACGATTTGATCAGCGATCAGGAGCTGCACGATGAAGGCGTGGA
ACTGTTGGAAGGCTATCGGACGCTGATGACCGGCTCTCATCCCGAATACCACACCCCGAAGACTCTGGAT
GCGATCGAGGCTTGGAGGGATCGAGGCGGGCGGTTGTGCTATCTCGGCGGGAACGGTTTTTATTGGAAGA
TCGCCCTTTCGCCGGAAAAGGAAGGGGTGATCGAGATTCGTCGGGGAGAGGGCGGTATCCGCGCATGGGC
GGCGGAAGCGGGTGAGTACTACAACCAATTCGATGGGGAATACGGTGGCTTGTGGCGGCGCAACGGCCGT
CCACCGCAGAATCTTTGCGGGGTCGGCTTCACCGCGCAGGGAAATTACGCAGGCTCCTATTATCGAAAGC
GTAGCGAGGCTTGCGATCCCCGAGTGGCGTGGATCTTCGAGGGTATCGAGGGGGGATATTNTCGGCGACC
ACGGTTTGCGAGGGCACGGGGCGGCGGGTTTCGAGCTGGATCGGGCCGACAAGCGACTCGGCACGCCGGC
GCATGCGCTGATCGTCGTTGCCTCGGAAAACCATTCCGCCGGATACGCCTTGGGTCCTCGTGCCCGAGGA
GCAGTTTGACGCATATCGTCCTTTGGCCCGGTGAGCCCTATCGGGAATTGATCCGGGCCGATATGCTTCC
TTCCGAACCCGGGCGGGCGGTTCGGTTACCCGGGCAGCGTCATCACTTTTTGCGGGACGCTTGCGAGCGA
GGGGTTCGGTACCCATTTTTCGGGTTGGTTGAAACCTA
dbj|BA000040.2| Bradyrhizobium japonicum USDA 110 DNA, comple... 179 6e-72 1
emb|BX640445.1| Bordetella bronchiseptica strain RB50, comple... 122 3e-67 1
emb|BX640431.1| Bordetella parapertussis strain 12822, comple... 122 3e-67 1

9 >gnl|ti|858327842 name:BAYA44569.x1 mate:858334743
AAAAAAAAAAAAAAAAAGAAAAAGAGAAAGAAAAAAAGAGGAAAAGAAAGAAAGAAAAAGAAAAAAGAGG
GNNNNNNNNCCCGCTCCTCAGACGGCAGGCAGCTGCATGCTGCAGGCNACTCTAGAGGATCCCCGNCGC
TGGCTGATGGTCGAACGCGATACGCTGACCCATCNGTATCGTGCCACGGCGCCTCCGGGGAAGGAGCCGG
GCGAGCCGATTGCCGCCGCTTCGGATCCGGGGCTTGTGCGCCATCGCCGCAAGCTGCCCAAGCTCGGTGC
CGAGGCCGCGGATATCAAAGCCGCCTGCGGAAAGATAAGCGAAGCGGATTCGCCTTCGACGAGCGCCGAG
GCGAACGAGGAATCGGGCGCGGTGCGAAGGCCATCCGAGTCCCTCGAAGTATCCGAATCCCCCGAATCCC
CCGAATCCCCCGCACCATCATTGCCCGCCCTCGAACAGCCGCGCCGTACCGCCCAAGGAGGTCGGATCGA
TCGCTCGGCTCCCTTGCGCTTTACCTTCGACGGCATCGAATACGGCGGCTATCAAGGAGATACCCTGGCC
TCGGCGTTGCTGGCCAACGGGGTGCGGCGGGTCGGGCGCAGTTTCAAATACCACCGTCCGCGCGGAATCC
TCGGCATCGGCGCCGAGGAGCCCAATGCGCTGGTGCGATTGGGGGAAGGGGCGTACGCCGCCCCCAACCA
CAAAGCCACCGAAGTCGAACTCTTCGATGGCTTGGTCGCCCACAGCCAGAACCGGTGGCCCTCGAGGGAT
TTCGATATCGGGGTTCTGGCCGATTTCGCCTCGCGCCTGCTCCCGGCGGGCTTTTACTACAAGACCTTTA
TGTGGCCGGGCTCGTGGTGGCGCTTTTACGAGCGCTTCATCCGCAAGGCGGCGGGTTGGGGCGATCGGC
GAGGGCTGCGGATCCGGACGCCTATGACCATCGCCATGCCTTTTGCGATGTGCTGGTGGTCGGGGCCGGC
CCGGCGGGCTTGATGGCGGCGCAAGCGGCGGCCGAGTCCGGGGCGAGAACGATCCTGATCGACGATGCCG
TCGAGCCCGGCGGGGAGCTGTCGCCACCGACCGAGCGATCGCCATCGACGGACTTCCCCCCCCGTTTCTTC
GGCCGAAAGAACGCTCGAAACCCTTCGNNNNGG
dbj|AP009384.1| Azorhizobium caulinodans ORS 571 DNA, complet... 215 2e-66 1
emb|AL646052.1| Ralstonia solanacearum GMI1000 chromosome com... 239 2e-63 1
gb|CP001298.1| Methylobacterium chloromethanicum CM4, complet... 197 2e-63 1

10 >gnl|ti|858329158 name:BAYA47039.x1 mate:858336443
NNNCCCGCGTTNTGAAGCAGCACTGTGCTGGGTACTCTAAGATCCCAGGGTAATATTGATCGCGTATCGC
AGCGAATTCCATCCCAGATAATGCTCTGCGATATCGCCGGTGGTGATGGTATAGCCGCCGTCTTTCCGGC
GGCGCAAGGCAAGACCGGGTGCGAAAAGATTGGGAAGCCCTCGATCCGGCGCCGGTCCAGTGCGCCCGAC
CGAGGATCGCACCCCCAACTGCGGGAGGTCGAGGCCGAGATTGCGGGCGAAAATACTCGTCCATACCCCG
GCGGTGAGGGCCACGCGTTCGCAGCGGATACGGCCGCGCTCGGTATGAAGGGCGCGGACCTCGCCCTTTT
CGCTTTCGATGGTGCGCGCCGCGCAGCCCTCGGTGATGCGGACTCCGAGCCGTTGCGCAGCGCGCGCCAG
CGCCGGAACGGCGACGAAGGGCTCGGCCCGGCCATCGCTTTTGGTGAGCAGCCCACCGACCCAAGGGCCC
TCGATCCCAGGGGCGGCGGCTTTGGCTTGTTCGGCCGAGAGCATTTCGGTATCGAGCTGGTATTCCTTGG
CCCGCTCCCGCCATGCTTCGTGGCCGGCGAGGCTTTGGGGGTCATCCGAAAGATAAAGGCAGCCCGAAGC
GGTGAAGGTCAGATCGCGCTCCCCGGTCTCGCGCGCCAGACCCTCCCAGATCCTTCTCGATTCCATCGCG
ATCGGGATCTCCGCCGGATCCCGTCCTTGCTGGCGGATCCAACCCCAGTTGCGGCTCGATTGCTCACCCG
CGATCCGACCCTTTTCGACGAGGGCAACCCGAAGCCCCCTTCGGGCGAGAAACCATGCGGTGGAGACCCC
GACGATCCCGCCCCCGATGACCGTGACATCGACCGCCTCGGGCAAGGGATCGTCGAAGCGGATCGAGGAG

AGATTCGGGATCGAAGGCGTAAGGGACATGGTCGTTGCCGGGGCTGCTCGATCGCCGATACTTAAGCGAT
CTCGCTCTCGCCTTTGCGGGCGGATCGAAAACCCTTGCTAAACGGGAGCCGTTTGCTGCATCGATGGACG
GGCGGCGAAGTCCTCGTACCAACGGGCAATGCCGGGGCGGGAATCGCGCAAGTTCTCTTCGGAATAGGAA
AATCGAGATAGCGGAGGCGGCGGGCACAGTGACCAAGCCGATATCGGGTTCGAATCAGGGCATCGGAAGG
GATTGATCGATTCAGCCCGGGGGTAACTGCCCAGTTGN
gb|CP000661.1| Rhodobacter sphaeroides ATCC 17025, complete g... 158 1e-73 1
gb|CP000143.1| Rhodobacter sphaeroides 2.4.1 chromosome 1, co... 158 4e-73 1
gb|CP001150.1| Rhodobacter sphaeroides KD131 chromosome 1, co... 156 1e-72 1

B

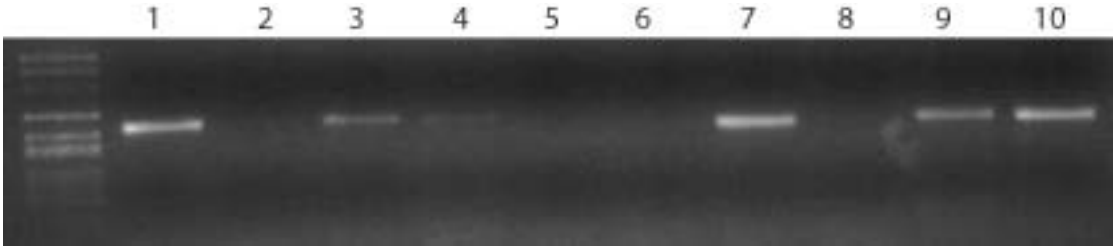


Fig. S2.7.2 (A) Putative bacterial traces assessed for presence in *Amphimedon* larvae by PCR. Most significant Blast hits are listed below the trace sequences. Oligonucleotide primer sequences are highlighted in yellow. (B) PCR amplicons of 10 primer sets corresponding to the 10 sequences shown in A.

Table S2.8.1: Mean values for all dinucleotides.

AA	0.98
AC	0.99
AG	1.09
AT	0.97
CA	1.27
CC	1.00
CG	0.36
CT	1.09
GA	0.98
GC	1.08
GG	1.00
GT	0.99
TA	0.87
TC	0.98
TG	1.27
TT	0.98

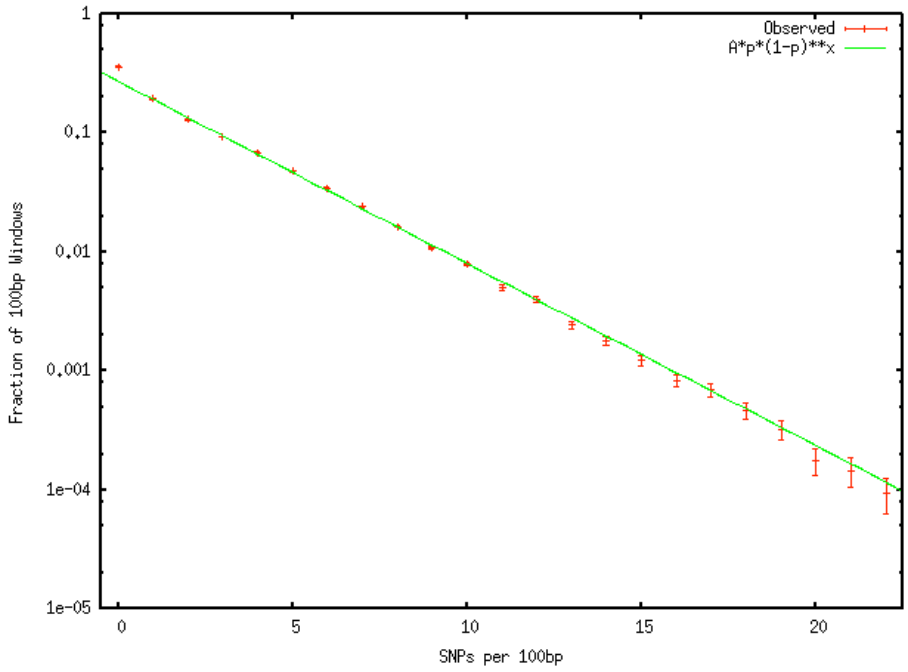


Figure S3.1: SNP distribution in 100bp windows.

Table S4.1: Summary statistics for gene models.

	Median	**Mode**
Peptide Length	280aa	118
exons	5	1
Intron Length	80	50
Gene span	1365 bp	384

Table S4.2: Support for *Amphimedon queenslandica* gene models

Support	**Number of gene models**
EST	6724
PFAM domain	19469
Human hit (1e-5 or better)	12759
Domain or EST	21980
human hit or EST	15756
Domain or human hit or EST	23560
SwissProt hit	18643
SwissProt hit or EST	21341
SwissProt hit or EST or Domain	23887
SwissProt hit or EST or Domain or human hit	24743
Total	29867

Table S6.1: p-value grid for synteny between *Amphimedon* scaffolds and ancestral linkage groups.

	p19_pal_15	p19_pal_11	p19_pal_2	p19_pal_10	p19_pal_9	p19_pal_14	p19_pal_1	p19_pal_4	p19_pal_13	p19_pal_8	p19_pal_7	p19_pal_3	p19_pal_5	p19_pal_12	p19_pal_6
Contig13436	32	2	1	2	2	2	2		2	3	1	2	3	2	4
Contig13482	26	2	2	6	12	8	3	2	4	3	5	6	4	4	2
Contig13470	20	3	1	3	6	4	2		2	3	1	1	3		1
Contig13315	16		2	3	3		1		3	1	2	4	1	1	1
Contig13289	10					1	2	1		1	1				
Contig13329	10	1		2		1				1			1	1	1
Contig13307	8	1	1	1	1	1	1	2		1	2	1	1	2	1
Contig13161	7	2	2	1		1	2			1	1		1	1	1
Contig13514	40	22	5	17	7	8	20	6	12	9	10	6	3	10	11
Contig13520	31	23	4	10	10	12	21	4	10	5	8	6	3	11	7
Contig13508	20	25	6	7	7	7	7	3	4	7	2	3	7	5	12
Contig13373	2	12	1	3	4	1	3		1	4	3		1	1	2
Contig13519	12	8	80	12	9	8	9	4	13	6	14	5	10	7	20
Contig13490	4		25	3	3	2	3		4	1	3	1	3	4	4
Contig13465	10	4	23	1	4	6	7	1	5	2	1	1	5	3	3
Contig13392	2	1	18	5	2	5	1	2	2	1	1	1		2	1
Contig13335	4		18	2	1	1	2		3			1		2	2
Contig13463	3	2	12				2	1	4	1	9		2	4	6
Contig13217	1		7	1		3		1	1			1	1		2
Contig13500	6	8	3	48	13	9	12	2	5	5	5	1	6	6	3
Contig13471	3	1		32	4	5	2		1	2	3	2	4		3
Contig13447	5	2	2	17	3	1	2		1		2	1	2	2	2
Contig13372	1	2	1	14	1	3	1		2	1	2	3		1	
Contig13437	2	2	2	13	4	3	2	2	1	1	3	2	1	1	3
Contig13192	2	3	1	11	2		1	1	3		2	1	1	1	3
Contig13281	3	2	1	10	1	1		1			3	1	3	3	1
Contig13313	1	1		10	2	1	1			1		1	2	1	1
Contig13320	1	1		9		1	1		1			1		1	
Contig13506	8	7	4	10	50	6	7	2	4	1	7	4	9	6	4
Contig13513	10	2	7	7	27	7	4	3	8	3	9	4	8	7	4
Contig13487	8	2	1	3	20	1	4	1	2	2	3	5	2	2	4
Contig13456	7	1	4	1	14	2	2		3	3	3	3			5
Contig13516	16	9	11	11	10	54	9	7	8	3	8	4	10	9	4
Contig13481	4	4	1	5	1	23	2	2	3	2	3		4	1	2
Contig13504	5	2	5	4	3	16	4	2		2	4	1	2	4	4
Contig13348		1		1		11			2	1			1		1
Contig13509	8	4	2	7	8	8	31	3	10	8	11	2	10	7	7
Contig13448	7	8	5	5	2	4	16		5	2	4		2	3	1
Contig13512	3	2	1	1	1	2	12				2		2		2
Contig13384	1	2		1	1		11	1				1			
Contig13316	1	5	2	2	2	4	11	3	1		1			1	3
Contig13345			1				8	1		1			1		
Contig13157	1	1				1	7		2						1
Contig13445	2	2		4		1	2	16	1	1	2	1	2	3	3
Contig13409	5	2	2	3	4	4	3	14	3	4	1	4	1	1	1
Contig13374	2	3	2	3	2	5	2	14	2			1	2	2	5
Contig13479	5		4	3	3	3	4	10	3	1	3	2	3	3	6
Contig13308	2	1	2	1	1	1	1	8	1	1			1	1	
Contig13355	2	2		1		4	1	7	1	1			1		3
Contig13484	8	4	5	3	4	6	5	1	18	4	11	2	7	3	4
Contig13467	3	1	2	3	2	3	4		12	3	4	3	3	2	2
Contig13347	2				2	2	1	2	11	2	1	2		2	3
Contig13251	3			1	2	2	1		11	2	4	2	1		3
Contig13441	7			2	5	4	2	2	11	8	1	3	1	3	5
Contig13501	8	3	2	5	4	5	4	4	8	28	1	1	4	1	4
Contig13433	6	4		4	3	4	3	1	1	12	5			4	1

Contig13429	4	4		4	1	2	4	1		11	1	1	1	2	2
Contig13435	6				3	1	2	1		8	2		2	3	5
Contig13401	3	2	1	4	4	2	1	3	3	1	18	1	2	2	4
Contig13111				3	2	1	2	1	1	1	10	1		2	1
Contig12971	4	2					1		2	1	8		2		
Contig13511	15	5	7	8	15	11	6	2	7	6	6	14	7	5	11
Contig13478	4	6	3	2	3	4	6	4	1	5	3		28	11	5
Contig13457	5		1	4	4	3	1		4	2	3	1	23	3	3
Contig13492	7	5	1	9	4	7	6	4	3	5	4	3	21	11	7
Contig13416	5	3		3	2	3	1		2	1	2	1	16	1	3
Contig13356	2	1		3	1	1		1	3		3	1	14	3	1
Contig13338	3		1	3	1	1	1		3	2	1	1	11		2
Contig13248	5	1	1	2		4	2	1		1	3	1	10	3	
Contig13420	1		1	6		5	3	1	1	2	3		10	1	2
Contig13365	1	1		1		1				1			8		
Contig13507	7	3	4	5	2	7	3		5	4	6		26	27	6
Contig13489	11	3	5	4	9	5	7	3	4	5	7	5	17	31	6
Contig13453	5	6	3	5	8	10	4	2	4	4	2	2	4	20	7
Contig13485	6	2	2	3	4	5	2	2	4	2	8		7	16	2
Contig13260		1	1	1	1		1				2		2	7	2
Contig13521	16	10	8	19	22	23	16	5	15	14	13	8	25	29	30
Contig13468	6	1	2	7	3	7	2	2	3	5	6	1	4	1	18
Contig13434	7	3	5	5	5	6	4	3	4	4	6	1	8	6	16
Contig13499	3		3	2	5	8	1	4	7	2	3	2	2	3	14
Contig13314	2	1		7	3	1	3	2	2	1	2		4	1	12
Contig13380	2	2	1	1	2	1			1	1	1		2	1	10
Contig13510	7	7	5	16	2	7	5	4	4	7	3	1	7	3	26

Multiple test correction = 1/2448

p <= 0.05

p <= 0.09

p <= 0.50

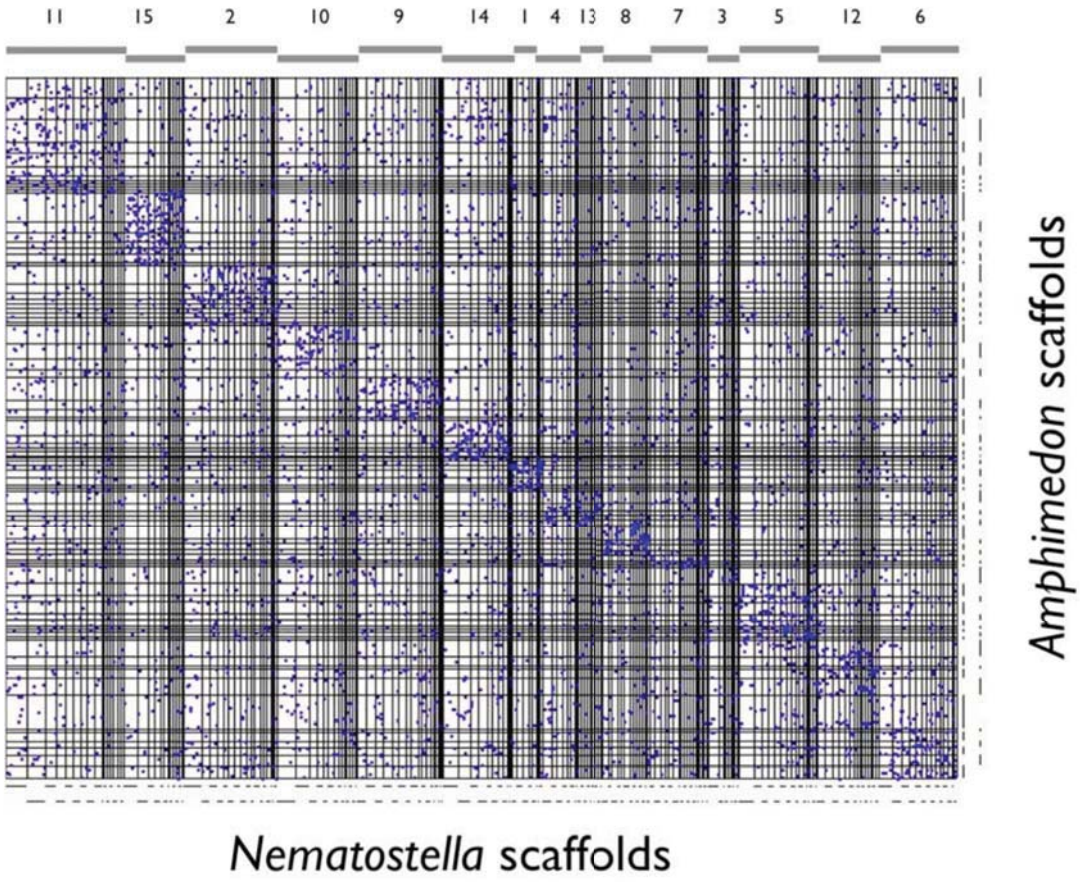


Figure S6.1: Dot-plot of orthologous genes between *Nematostella* and *Amphimedon* scaffolds. Blue dots represent the genomic locations (number of genes from the end of the scaffold) of orthologous genes in the *Nematostella vectensis* (horizontal coordinate) and *Amphimedon queenslandica* genomes (vertical coordinate). Horizontal and vertical lines mark the boundaries of draft genome scaffolds of the *Amphimedon* and *Nematostella* genome assemblies respectively. *Amphimedon* scaffolds are ordered as in Table S6.1, and *Nematostella* scaffolds are ordered to group scaffolds assigned to the same ancestral cnidarian-bilaterian ancestral chromosome [16]. Numbers and alternating horizontal gray bars indicate the partitioning of *Nematostella scaffolds* into ancestral linkage groups.

Table S6.2: Examples of genes shared between syntenic blocks in *Amphimedon* and chordates.

Chordate linkage group	Amphimedon scaffold	Number of shared named genes	Names of well-known genes
cho1	Contig13513	6	BTRC | CTBP1 | DOCK1 | DUSP1 | PDCD4 | WFS1 |
cho10	Contig13489	6	GABBR1 | LHX3 | NOTCH1 | NR5A1 | PPARD | RXRA |
cho10	Contig13521	6	GABBR1 | POU2F1 | POU5F1 | PSMB8 | PSMB9 | TRAF1 |
cho11	Contig13421	5	CTSH | CTSS | CYP1A1 | CYP4F2 | PTBP1 |
cho11	Contig13511	5	ANXA2 | CA9 | HINT1 | MCL1 | TP53BP1 |
cho11	Contig13514	5	CYP1A1 | CYP4F2 | NEDD4 | RECK | UPF1 |
cho11	Contig13519	7	ACO1 | CTSH | CTSS | GNAQ | IREB2 | NEDD4 | PRPF3 |
cho12	Contig13313	4	AIP | AIPL1 | RPA1 | STX1A |
cho12	Contig13447	7	BLMH | C1QBP | EWSR1 | PPP1CA | PPP4C | YWHAE | YWHAG |
cho12	Contig13500	9	ADRBK2 | CDC45L | DYNLL1 | MLXIPL | NF2 | ORAI1 | P2RX1 | P2RX7 | PRODH |
cho13	Contig13514	6	ACSL4 | CUL3 | CUL4A | CYP3A4 | LAMP1 | LIG4 |
cho13	Contig13520	6	ABCC4 | ABCC5 | CUL3 | CUL4A | HDAC4 | TPT1 |
cho15	Contig13248	4	GLO1 | PSMA6 | SNAP23 | SNAP25 |
cho15	Contig13438	5	CBL | FLI1 | GSTZ1 | SIRT2 | XRCC1 |
cho15	Contig13489	5	BRF1 | ESR2 | PRDX5 | SLC12A6 | SLC3A1 |
cho15	Contig13507	7	ERCC1 | GSTZ1 | KAT5 | MTA1 | MUS81 | POMT2 | PSMA6 |
cho15	Contig13521	8	ATP1A2 | BMP4 | H2AFX | PLCB1 | POLH | TGFB3 | TJP1 | ZW10 |
cho16	Contig13289	4	COL1A1 | DBI | HUS1 | TIMELESS |
cho16	Contig13470	6	ESPL1 | HEXIM1 | STAT1 | STAT2 | STAT5B | STAT6 |
cho16	Contig13482	6	ATF2 | BIN1 | ETV1 | GLB1 | SEMA3A | TFCP2 |
cho16	Contig13508	5	GLB1 | ITGA3 | ITGA4 | ITGB3 | KAT2A |
cho16	Contig13520	6	ABCB1 | ABCB11 | BARD1 | CREB1 | ITGB3 | TUBA1A |
cho17	Contig13315	4	ADARB1 | CTSC | RPGR | TCP1 |
cho17	Contig13501	10	ACAT1 | ACAT2 | CUL5 | FZD4 | PICALM | ROS1 | RPS6KA3 | TCP1 | TNFAIP3 | U2AF1 |
cho17	Contig13511	6	CHMP2B | IL18R1 | IL18RAP | IL1R1 | IL1RL1 | RPGR |
cho17	Contig13514	5	ACAT1 | ACAT2 | CUL5 | GAB2 | MYO6 |
cho17	Contig13520	6	CTSC | CUL5 | HDAC1 | INPPL1 | MYO6 | TCP1 |
cho2	Contig13516	5	FTH1 | FTL | HSD3B1 | HSP90B1 | TDG |
cho3	Contig13514	5	ADH1B | CDH23 | CYP4V2 | EIF4E | PRKG1 |
cho4	Contig13504	5	ALDH1L1 | ATXN7 | COL7A1 | GNL3 | TRO |
cho4	Contig13514	6	CAV1 | CAV2 | CUL1 | CYP27B1 | TBXAS1 | WNT2 |
cho4	Contig13516	10	ALAS2 | APPL1 | BCAP31 | CERK | HUWE1 | IKBKE | LTA4H | MCM10 | MCM2 | PTPN22 |
cho4	Contig13520	5	CHIT1 | CUL1 | HDAC6 | IKBKE | SLC26A4 |
cho5	Contig13490	10	AXIN1 | AXIN2 | CBY1 | GNA12 | HGS | KIAA1303 | MYH1 | MYH10 | PDIA2 | SEPT9 |
cho5	Contig13519	17	ABCC1 | ABCC6 | ACTB | GNA12 | MAD1L1 | MAP3K7IP1 | MKL1 | MPG | MYH1 | MYH10 | NUDT1 | PDIA2 | PRKCA | SPHK1 | TK1 | TSC2 | UNC13D |
cho5	Contig13520	9	ABCC1 | ABCC6 | CARD11 | DNASE1 | ITGB4 | MYH1 | MYH10 | PLA2G6 | SFRS2 |
cho7	Contig13511	6	ANXA6 | CAMK2A | KDR | PPP2R2B | TLR1 | TLR2 |
cho9	Contig13360	5	CETP | CNDP1 | LBP | PLTP | SLC7A9 |
cho9	Contig13519	5	GNAS | LMAN1 | MC1R | RALBP1 | WWP1 |
cho9	Contig13521	9	BRD7 | CDT1 | CEBPB | CEBPE | FKBP1A | GNAS | NP | PLCG1 | WWOX |

Table S7.1: Datasets generated using the four taxon kernel (FTK) and filtered mutual best hit (fMBH)

methods.

	FTK method			fMBH method		
Number of species allowed to be missing	None	One	Two	One	Two	Three
Number of orthologous gene clusters matching criteria	38	118	229	25	112	242
Number of genes with alignments after GBlocks	38	116	226	23	108	237
Number of amino acids in alignment	8,191	24,520	44,616	4,339	20,099	44,707
Name of dataset	FTK small	FTK medium	FTK large	fMBH small	fMBH medium	fMBH large
Number of clusters with *Oscarella* genes	-	-	64	-	-	53
Number of clusters with *Mnemiopsis* genes	-	-	46	-	-	48

Table S7.2: Topologies tested with fMBH and FTK datasets including nematodes

Method	Dataset	# genes	# amino acids	Topology with nematodes (a)	Topology without nematodes (b)	Can other topologies be rejected?
Filtered mutual best hits (fMBH) method	small	25	4,339	*Trichoplax*=cnidarian; nematodes=early animal	*Trichoplax*= eumetazoan	No
	medium	112	20,099	*Trichoplax*=cnidarian; nematodes=early animal	*Trichoplax*= eumetazoan	(a) cannot reject *Trichoplax*=eumetazoan; (b) rejects all other topologies
	large	242	44,707	*Trichoplax*=eumetazoan; nematodes=protostomes	*Trichoplax*= eumetazoan	Both reject all other topologies
Four-taxon kernel (FTK) method	small	38	8,191	*Trichoplax*=early animal; nematodes=protostomes	*Trichoplax*= early animal	No
	medium	118	24,520	*Trichoplax*=eumetazoan; nematodes=protostome	*Trichoplax*= eumetazoan;	(a) cannot reject other topologies; (b) rejects all other topologies
	large	229	44,616	*Trichoplax*=eumetazoan; nematodes=protostomes	*Trichoplax*= eumetazoan; urchin=early bilaterian	(a) cannot reject *Trichoplax* as cnidarian; (b) rejects all other topologies

All topologies placing sponges, placozoans and cnidarians in a clade sister to bilaterians are rejected by all datasets (12 datasets described above).

Table S7.3: Maximum likelihood bootstrap support values and Bayesian posterior probabilities for clades of the tree in Figure 2.1 using various methods. ML = maximum likelihood bootstrap values, PP = posterior probabilities from Bayesian inference analyses.

Clade	FTK large dataset				fMBH large dataset			
	ML	ML (no nematodes)	PP	PP (aamodel)	ML	ML (no nematodes)	PP	PP (aamodel)
Opisthokont	99	100	100	100	93	100	100	100
Holozoa	100	100	100	100	100	100	100	100
Metazoa	100	100	100	100	100	100	100	100
***Trichoplax* + Eumetazoa**	25	99	100	95	71	98	100	100
Eumetazoa	25	92	100	96	73	100	100	100
Cnidaria	100	100	100	100	100	100	100	100
Bilateria	7	100	100	100	79	100	100	100
Protostomia	25	100	100	92	77	100	100	100
Ecdysozoa	27	NA	100	100	79	NA	100	100
Lophotrochozoa	100	100	100	100	100	100	100	100
Deuterostomia	88	*	100	100	100	100	100	100

*In the FTK topology without nematodes, *Strongylocentrotus* appears as a long branch at the base of the Bilateria, thus deuterostomes are not recovered as a monophyletic group.

Table S7.4: Summary of statistical tests done to determine plausible positions for *Oscarella* and *Mnemiopsis*. Values in parentheses are for the datasets without nematodes (only indicated if different from the dataset that included nematodes).

***Oscarella* (X) genes added to 64 of the 229 genes of FTK method**	***Oscarella* (X) genes added to 53 of the 242 genes of fMBH method**	**Placement of *Oscarella*/*Mnemiopsis* (X)**	***Mnemiopsis* (X) genes added to 48 of the 242 genes of fMBH method**	***Mnemiopsis* (X) genes added to 46 of the 229 genes of FTK method**
not rejected	rejected	X as the earliest animal branch	best	best
rejected	not rejected (rejected)	X as sister to cnidarians	not rejected (rejected)	rejected
rejected	best	X as branch after *Trichoplax* but before cnidarians	not rejected (rejected)	rejected
not rejected	not rejected	X as sister to *Trichoplax*	not rejected	rejected
not rejected	not rejected	X as branch after *Amphimedon* but before *Trichoplax*	not rejected	rejected
best	not rejected	X as sister to *Amphimedon*	not rejected	not rejected
rejected	rejected	X as sister to bilaterians	rejected	rejected

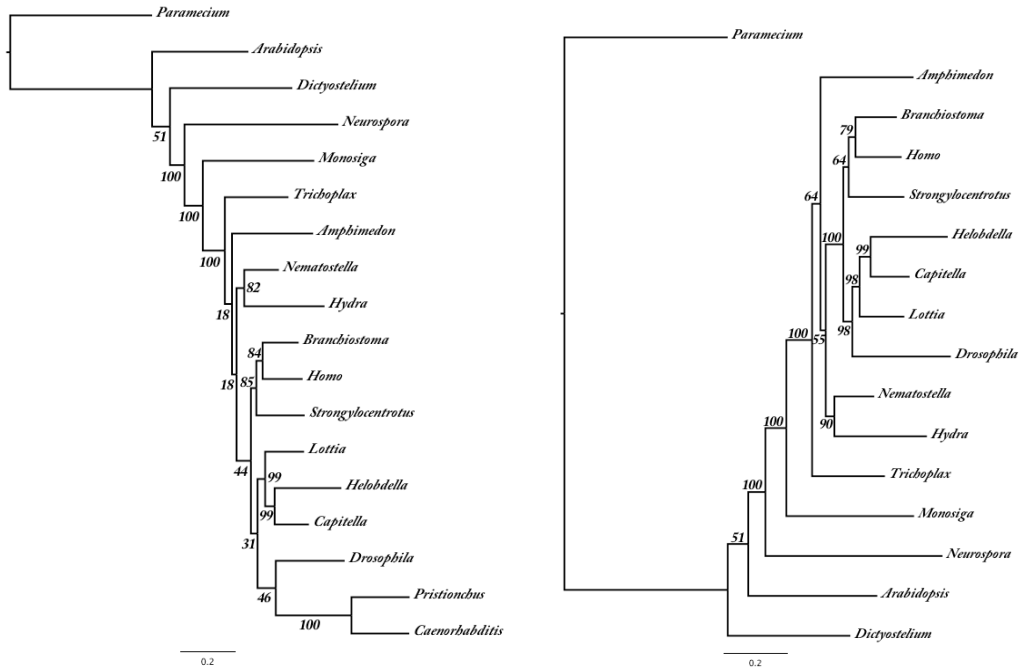


Figure S7.1: Maximum likelihood topology obtained for small FTK dataset including (left) and excluding (right) nematodes.

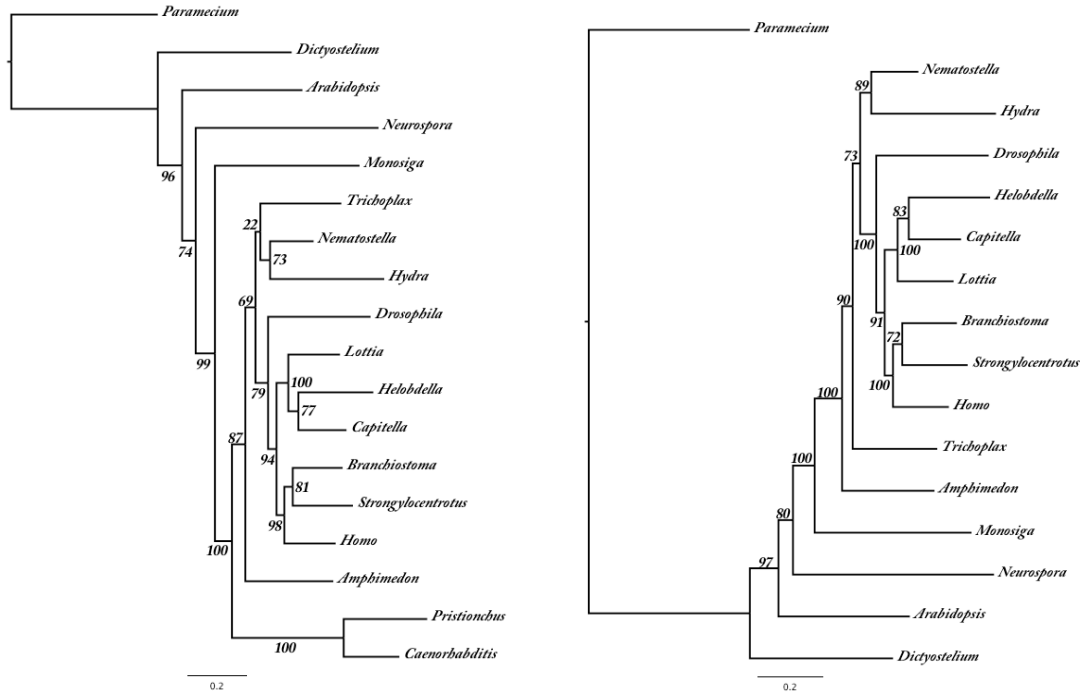


Figure S7.2: Maximum likelihood topology obtained for small fMBH dataset including (left) and excluding (right) nematodes.

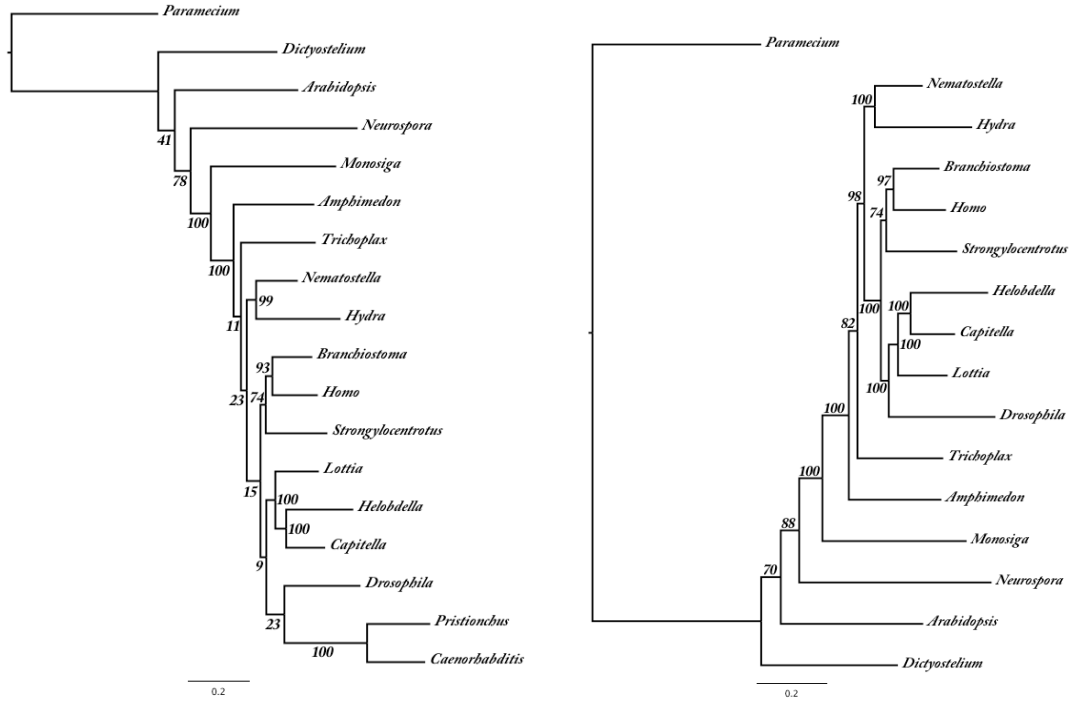


Figure S7.3: Maximum likelihood topology obtained for medium FTK dataset including (left) and excluding (right) nematodes.

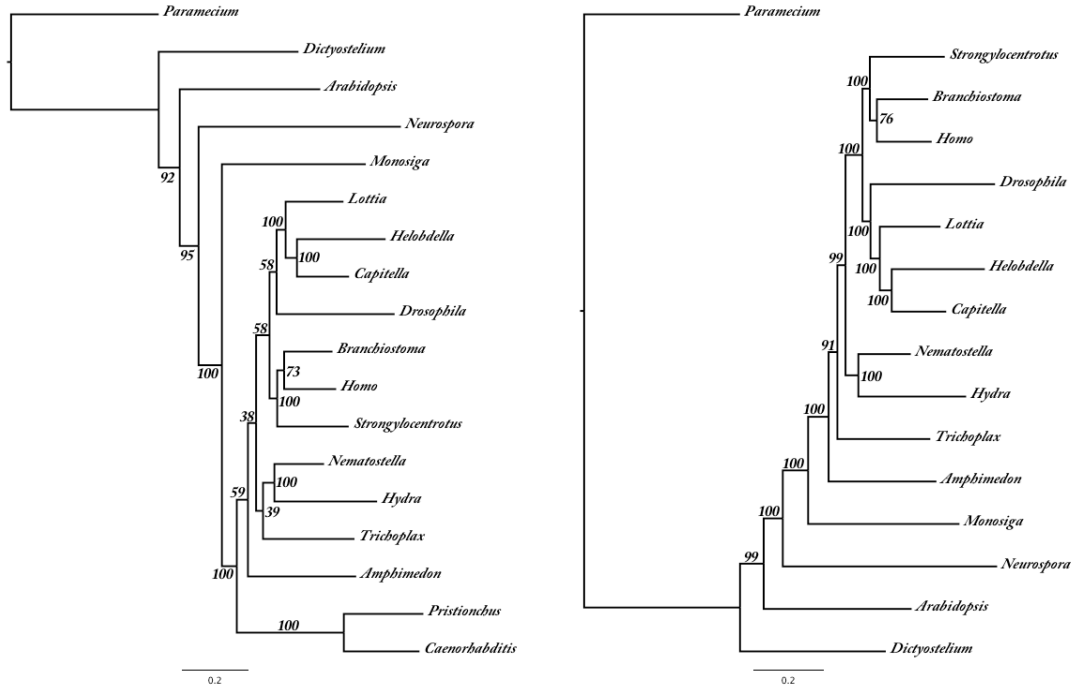


Figure S7.4: Maximum likelihood topology obtained for medium fMBH dataset including (left) and excluding (right) nematodes.

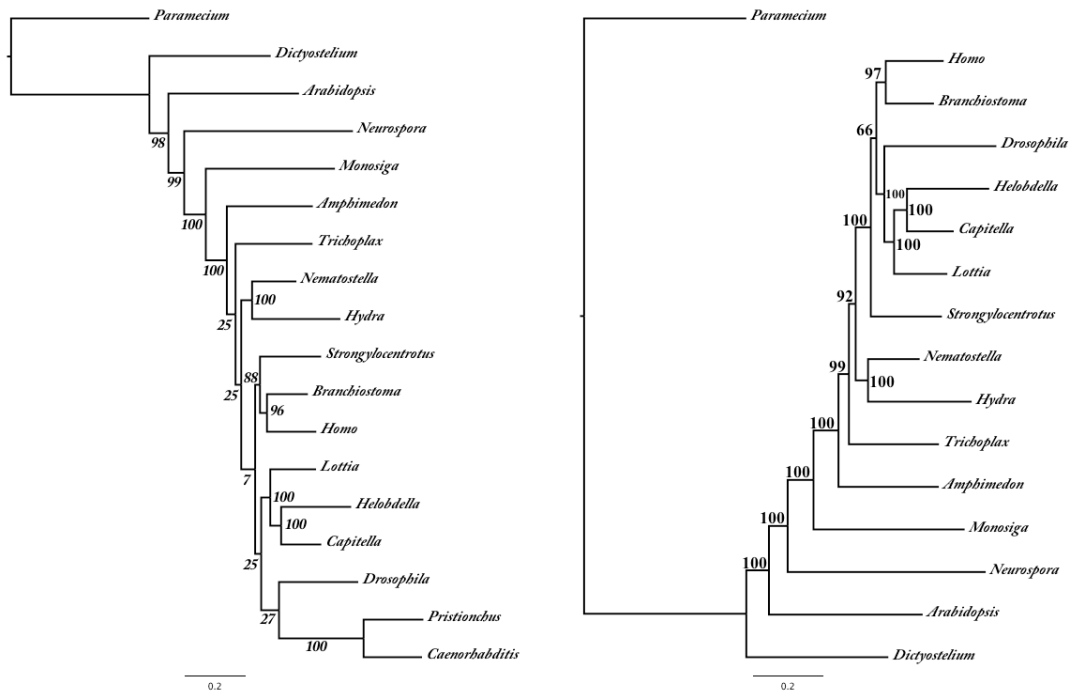


Figure S7.5: Maximum likelihood topology obtained for large FTK dataset including (left) and excluding (right) nematodes.

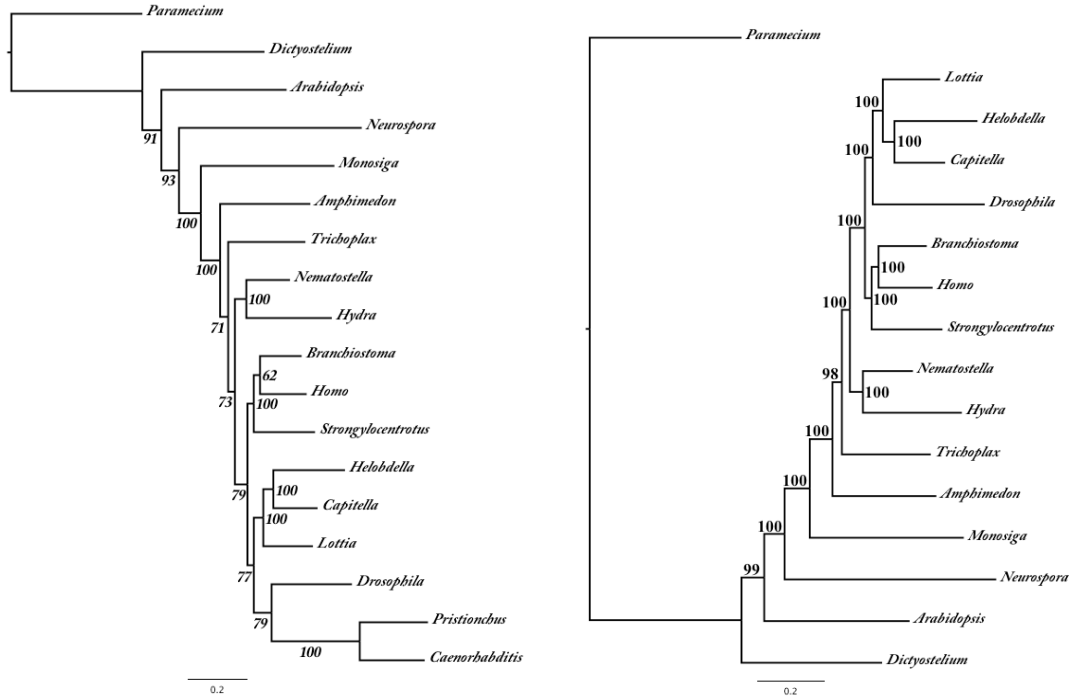


Figure S7.6: Maximum likelihood topology obtained for large fMBH dataset including (left) and excluding (right) nematodes.

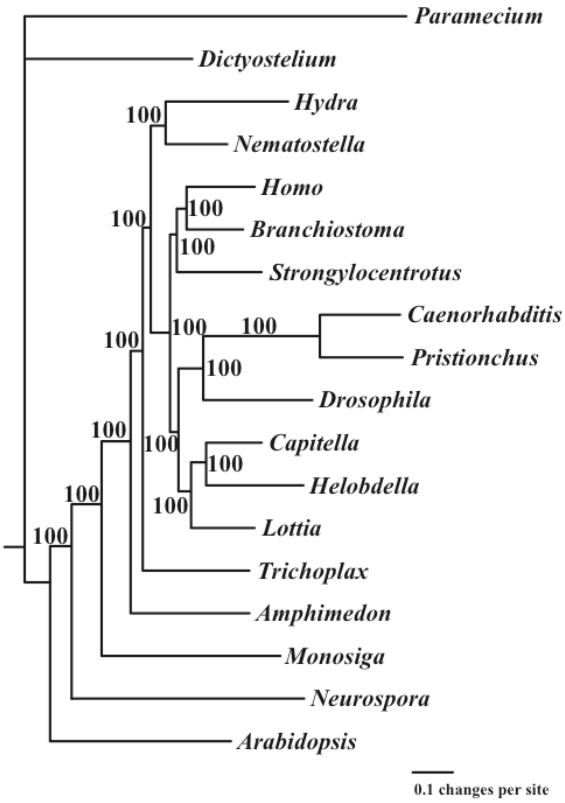


Figure S7.7: Bayesian inference-derived topology obtained for large fMBH dataset including nematodes.

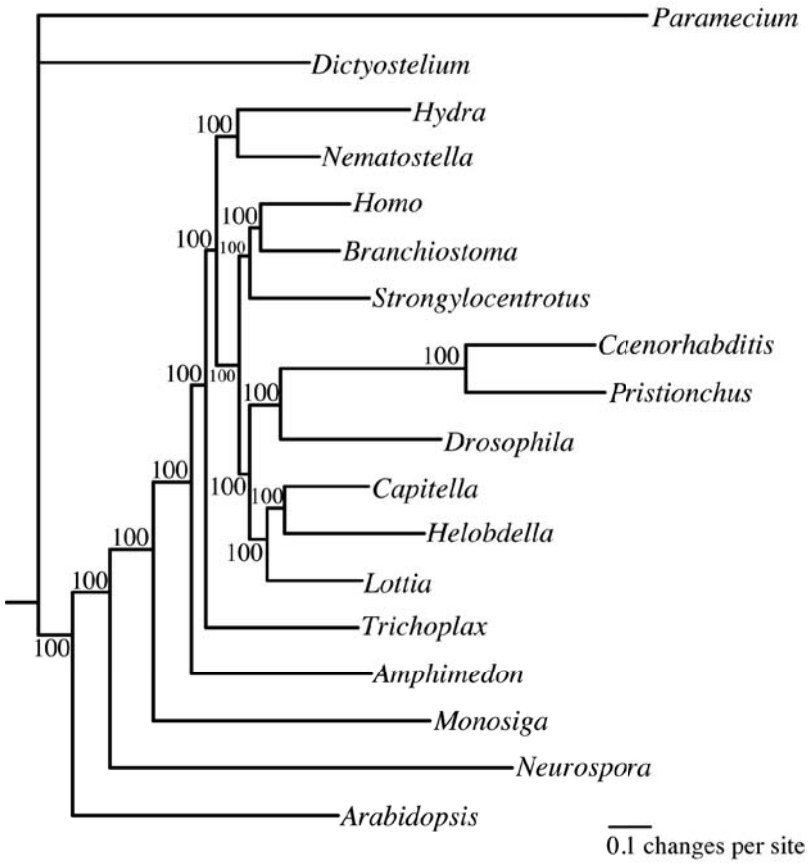


Figure S7.8: Bayesian inference topology obtained using a GTR+γ model of amino acid evolution in aamodel for large fMBH dataset including nematodes.

Trees in the 95% clade credibility set:

tree tree_1 [p = 1.000, P = 1.000] = [&W 1.000000]
(*Paramecium*,*Dictyostelium*,(*Arabidopsis*,(*Neurospora*,(*Monosiga*,(*Amphimedon*,(*Trichoplax*,((*Hydra*,*Nematostella*),((*Strongylocentrotus*,(*Homo*,*Branchiostoma*)),((*Drosophila*,(*Caenorhabditis*,*Pristionchus*)),(*Lottia*,(*Capitella*,*Helobdella*))))))))));

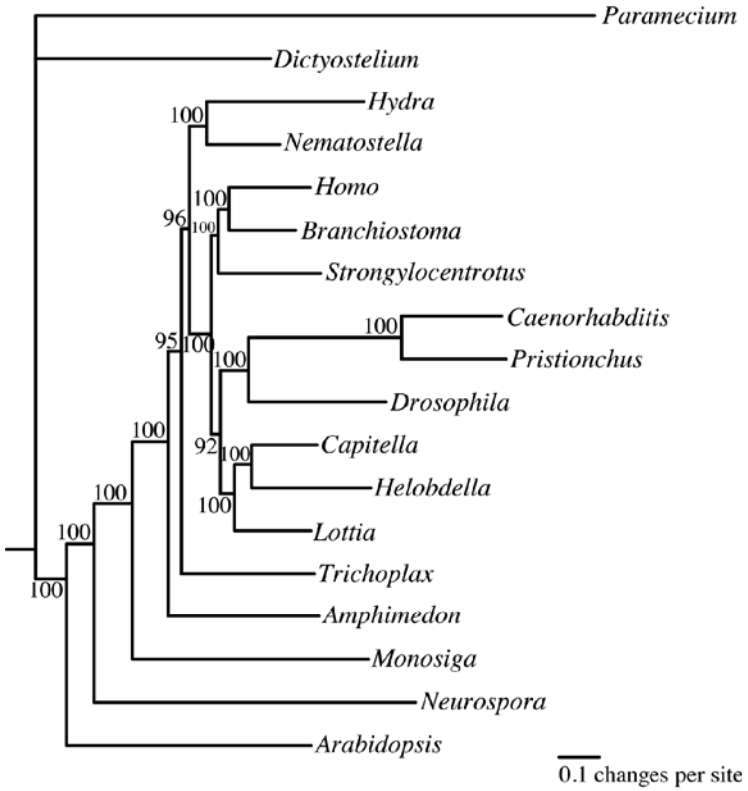


Figure S7.9: Bayesian inference topology obtained using a GTR+γ model of amino acid evolution in aamodel for large FTK dataset including nematodes.

Trees in the 95% clade credibility set:

tree tree_1 [p = 0.920, P = 0.920]
(*Paramecium*,*Dictyostelium*,(*Arabidopsis*,(*Neurospora*,(*Monosiga*,(*Amphimedon*,(*Trichoplax*,((*Hydra*,*Nematostella*),((*Strongylocentrotus*,(*Homo*,*Branchiostoma*)),((*Drosophila*,(*Caenorhabditis*,*Pristionchus*)),(*Lottia*,(*Capitella*,*Helobdella*))))))))));

tree tree_2 [p = 0.041, P = 0.961]
(*Paramecium*,*Dictyostelium*,(*Arabidopsis*,(*Neurospora*,(*Monosiga*,(((*Hydra*,*Nematostella*),(*Amphimedon,Trichoplax*)),((*Drosophila*,(*Caenorhabditis*,*Pristionchus*)),((*Strongylocentrotus*,(*Homo*,*Branchiostoma*)),(*Lottia*,(*Capitella*,*Helobdella*)))))))));

tree tree_3 [p = 0.030, P = 0.991]
(*Paramecium*,*Dictyostelium*,(*Arabidopsis*,(*Neurospora*,(*Monosiga*,(*Amphimedon*,(*Trichoplax*,((*Hydra*,*Nematostella*),((*Drosophila*,(*Caenorhabditis*,*Pristionchus*)),((*Strongylocentrotus*,(*Homo*,*Branchiostoma*)),(*Lottia*,(*Capitella*,*Helobdella*))))))))));

tree tree_4 [p = 0.009, P = 1.000]
(*Paramecium*,*Dictyostelium*,(*Arabidopsis*,(*Neurospora*,(*Monosiga*,(((*Hydra*,*Nematostella*),((*Drosophila*,(*Caenorhabditis*,*Pristionchus*)),((*Strongylocentrotus*,(*Homo*,*Branchiostoma*)),(*Lottia*,(*Capitella*,*Helobdella*))))),(*Amphimedon,Trichoplax*))))));

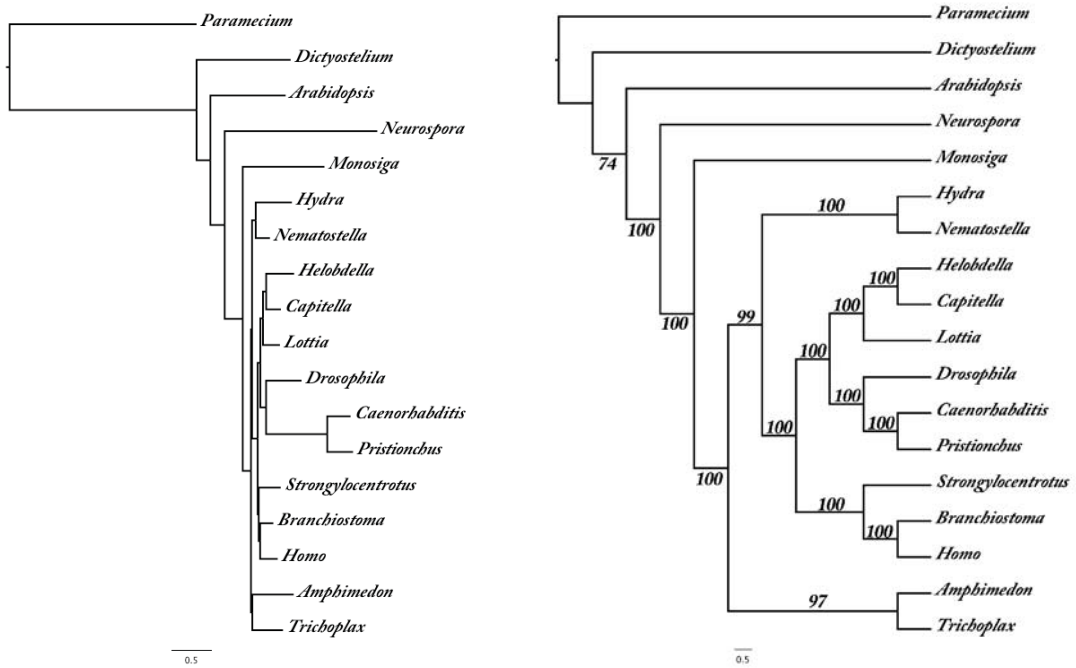


Figure S7.10: Bayesian inference topology obtained using CAT+Poisson model of amino acid evolution in PhyloBayes for large fMBH dataset including nematodes. Phylogram on left; cladogram on right shows frequency of taxon bipartitions with clarity.

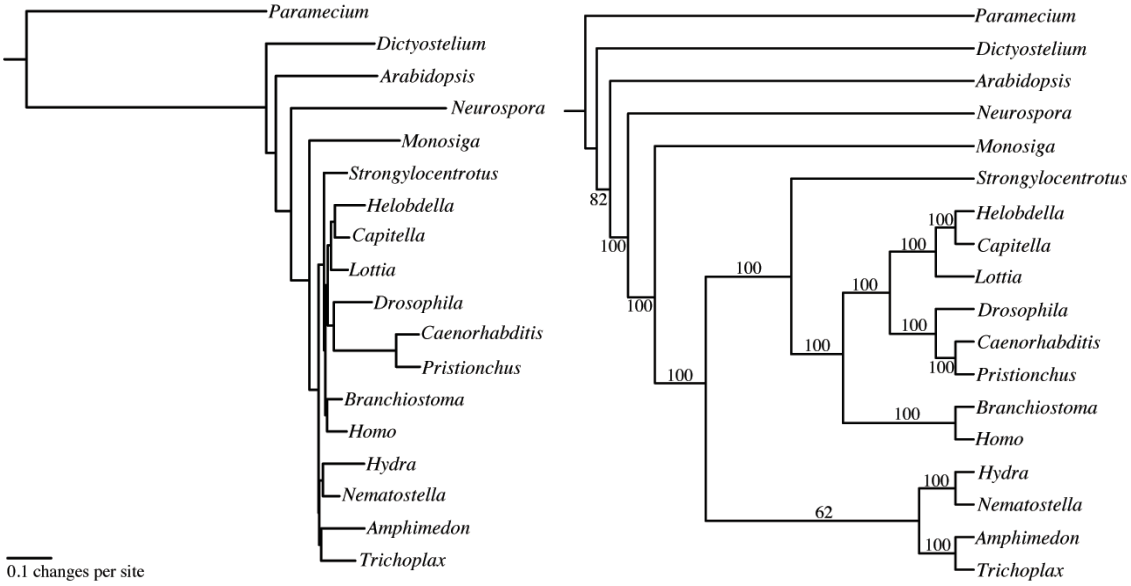


Figure S7.11: Bayesian inference topology obtained using CAT+Poisson model of amino acid evolution in PhyloBayes for large FTK dataset including nematodes. Phylogram on left; cladogram on right shows frequency of taxon bipartitions with clarity.

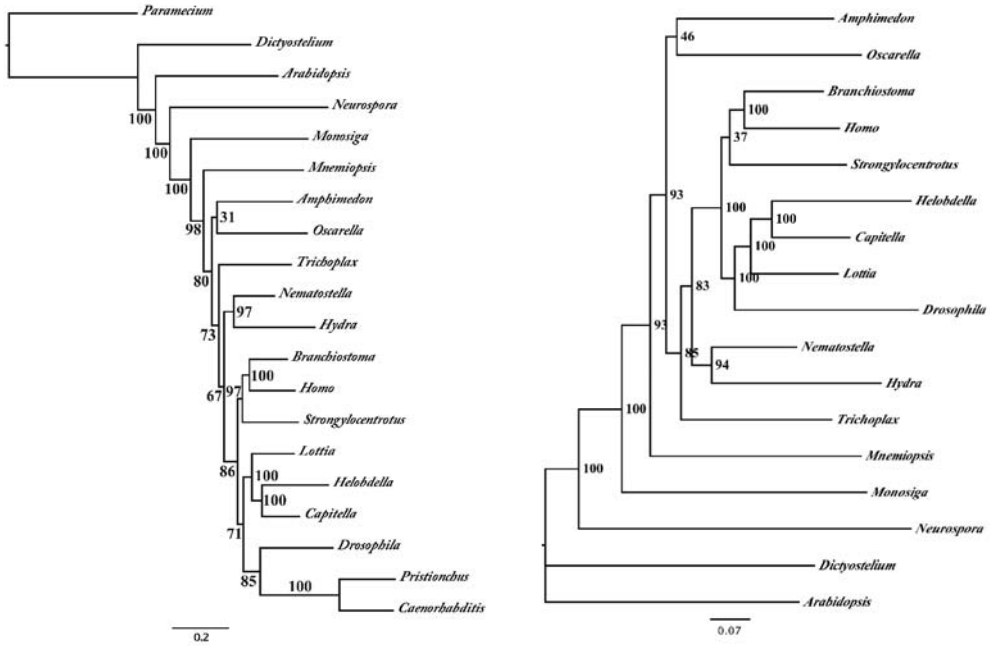


Figure S7.12: Maximum likelihood topology obtained for large FTK dataset with (left) and without (right) nematodes, with *Oscarella* and *Mnemiopsis* genes added where possible.

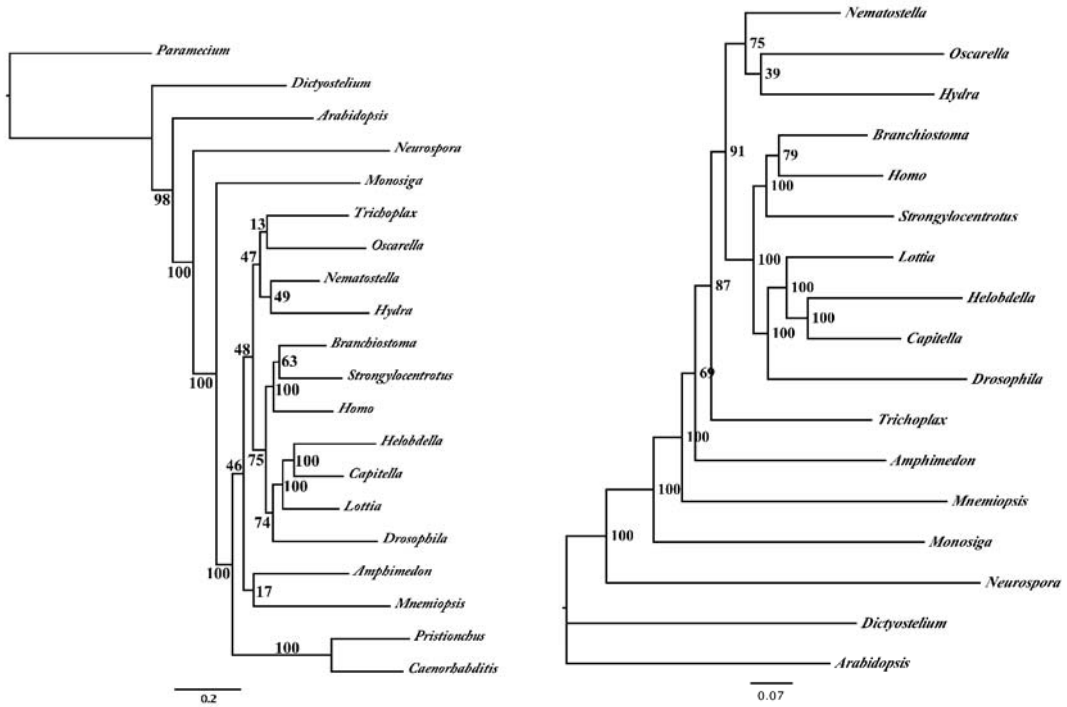


Figure S7.13: Maximum likelihood topology obtained for large fMBH dataset with (left) and without (right) nematodes, with *Oscarella* and *Mnemiopsis* genes added where possible.

Table S7.5: Key for species symbols used in topologies in Tables S7.6, S7.7, S7.13, S7.14.

Symbol	Species
p11	*Homo sapiens*
p17	*Brachiostoma floridae*
p24	*Strongylocentrotus purpuratus*
p3	*Drosophila melanogaster*
p12	*Caenorhabditis elegans*
p111	*Pristionchus pacificus*
p57	*Lottia gigantea*
p60	*Capitella sp. 1*
p65	*Helobdella robusta*
p19	*Nematostella vectensis*
p97	*Hydra magnipapillata*
p63	*Trichoplax adhaerens*
p100	*Amphimedon queenslandica*
p21	*Monosiga brevocollis*
p80	*Neurospora crassa*
p98	*Arabidopsis thaliana*
p101	*Paramecium tetraurilia*
p48	*Dictyostelium discoideum*
p93	*Oscarella carmella*
p104	*Mnemiopsis leidyi*

Table S7.6: Topologies tested with fMBH and FTK datasets without nematodes

Tree 1 = *Trichoplax* sister to cnidarians and bilaterians
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 2 = *Trichoplax* sister to cnidarians and bilaterians with *Drosophila* as early-branching bilaterian
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 3 = *Trichoplax* as earliest animal branch
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 4 = *Trichoplax* sister to cnidarians and bilaterians with *Homo* as early-branching deuterostome
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 5 = *Trichoplax* as earliest animal branch with *Drosophila* as early-branching bilaterian
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 6 = *Trichoplax* as earliest animal branch with *Homo* as early-branching deuterostome
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 7 = *Trichoplax* sister to cnidarians
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 8 = *Trichoplax* sister to cnidarians with *Drosophila* as early-branching bilaterian
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 9 = *Trichoplax* sister to cnidarians with *Homo* as early-branching deuterostome
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 10 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians
(((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 11 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians with *Drosophila* as early-branching bilaterian
(((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 12 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians with *Homo* as early-branching deuterostome
(((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);

Table S7.7: Topologies tested with fMBH and FTK datasets including nematodes

Tree 1 = *Trichoplax* sister to cnidarians and bilaterians
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 2 = *Trichoplax* sister to cnidarians and bilaterians with nematodes as early branching animals
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 3 = *Trichoplax* sister to cnidarians and bilaterians with nematodes as early branching animals, *Drosophila* as early branching bilaterian
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 4 = *Trichoplax* sister to cnidarians and bilaterians with nematodes as early branching animals, *Drosophila* as early branching bilaterian, *Homo* as early branching deuterostome
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 5 = *Trichoplax* as earliest animal branch
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 6 = *Trichoplax* as earliest animal branch with nematodes as early branching animals
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 7 = *Trichoplax* as earliest animal branch with nematodes as early branching animals, *Drosophila* as early branching bilaterian
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 8 = *Trichoplax* as earliest animal branch with nematodes as early branching animals, *Drosophila* as early branching bilaterian, *Homo* as early branching deuterostome
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p100:1.0):1.0,p63:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 9 = *Trichoplax* sister to cnidarians
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 10 = *Trichoplax* sister to with nematodes as early branching animals
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 11 = *Trichoplax* sister to cnidarians with nematodes as early branching animals, *Drosophila* as early branching bilaterian
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 12 = *Trichoplax* sister to cnidarians with nematodes as early branching animals, *Drosophila* as early branching bilaterian, *Homo* as early branching deuterostome
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0):1.0):1.0,p100:1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 13 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians
(((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 14 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians with nematodes as early branching animals
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,p3:1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 15 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians with nematodes as early branching animals, *Drosophila* as early branching bilaterian
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);
Tree 16 = *Trichoplax, Amphimedon* and cnidarians sister to bilaterians with nematodes as early branching animals, *Drosophila* as early branching bilaterian, *Homo* as early branching deuterostome
(((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,((p17:1.0,p24:1.0):1.0,p11:1.0):1.0):1.0,p3:1.0):1.0,((p19:1.0,p97:1.0):1.0,p63:1.0,p100:1.0):1.0):1.0,(p111:1.0,p12:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p101:1.0,p48:1.0);

Table S7.8: p-values for statistical tests of topologies using the FTK datasets including nematodes.

Topology Tested	Four taxon kernel (FTK) method (including nematodes)								
	Small			Medium			Large		
	WKH	ELW	AU	WKH	ELW	AU	WKH	ELW	AU
Tree 1	0.275	0.1691	0.380	0.951	0.9119	0.981	0.952	0.9410	0.950
Tree 2	0.004	0.0020	0.004	0.005	0.0010	0.007	0	0.0000	3e-06
Tree 3	5e-04	0.000	0.001	0	0.0000	4e-04	0	0.0000	7e-51
Tree 4	4e-04	0.0000	3e-07	2e-05	0.0000	1e-04	0	0.0000	4e-38
Tree 5	0.725	0.6747	0.785	0.039	0.0321	0.043	0	0.0000	1e-04
Tree 6	2e-04	0.0000	4e-05	0.001	0.0000	1e-04	0	0.0000	3e-45
Tree 7	3e-05	0.0000	6e-35	0	0.0000	9e-56	0	0.0000	5e-64
Tree 8	5e-05	0.0000	6e-15	0	0.0000	1e-05	0	0.0000	1e-51
Tree 9	0.197	0.1482	0.230	0.049	0.0465	0.064	0.048	0.0590	0.050
Tree 10	0.009	0.0060	0.010	0.012	0.0085	0.015	0	0.0000	6e-06
Tree 11	0.002	0.0000	0.002	0.001	0.0000	2e-04	0	0.0000	4e-63
Tree 12	0.001	0.0000	0.001	3e-04	0.0000	5e-04	0	0.0000	2e-14
Tree 13	0.001	0.0000	2e-04	2e-05	0.0000	3e-34	0	0.0000	4e-05
Tree 14	0.001	0.0000	8e-08	3e-04	0.0000	2e-62	0	0.0000	5e-65
Tree 15	4e-04	0.0000	2e-04	0	0.0000	1e-45	0	0.0000	3e-48
Tree 16	5e-04	0.0000	1e-05	0	0.0000	5e-05	0	0.0000	6e-39

WKH – Weighted Kishino Hasegawa Test
ELW – Expected Likelihood Weight Test
AU – Approximately Unbiased Test

Table S7.9: p-values for statistical tests of topologies using the FTK datasets without nematodes.

Topology Tested	Four taxon kernel (FTK) method (without nematodes)								
	Small			Medium			Large		
	WKH	ELW	AU	WKH	ELW	AU	WKH	ELW	AU
Tree 1	0.335	0.2443	0.481	0.974	0.9602	0.990	0.983	0.9829	0.980
Tree 2	0.004	0.0000	0.008	0	0.0000	8e-07	0	0.0000	3e-04
Tree 3	0.006	0.0034	0.018	9e-05	0.0000	2e-04	0	0.0000	4e-09
Tree 4	0.665	0.6456	0.726	0.026	0.0233	0.025	0	0.0000	3e-11
Tree 5	0.004	0.0011	0.003	0	0.0000	3e-05	0	0.0000	3e-78
Tree 6	0.192	0.0524	0.182	0.016	0.0082	0.014	0.017	0.0085	0.021
Tree7	0.192	0.0524	0.182	0.016	0.0082	0.014	0.017	0.0085	0.021
Tree 8	0.003	0.0000	0.003	0	0.0000	1e-05	0	0.0000	9e-05
Tree 9	0.005	0.0010	0.002	0	0.0000	1e-06	0	0.0000	7e-05
Tree 10	3e-04	0.0000	6e-10	0	0.0000	0.001	0	0.0000	1e-56
Tree 11	0	0.0000	3e-04	0	0.0000	2e-04	0	0.0000	9e-52
Tree 12	0	0.0000	4e-04	0	0.0000	1e-48	0	0.0000	2e-05

Table S7.10: p-values for statistical tests of topologies using the fMBH datasets including nematodes.

Topology Tested	Filtered mutual best hits (fMBH) method (including nematodes)								
	Small			Medium			Large		
	WKH	ELW	AU	WKH	ELW	AU	WKH	ELW	AU
Tree 1	0.801	0.6204	0.913	0.974	0.9538	0.984	0.986	0.9788	0.993
Tree 2	0.035	0.0042	0.025	0	0.0000	8e-05	0	0.0000	4e-73
Tree 3	0.089	0.0125	0.114	0	0.0000	5e-62	0	0.0000	4e-123
Tree 4	0.122	0.0432	0.205	0	0.0000	6e-53	0	0.0000	1e-47
Tree 5	0.199	0.1769	0.270	0.016	0.0196	0.018	0.002	0.0024	0.001
Tree 6	0.004	0.0000	0.001	0	0.0000	7e-52	0	0.0000	1e-99
Tree 7	0.009	0.0000	0.005	0	0.0000	1e-56	0	0.0000	7e-75
Tree 8	0.012	0.0000	5e-04	0	0.0000	7e-66	0	0.0000	3e-54
Tree 9	0.070	0.0497	0.122	0.026	0.0266	0.033	0.014	0.0188	0.013
Tree 10	0.036	0.0033	0.018	0	0.0000	1e-04	0	0.0000	5e-79
Tree 11	0.099	0.0166	0.154	0	0.0000	1e-07	0	0.0000	3e-43
Tree 12	0.141	0.0724	0.262	0	0.0000	1e-05	0	0.0000	3e-78
Tree 13	0.001	0.0000	0.001	0	0.0000	5e-07	0	0.0000	0.001
Tree 14	0.003	0.0000	0.002	0	0.0000	0.002	0	0.0000	6e-52
Tree 15	0.016	0.0000	0.003	0	0.0000	3e-04	0	0.0000	5e-57
Tree 16	0.029	0.0007	0.015	0	0.0000	3e-11	0	0.0000	2e-81

Table S7.11: p-values for statistical tests of topologies using the fMBH datasets without nematodes.

Topology Tested	Filetered mutual best hits (fMBH) method (without nematodes)								
	Small			Medium			Large		
	WKH	ELW	AU	WKH	ELW	AU	WKH	ELW	AU
Tree 1	0.164	0.0925	0.264	0.984	0.9659	0.996	0.998	0.9979	1.000
Tree 2	0.283	0.1849	0.522	0.002	0.0020	0.001	0	0.0000	1e-06
Tree 3	0.717	0.6081	0.854	0.002	0.0000	0.004	0	0.0000	4e-07
Tree 4	0.085	0.0273	0.109	0.016	0.0201	0.017	0.002	0.0020	0.002
Tree 5	0.122	0.0400	0.152	0	0.0000	7e-05	0	0.0000	3e-32
Tree 6	0.042	0.0030	0.030	0.010	0.0060	0.008	0.002	0.0000	3e-04
Tree7	0.042	0.0030	0.030	0.010	0.0060	0.008	0.002	0.0000	3e-04
Tree 8	0.054	0.0116	0.063	4e-04	0.0000	0.002	0	0.0000	1e-63
Tree 9	0.056	0.0298	0.098	0.001	0.0000	0.001	0	0.0000	7e-53
Tree 10	2e-04	0.0000	6e-51	0	0.0000	3e-05	0	0.0000	1e-52
Tree 11	2e-04	0.0000	5e-05	0	0.0000	2e-06	0	0.0000	2e-06
Tree 12	3e-04	0.0000	0.002	0	0.0000	3e-39	0	0.0000	9e-43

Table S7.12: Details of the Phylobayes analyses

	FTK method		fMBH method	
	Chain 1	Chain 2	Chain 1	Chain 2
# cycles	31935	31744	64623	65484
# generations	1014263	1004700	1871779	1894504
# trees considered after burnin	309	307	636	644
Mean difference	0.00155729		0.00484081	
Max difference	0.0574934		0.3222864	

Though the Max difference for the fMBH dataset is not < 0.1, it is at the acceptable level of 0.3, giving a good qualitative picture of the posterior consensus. The source of disagreement between the two runs is the alternative placement of *Paramecium* and *Dictyostelium* in a group, vs. placing *Paramecium* and *Arabidopsis* in a group.

Table S7.13: Topologies with different placements for *Oscarella* tested with fMBH and FTK datasets. The topologies were also tested by removing nematodes (p12 and p111).

Tree 1 = *Oscarella* as the earliest animal branch
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p93:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 2 = *Oscarella* as sister to cnidarians
((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p93:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 3 = *Oscarella* as branch after *Trichoplax* but before cnidarians
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p93:1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 4 = *Oscarella* as sister to *Trichoplax*
((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,(p63:1.0,p93:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 5 = *Oscarella* as branch after *Amphimedon* but before *Trichoplax*
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p93:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 6 = *Oscarella* as sister to *Amphimedon*
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,(p93:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 7 = *Oscarella* as sister to bilaterians
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p93:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);

Table S7.14: Topologies with different placements for *Mnemiopsis* tested with fMBH and FTK datasets. The topologies were also tested by removing nematodes (p12 and p111).

Tree 1 = *Mnemiopsis* as the earliest animal branch
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p104:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 2 = *Mnemiopsis* as sister to cnidarians
((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,((p19:1.0,p97:1.0):1.0,p104:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 3 = *Mnemiopsis* as branch after *Trichoplax* but before cnidarians
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p104:1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 4 = *Mnemiopsis* as sister to *Trichoplax*
((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,(p63:1.0,p104:1.0):1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 5 = *Mnemiopsis* as branch after *Amphimedon* but before *Trichoplax*
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p104:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 6 = *Mnemiopsis* as sister to *Amphimedon*
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,(p104:1.0,p100:1.0):1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);
Tree 7 = *Mnemiopsis* as sister to bilaterians
((((((((((p57:1.0,(p60:1.0,p65:1.0):1.0):1.0,(p3:1.0,(p111:1.0,p12:1.0):1.0):1.0):1.0,((p17:1.0,p11:1.0):1.0,p24:1.0):1.0):1.0,p104:1.0):1.0,(p19:1.0,p97:1.0):1.0):1.0,p63:1.0):1.0,p100:1.0):1.0,p21:1.0):1.0,p80:1.0):1.0,p98:1.0,p48:1.0);

Table S7.15: p-values for statistical tests for alternative positions for *Oscarella* using fMBH and FTK datasets. Values in parentheses are for the datasets without nematodes.

Topology Tested	FTK method			fMBH method		
	WKH	ELW	AU	WKH	ELW	AU
Oscarella as earliest animal branch (Tree 1)	0.087 (0.260)	0.0399 (0.1391)	0.087 (0.307)	0.011 (0.016)	0.0001 (0.0017)	0.001 (0.001)
Oscarella as sister to cnidarians (Tree 2)	0.012 (0.004)	0.0024 (0.0000)	0.011 (0.006)	0.053 (0.039)	0.0344 (0.0258)	0.076 (0.053)
Oscarella as branch after *Trichoplax* but before cnidarians (Tree 3)	0.010 (0.006)	0.0001 (0.0001)	0.012 (0.006)	0.567 (0.602)	0.3271 (0.4125)	0.749 (0.805)
Oscarella as sister to *Trichoplax* (Tree 4)	0.130 (0.243)	0.0925 (0.1731)	0.153 (0.259)	0.379 (0.267)	0.1895 (0.1212)	0.423 (0.299)
Oscarella as branch after *Amphimedon* but before *Trichoplax* (Tree 5)	0.212 (0.386)	0.1454 (0.1772)	0.395 (0.581)	0.433 (0.398)	0.1364 (0.1668)	0.547 (0.565)
Oscarella as sister to *Amphimedon* (Tree 6)	0.788 (0.614)	0.7179 (0.5095)	0.865 (0.709)	0.364 (0.317)	0.3001 (0.2492)	0.393 (0.341)
Oscarella as sister to bilaterians (Tree 7)	0.007 (0.007)	0.0018 (0.0011)	0.010 (0.008)	0.035 (0.044)	0.0125 (0.0228)	0.041 (0.050)

Table S7.16: p-values for statistical tests for alternative positions for *Mnemiopsis* using fMBH and FTK datasets. Values in parentheses are for the datasets without nematodes.

Topology Tested	FTK method			fMBH method		
	WKH	ELW	AU	WKH	ELW	AU
Mnemiopsis as earliest animal branch (Tree 1)	0.714 (0.766)	0.6911 (0.7372)	0.750 (0.795)	0.727 (0.809)	0.5274 (0.6705)	0.766 (0.857)
Mnemiopsis as sister to cnidarians (Tree 2)	0.002 (0.017)	0.0000 (0.0010)	0.002 (0.003)	0.057 (0.029)	0.0218 (0.0158)	0.084 (0.045)
Mnemiopsis as branch after *Trichoplax* but before cnidarians (Tree 3)	0.001 (0.001)	0.0000 (0.0000)	7e-05 (9e-10)	0.060 (0.027)	0.0074 (0.0031)	0.114 (0.030)
Mnemiopsis as sister to *Trichoplax* (Tree 4)	0.031 (0.019)	0.0194 (0.0115)	0.035 (0.015)	0.236 (0.128)	0.1678 (0.0846)	0.309 (0.189)
Mnemiopsis as branch after *Amphimedon* but before *Trichoplax* (Tree 5)	0.015 (0.017)	0.0002 (0.0019)	0.003 (0.003)	0.276 (0.163)	0.0997 (0.0666)	0.482 (0.335)
Mnemiopsis as sister to *Amphimedon* (Tree 6)	0.286 (0.234)	0.2893 (0.2485)	0.347 (0.292)	0.273 (0.191)	0.1733 (0.1578)	0.379 (0.289)
Mnemiopsis as sister to bilaterians (Tree 7)	2e-04 (4e-04)	0.0000 (0.0000)	2e-04 (2e-04)	0.024 (0.011)	0.0027 (0.0015)	0.025 (0.014)

Table S7.8.1: Estimated divergence times for various nodes using the r8s[73] program.

Estimation done on tree from dataset	Holozoan divergence (mya)	Metazoan divergence (mya)	Placozoan-Cnidarian-Bilaterian divergence (mya)	Cnidarian-Bilaterian divergence (mya)
FTK large with nematodes	923	755	697	644
FTK large without nematodes	952	785	715	656
fMBH large with nematodes	958	777	702	656
fMBH large without nematodes	990	798	718	667

Table S8.2.1: Classification of cell cycle genes by origin.

Gene	Origin	Methods used to determine origin
CDKN1A-C or p21/p27/p57	eumetazoan	No putative members in sponges or non-animals. *Nematostella* and *Trichoplax* candidates have the DCI domain and pick up known CDKN1 proteins by BLAST; plants have CDI domains but those CDK inhibitors are fundamentally different proteins. Unlrelated CDK inhibitors are known in yeast and plants.[83]
CDKN2A-D or p15/p16/p18/p19	chordate	No hits found outside of chordates.
CyclinD	eukaryotic	Members of this group of cyclins are known in plants and animals (see Figure S8.2.1).
CylinE	metazoan subfamily	Cyclins are ancient; the well-supported cyclinE node on the tree has only animal sequences, but one *Monosiga* sequence picks up CyclinE as its best hits, though there is no support on the tree for the monophyly of this sequence with other CyclinE sequences (see Figure S8.2.1).
cyclinA	eukaryotic	Members of this group of cyclins are known in plants and animals (see Figure S8.2.1).
cyclinH	eukaryotic	Members of this group of cyclins found in plants and animals; one fungal sequence (clp1) has cyclinH sequences as its best hits (see Figure S8.2.1).
CyclinB	eukaryotic	Members of this group of cyclins found in plants, fungi and animals (see Figure S8.2.1).
CDK4/6	eumetazoan subfamily	CDKs are ancient, but this subfamily has members only in eumetazoans (see Figure S8.2.2).
PFTAIRE CDK	metazoan subfamily	The PFTAIRE family is unique to animals (see Table S8.7.2).
PCTAIRE CDK	holozoan subfamily	The PCTAIRE family is unique to holozoans (see Table S8.7.2).
CDK10	holozoan subfamily	CDKs are ancient, but this subfamily has members only in choanoflagellates and animals (see Figure S8.2.2).
CDK9	opisthokont	CDKs are ancient, but this subfamily has members only in animals and fungi. SGV1 in *S. cerevisiae* is orthologous to CDK9.
CDK2	metazoan	Cdk2 is an animal-specific subfamily of the Cdc2 group of cyclin dependent kinases (see table S8.7.2).
CCRK CDK	holozoan subfamily	CDKs are ancient, but this subfamily has members only in choanoflagellates and animals (see Figure S8.2.2).
CDC25	eukaryotic	Members of this family present in all eukaryotes; differentiation into A,B,C groups is specific to vertebrates.[79,80]
CDK7	eukaryotic	Members of this subfamily of CDKs found in plants, fungi and animals (see Figure S8.2.2).
Wee1	eukaryotic	Wee1 subfamily of the Wee family is known to be in all eukaryotes; Mik1 subfamily is unique to fungi; Myt1 subfamily is found in animals including sponges and cnidarians. Choanoflagellates only have Wee1, suggesting that Myt1 is an animal-specific subfamily.[156]
Rb	eukaryotic	Yeast do not have this ancient transcription factor, but animals, plants and choanoflagellates do.[157]
Src subgroup	holozoan	Tyrosine kinases are unique to holozoans. As published, *Monosiga* has TKs belonging to the Abl, the Src and the Tec families. The Frk subfamily appears to be new to animals (see Table S8.7.2).
E2F/Dp	eukaryotic superfamily	The E2F/DP group of transcription factors are ancient. The genomes of *Nematostella*, *Trichoplax*, *Monosiga* and *Amphimedon* each have one DP ortholog. Multiple members of the E2F sub-group are found in *Nematostella*, *Trichoplax*, *Monosiga* and *Amphimedon* (see Figure S8.2.4).[156]
Transcriptional regulator myc	metazoan	Plants have Myc-like bHLH proteins, however, only animals and choanoflagellates have putative Myc and Max subfamily members. All animal Myc sequences have similarity in the region N-terminal to the bHLH and zipper regions (including the MycBoxII with the 'DCMW' motif and the acidic 'EIDVVS' motif). However, the *Monosiga* protein lacks any similarity to animal Mycs in the N-terminal regions (see Figure S8.2.3).
NEK	eukaryotic superfamily	Different families appear to have varied origins. Nek1, Nek6 and Nek11 are metazoan novelties. The other subfamilies are ancient (see Table S8.7.2).
MDM2	eumetazoan	The bilaterian sequences contain the unique combination of SWIB/MDM2 and Zinc_finger_in_RanBP domains. While *Nematostella* and *Trichoplax* hits lack recognizable Pfam domains they have reciprocal best alignments to MDM2/4 and are recognized by the MDM2 Panther HMM.
ATM	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.

ATR	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
Checkpoint S/T protein kinase CHK1	opisthokont	Known to be present in fungi and animals.[158]
Checkpoint S/T protein kinase CHK2 (RAD53)	opisthokont	Known to be present in fungi and animals.[159]
CAK CDK activating kinase	holozoan	This is a complex of Cdk7, cyclinH and ccrk,[160] which are present in holozoans.
Transcriptional regulator myb	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
NBS1/nibrin	metazoan	No hits outside of animals. One gene each in *Trichoplax*, *Nematostella*, *Amphimedon*.
FANCD2	eukaryotic	Cannot find in *Nematostella*, but present in choanoflagellates, *Amphimedon* and plants.[161]
HIPK	metazoan	No hits outside of animals. One gene each in *Trichoplax*, *Nematostella*, *Amphimedon*.
DNAPK	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
BRCA1	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.[162]
p53	holozoan	Reported in choanoflagellates and cnidarians, found in Amphimedon and *Trichoplax*.[81] Absent in fungi and other eukaryotes.
Polo like kinase (PLK)	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
CDK/cyclin regulator Mat1	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
Aurora kinase	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.

Figure S8.2.1: Neighbor-joining tree for cyclin genes. CyclinE (red) and CyclinJ (green) appear to be animal-specific subfamilies. The names of *Monosiga* genes that have best hits to these groups are highlighted in the respective colors. However, they do not form monophyletic groupings with their corresponding animal proteins. All other cyclin subfamilies appear to be ancient eukaryotic genes. Only cyclins A, B, D, and E are implicated in cell cycle control. Hs, *Homo sapiens*; Dr, *Danio rerio*; Sp, *Strongylocentrotus purpuratus*; Ce, *Caenorhabditis elegans*; Cb, *Caenorhabditis briggsae*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Hm, *Hydra magnipapillata*; Ta, *Trichoplax adhaerens*; Aqu, *Amphimedon queenslandica*; Mb, *Monosoga brevicollis*. At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Sc, *Saccharomyces cerevisiae*.

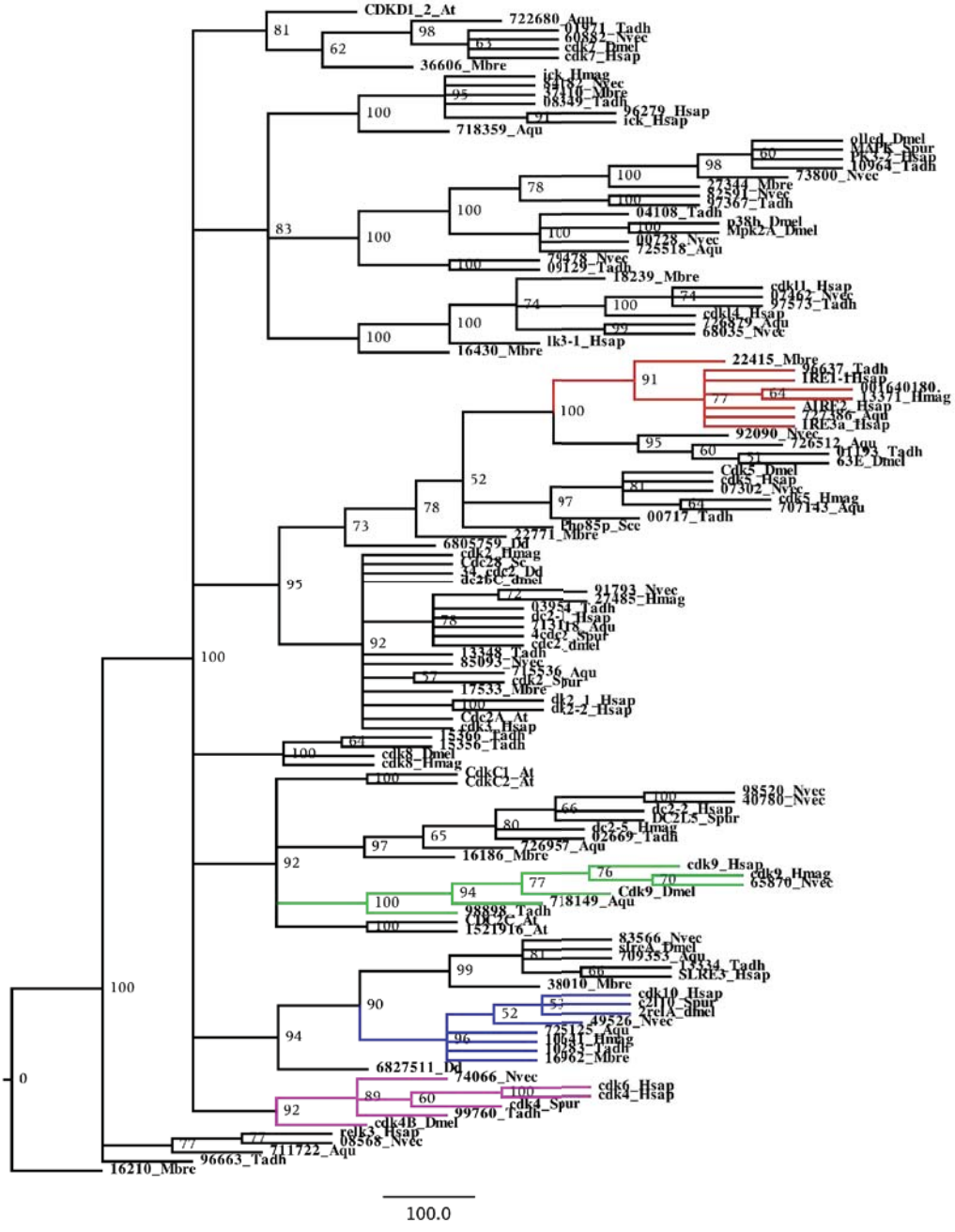


Figure S8.2.2: Neighbor-joining bootstrap tree for cyclin dependent kinase (CDK) genes. PFAIRE and PCTAIRE (red) and Cdk10 (blue) appear to be holozoan-specific subfamilies, Cdk9 (green) appears to be animal-specific, Cdk4/6 (pink) appears to be a eumetazoan-specific family. All other families are known to be ancient to eukaryotes. Hsap, *Homo sapiens*; Spur, *Strongylocentrotus purpuratus*; Dmel, *Drosophila melanogaster*; Nvec, *Nematostella vectensis*; Hmag, *Hydra magnipapillata*; Tadh, *Trichoplax adhaerens*; Aqu, *Amphimedon queenslandica*; Mbre, *Monosiga brevicollis*. At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum* .

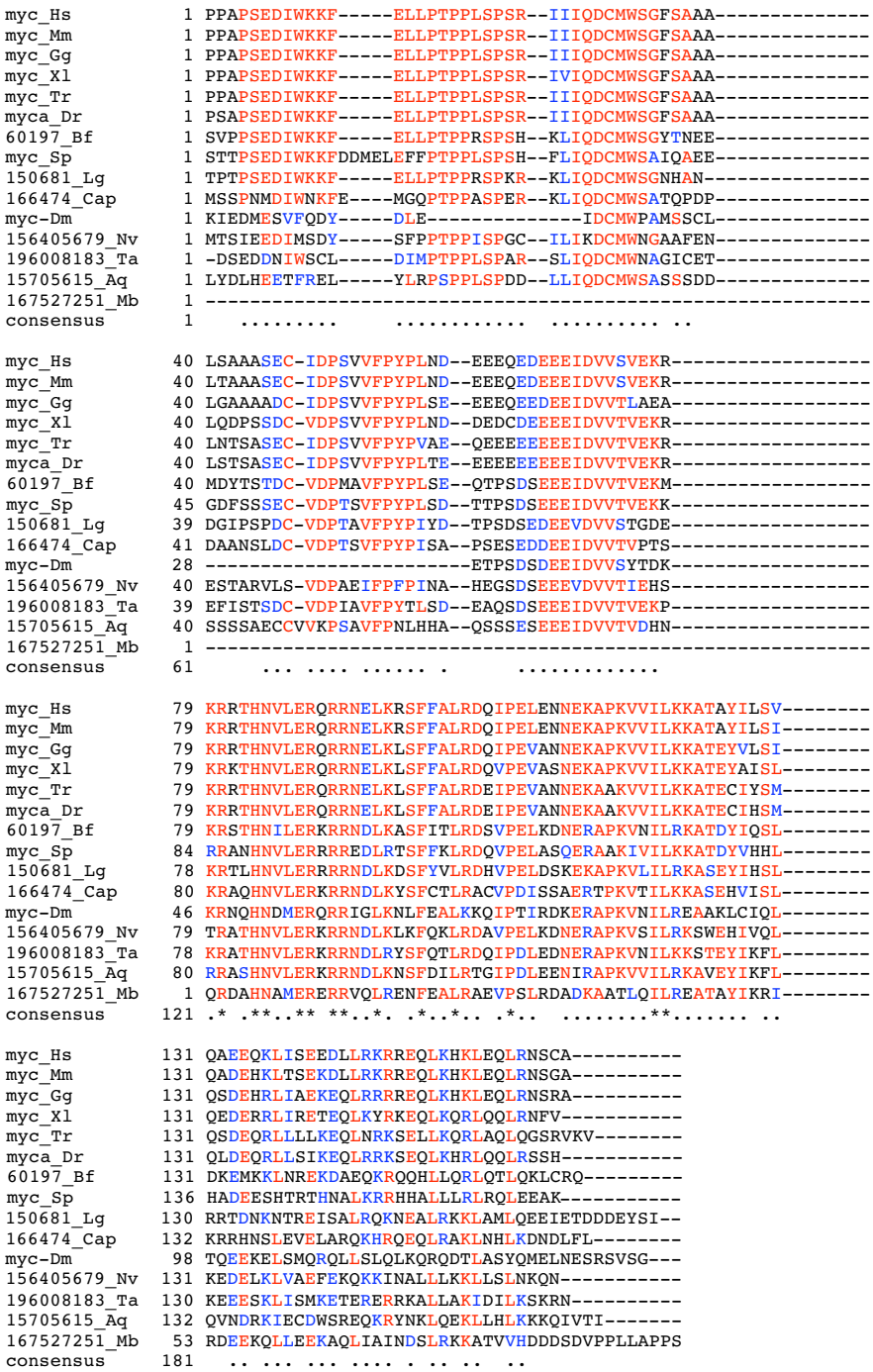


Figure S8.2.3: Alignment of animal and *Monosiga* Myc orthologs. The *Monosiga* protein aligns in the bHLH and zipper regions, but does not share conserved motifs found in the N-terminal region of all animal Myc proteins. Hs, *Homo sapiens*; Mm, *Mus musculus*; Gg, *Gallus gallus*; Xl, *Xenopus laevis*; Tr, *Tetraodon rubripes*; Dr, *Danio rerio*; Bf, *Branchiostoma floridae*; Sp, *Strongylocentrotus purpuratus*; Lg, *Lottia gigantea*; Cap, *Capitella*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosoga brevicollis*.

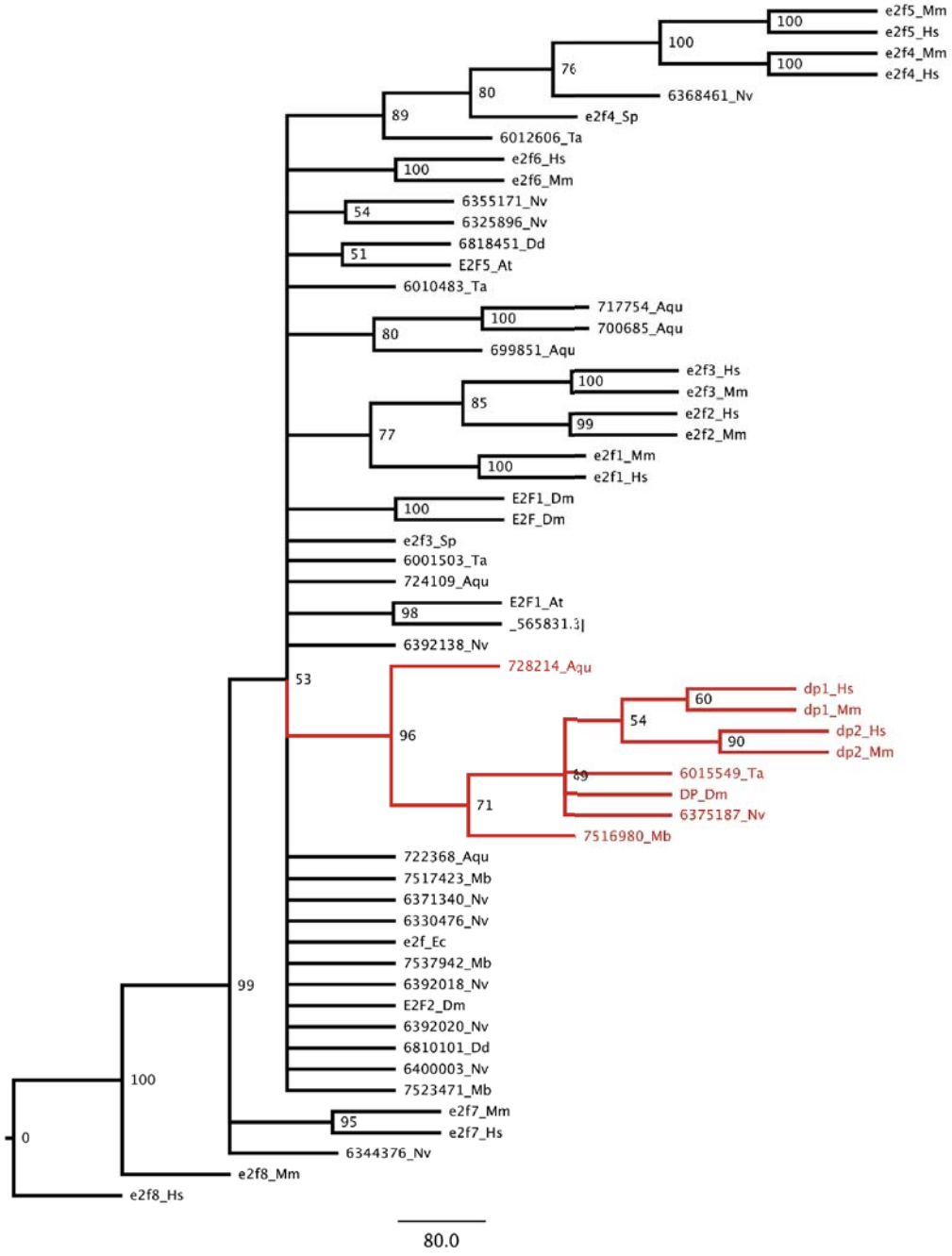


Figure S8.2.4: Neighbor-joining tree for E2F/Dp family of transcription factors. The E2F and DP subgroups are known to be ancient families with members known from plants and animals. Multiple members of the E2F group that likely fall into different E2F subfamilies are seen here from the genomes of *Nematostella*, *Trichoplax*, *Monosiga* and *Amphimedon*. One member of the DP group (red) was found in each of these genomes. Hs, *Homo sapiens*; Dr, *Danio rerio*; Sp, *Strongylocentrotus purpuratus*; Ce, *Caenorhabditis elegans*; Cb, *Caenorhabditis briggsae*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Hm, *Hydra magnipapillata*; Ta, *Trichoplax adhaerens*; Aqu, *Amphimedon queenslandica*; Mb, *Monosoga brevicollis*. At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Mm, *Mus musculus*.

Table S8.2.2: Classification of Akt signaling pathway genes by origin

Gene	Origin	Methods used to determine origin
PAK1	eukaryotic family	The PAK kinases break into two subfamilies: PAKA (PAK1-3 in human) is eukaryotic-wide (lost in plants) and PAKB (PAK4-6 in humans) is metazoan-specific (see Nichols 2004[163] and Table S8.8.1).
GSK3	eukaryotic	Known ancient protein.
AMPK	eukaryotic	Known ancient family[164] - called Snf1 in yeast.[165]
PDK1	eukaryotic	Known ancient family.[164]
Akt	eukaryotic	Known ancient family.[164] Known to be present in *Dictyostelium.*[166]
JAK	bilaterian	Known to exist in insects and vertebrates. No bonafide orthologs were found in cnidarians, placozoans or sponges.
STAT	eukaryotic	*Monosiga* has a STAT-like protein(king et al); also found in *Dictyostelium*[167]
RSK-p70 (S6K)	eukaryotic	No clear yeast ortholog[168], but present in plants[169].
LKB	eukaryotic	Stk11 is the AMPK/snf1 activating kinase in human. *Monosiga*, *Trichoplax*, *Amphimedon*, *Nematostella* have kinases with best hits to Stk11. Stk11 also known to be present in *Dictyostelium*[167],
mTOR	eukaryotic	Known ancient protein.[170,171]
PI3K	eukaryotic	Known ancient protein.[172]
c-Raf	metazoan	Putative orthologs with the RBD, C1, and STKC domains are present in *Nematostella*, *Trichoplax*, and *Amphimedon.*
TSC1	ophisthokont	*Trichoplax* and *Nematostella* have a putative ortholog with the hamartin domain. No putative ortholog found in *Monosiga*. Tsc1 known from *S. pombe*[173].
Gab family	metazoan	As reported previously, *Nematostella* has a Gab ortholog[174]. Congruent with Wohrle et al. 2009, we could not identify an ortholog in *Monosiga. Amphimedon* has a putative ortholog with a PH domain and top hits to Gab1.
TRAF3	metazoan	See table S8.10.1.
IRS-1	metazoan	*Nematostella* and *Trichoplax* have proteins with 2 PH domains (N terminal) as the human IRS-1 and have best hits to irs1. No similar proteins were found in *Monosiga* and yeast. *Amphimedon* has a putative IRS-1 ortholog with a partial PH domain.
Bax	eumetazoan	See Table S8.3.1.
Rheb	eukaryotic	Fungi, *Dictyostelium* and animals have rheb family of GTPases.[175]
Raptor	eukaryotic	Orthologs known in plants[169] and yeast (Kog1).
PTEN	eukaryotic	Proteins from basal animals, yeast and plants have the right domains - PTP_PTEN-C.[176]
4E-BP	eukaryotic	Known ancient protein.
Ras	eukaryotic	Known ancient protein.[177]
SHP1/SHP2	holozoan	*Monosiga* has one putative SHP sequence. *Amphimedon* has an ortholog with similar domain structure as the human sequence (SH2+SH2+PTP).

GRB2	holozoan	Proteins with a Grb2-like domain structure and displaying sequence homology to Grb2 family proteins were only found in animal and *Monosiga* genomes.
SOS	holozoan	Proteins sharing sequence similarity and domain architecture with vertebrate SOS1 were found in *Monosiga* and animal genomes but not in the other eukaryotic genomes surveyed.
Shc	holozoan	*Monosiga* is known to have a Shc ortholog.[127]
Cbl	eukaryotic	*Dictyostelium* has a divergent ortholog (Manning, personal communication); *Monosiga* is reported to have an ortholog[127].
Rictor	opisthokont	Ortholog in *S. cerevisiae* is called Avo3.
TSC2	eukaryotic	There is a known clear ortholog in *S. pombe*[173] and a likely one in *Dictyostelium* (XP_641197).
Bcl2L11/BIM	vertebrate	See Table S8.3.1.
NOXA	vertebrate	See Table S8.3.1.
Class-I helical cytokine receptor family	bilaterian	This family includes class-I helical interleukin receptors (ILRs), Leukemia inhibitory factor receptors (LIFR), Ciliary neurotrophic factor receptor (CNTFR), Growth hormorne receptor (GHR), Prolactin receptor (PRLR) and Leptin receptor (LPR). All these receptors appear to be restricted to vertebrates. However, insects have a distant homolog of the vertebrate type I cytokine receptors, Dome, that possesses the characteristic domains and key amino acids required for signalling. This suggests that the vertebrate families evolved from a single ancestral receptor that also gave rise to Dome.[92]
Erythropoietin receptor (Epo R)	vertebrate	Not found in *Ciona*[178], *Branchiostoma*[179], *Strongylocentrotus*[180]or other earlier animal lineages.
Interferon receptor (INFR)	vertebrate	Not found in *Ciona*[178], *Branchiostoma*[179], *Strongylocentrotus*[180]or other earlier animal lineages.
Granulocyte colony stimulating factor receptor (CSFR)	vertebrate	Not found in *Ciona*[178], *Branchiostoma*[179], *Strongylocentrotus*[180]or other earlier animal lineages.
Suppres or of cytokine signaling (SOCS)	metazoan	*Amphimedon* has three proteins with the correct domain composition (SH2+SOCS_box) that are putative orthologs to SOCS proteins. This domain combination is known in *Nematostella* and was not found in *Monosiga*.[181]

Table S8.2.3: Classification of Warts/Hippo pathway components by origin

Gene	Origin	Methods used to determine origin
Warts	eukaryotic	Known as dbf2 in yeast. Also found in plants. The mechanism of hpo-wts-mats is functional in yeast but used in mitotis exit (MEN) and septation initiation (SIN).[93,94]
Discs overgrown/CSNK1e	eukaryotic	Casein kinase 1 is an ancient eukaryotic family. The *Drosophila* Dco is not a clear CSK1e ortholog, since *Drosophila* does not have a delta isoform, and the Dco sequence picks up human Csk1e and Csk1d proteins. Choanoflagellates, sponges, placozoans and cnidarians all have a putative CSK1e/d ortholog.
Merlin	metazoan	*Nematostella*, *Amphimedon*, *Trichoplax*, *Monosiga* have proteins with the right domain combination - ferm-n+germ-m+ferm-c+ERM. *Nematostella* and *Amphimedon* proteins have as best hits NF2/merlin. The *Monosiga* protein appears to be a radixin/moesin sequence.
Dachs	metazoan	This is an unconventional myosin in *Drosophila* that has similarity to human myosins X/V/VII but is quite divergent. *Drosophila* members of myosin families V and VII are known. [182] yeast have myosin classes I, II and V. *Dictyostelium* has I, II, VII and XI. Human Myosin X picks up crinkled/VII, XV, II, before picking up dachs. *Nematostella* protein picks up Dachs as best hit by BLAST against nr with very low e-value, as do the *Trichoplax* and *Amphimedon* orthologs. The most likely *Monosiga* myosin hit picks up myosin families III/V/VII/X as best hits.
salvador	metazoan/holozoan	*Nematostella* has a protein with the correct 2 WW domains and best hits as salvador. No proteins with the correct domains or best hits were found in *Trichoplax*. *Amphimedon* has one protein with best hits to salvador, but has only one ww domain. *Monosiga* has one protein with one ww domain that hits salvador and yap1.
expanded	*Drosophila*-specific	No putative human, *Nematostella*, *Trichoplax*, *Monosiga*, or *Amphimedon* orthologs were found with the FERM+half B41 + PH domain.
Hippo	eukaryotic	Known as CDC15 in yeast, also found in plants, the mechanism of hpo-wts-mats is functional in yeast but used in mitotis exit (MEN) and septation initiation (SIN).[93]
mats	eukaryotic	Known as mob1 in yeast, also found in plants, the mechanism of hpo-wts-mats is functional in yeast but used in mitotis exit (MEN) and septation initiation (SIN).[93]
yorkie/Yap	metazoan/holozoan	*Nematostella* has a protein with the correct 2 WW domains and best hits as salvador. No proteins with the correct domains or best hits were found in *Trichoplax*. *Amphimedon* has one protein with best hits to yorkie, but has only one ww domain. *Monosiga* has one protein with one ww domain that hits salvador and yap1.
fat	holozoan	Fat type kinases with the correct domains found in animals and choanoflagellates.[183]
dRASSF	holozoan	One *Monosiga* protein was found with the ubq domain that has best hits as rassf.

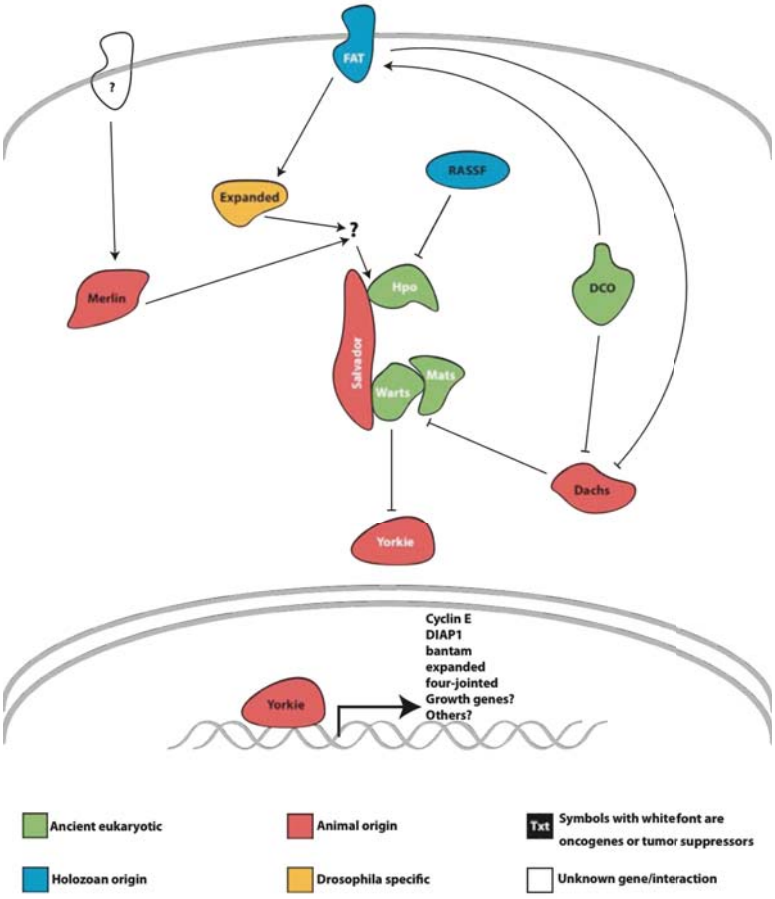


Figure S8.2.7: Schematic of the Warts/Hippo pathway with components colored by node of origin

Table S8.3.1: Classification of apoptosis signaling pathway genes by origin.

Gene	Origin	Methods used to determine origin
Initiator caspases	metazoan	Based on domain configuration, *Amphimedon* has three putative initiator caspases: one protein with two DEDs that groups within the caspase 8/10 clade with two other sponge candidates that lack the characteristic prodomain (poor branch support); and two proteins with a CARD that form a clade with other sponge and cnidarian caspase candidates but could not be assigned to any vertebrate subtypes. Caspases with a pyrin prodomain are restricted to vertebrates (Figure S8.3.1, Figure 2a and Section S9.3.1).
Effector caspases	metazoan	*Amphimedon*, *Nematostella* and *Trichoplax* encode putative effector caspases with a caspase domain only. *Amphimedon* seems to have undergone a lineage-specific expansion. Apart from three caspases that grouped within the caspase 8/10 clade (see notes for initiator caspases above), other sponge candidates could not be assigned to vertebrate subtypes (phylogenetic tree had poor branch support). All *Trichoplax* genes intriguingly group together within the caspase 3/6/7 clade. Lineage-specific expansion has also been observed in *Strongylocentrotus*, *Ciona* and *Branchiostoma* (Figure S8.3.1).[99,100]
CFLAR (FLIP)	vertebrate	No hit found outside of vertebrates.[100] CFLAR-like proteins present in poxvirus that counteract immune system of host.[184,185]
FASLG	chordate	No hit found outside of chordates. Detected in *Branchiostoma* but not in *Strongylocentrotus.*[99,101]
TRAIL	eumetazoan s.s.	No hit found outside of eumetazoans. Detected in *Nematostella* but not in *Trichoplax.* Present in *Branchiostoma* but not in *Strongylocentrotus.*[99,101]
NGFR	eumetazoa s.s.	NGFR is composed of a cysteine-rich TNFR domain and a death domain. This domain combination is only found in eumetazoans[99] with the exception of *Trichoplax*. *Amphimedon* encodes a putative membrane-associated NGFR-like protein with a TNFR domain but no death domain.
FAS	vertebrate	No hit found outside of vertebrates[99,101]. An unusual putative protein has been reported in *Geodia cydonium* that consists of 2 death domains with one domain showing homology to that of FAS however no TNFR or TM domain[97], but no homolog detected in *Amphimedon.* Not detected in *Nematostella* or *Trichoplax.*
TNFRSF1A	vertebrate	No hit found outside of vertebrates. No match in other animals and non-animal groups.
TNFRSF10A-B	vertebrate	No hit found outside of vertebrates. No match in other animals and non-animal groups.
ASK	holozoan	See Manning et al. 2008[127].
JNK	metazoan	See Table S8.5.2 and Table S8.7.2.
MKK/MEK7	metazoan	See Table S8.7.2.
FADD	metazoan	Putative FADD candidates with the characteristic DED-death domain combination encoded in *Amphimedon* and *Nematostella.* The protein candidate encoded in *Trichoplax* has a CARD instead of a DED. No hit outside of animals.
CRADD/RAIDD	eumetazoan	CRADD composed of CARD-death domain combination. No hit in *Amphimedon* or non-animal groups. A putative protein is encoded in *Nematostella* that has a similar domain organization to CRADD but low homology. No hit in *Hydra. Trichoplax* CRADD candidate possesses 2 CARD domains instead of one.
TRADD	vertebrate	No hit found outside of vertebrates.
TLR/ILR-like	deuterostome	No hit found outside of bilaterians. *Amphimedon* and *Nematostella* have putative transmembrane receptors related to TLRs with an Ig and a TIR domain combination[186](Gauthier, in prep). Not detected in *Trichoplax*. See Table S8.10.1.
IRAK	metazoan	See Table S8.10.1.
MyD88	metazoan	See Table S8.10.1.
TRAF	metazoan	See Table S8.10.1.
RIPK	deuterostome	See Bradham et al. 2006[187].
NIK	vertebrate	See Table S8.10.1.
IKK	metazoan	See Table S8.10.1.
NFkB	holozoa	See Table S8.10.1.
Rel	bilaterian	See Table S8.10.1.
14-3-3	eukaryotic	See table S8.5.5.

Bak-like	metazoan	At least one representative detected in *Amphimedon,* other sponges, *Nematostella* and *Trichoplax* (Figure S8.3.2).[98]
Bax-like	eumetazoan	One representative found in *Nematostella*, *Hydra* and *Trichoplax* but none in *Amphimedon* (Figure S8.3.2).
Bok-like	eumetazoan	At least one representative detected in *Nematostella* and *Trichoplax* but none in *Amphimedon* (Figure S8.3.2).
Bcl2-like	metazoan	Bcl2 and Bcl-X are defined by an N-terminal BH4 domain, a Bcl2 domain and a transmembrane region. The BH4 domain is necessary for anti-apoptotic activity. Putative proteins found in *Amphimedon, Nematostella* and *Trichoplax* lack a well-defined BH4 domain (not detected by prediction softwares) but their N-terminal domain weakly aligns to that of other metazoans' Bcl2/Bcl-X. These proteins group either separate to or at the base of the Bcl2/Bcl-X clade and cannot be confidently assigned to particular eumetazoan subtypes (Figure S8.3.2). Lineage-specific expansions have also occurred in other animal groups.[99,100]
Bad	vertebrate	No hit found outside of vertebrates.
Bim	vertebrate	No hit found outside of vertebrates.
Noxa	vertebrate/bilaterian	No non-bilateiran proteins were picked up by BLAST using human NoxA
Bid, tBid	vertebrate	No hit found outside of vertebrates.
Diablo/ Smac	bilaterian	No hit found outside of bilaterians.
cIAP1/BIRC2; cIAP2/BIRC3	vertebrate	No hit found outside of vertebrates.
AIF1	eukaryotic	Present in all holozoan representatives surveyed (except *Trichoplax*) and in *Dictyostellium* (CBP1).
APAF1	metazoan	All animal genomes surveyed have at least one APAF1 representative with conserved domain organization (CARD+NBARC+WD40 domains). *Nematostella, Strongylocentrotus* and *Amphioxus* have an expanded repertoire with proteins displaying additional domain combinations that are not found in vertebrates.[100]
API5	eukaryotic	Ancient protein found in plants, *Dictyostelium* and metazoans (*Amphimedon, Nematostella* and *Trichoplax*).
Akt	eukaryotic	See Table S8.2.2
XIAP/BIRC4	metazoan	BIRC4 is composed of a BIR domain type 1 + RING finger domain. Present in *Amphimedon* and *Nematostella* but not in *Trichoplax*. Related proteins found in baculoviral IAPs. Not detected outside of metazoan group.
BIRC1/NAIP	vertebrate	BIRC1 is composed of 3 BIR domains + NACHT domain. Not detected outside of vertebrate group.
AIFM1/PDCD8	eukaryotic	Found across animal genomes surveyed, as well as *Dictyostelium* and *Monosiga*. *Neurospora* and *Paramecium* only encode the related AIFM3 that possess an N-terminal Rieske domain; *Saccharomyces* has AIF1. Not found in plants.
NLRs	metazoan	See Table S8.10.1.
ICAD/DFFA CAD/DFFB	eumetazoan s.s.	The two subunits of the caspase-activated DNase/DNA fragmentation factor. A (also known as ICAD, inhibitor of CAD, or DFF45) and β (also known as CAD nuclease, or DFF40) are divergent (no significant similarity) despite both having an N-terminal CIDE_N domain. CIDE_N appears to be specific to cnidarian-bilaterian clade. Both DFF subunits are present in *Nematostella* but absent in *Trichoplax* and *Amphimedon*. DFFA is defined by the CIDE_N and DFF-C domain; DFFB is defined by the CIDE_N and DFF40 domains, as found in both *Nematostella* and human.
PARP1 (Poly (ADP-ribose) polymerase family, member 1)	eukaryotic	This is known to be an ancient eukaryotic protein[188]
LMNB1	eumetazoan	The lamin domain is metazoan-specific. However, its association with the C-terminal Intermediate filament tail domain is only found in *Nematostella* and *Trichoplax*. Two LMNB1-like proteins predicted in *Amphimedon* that possess only a lamin domain. Tunicate LMNB1 contains a lamin domain and a truncated C-terminal tail domain.[189]

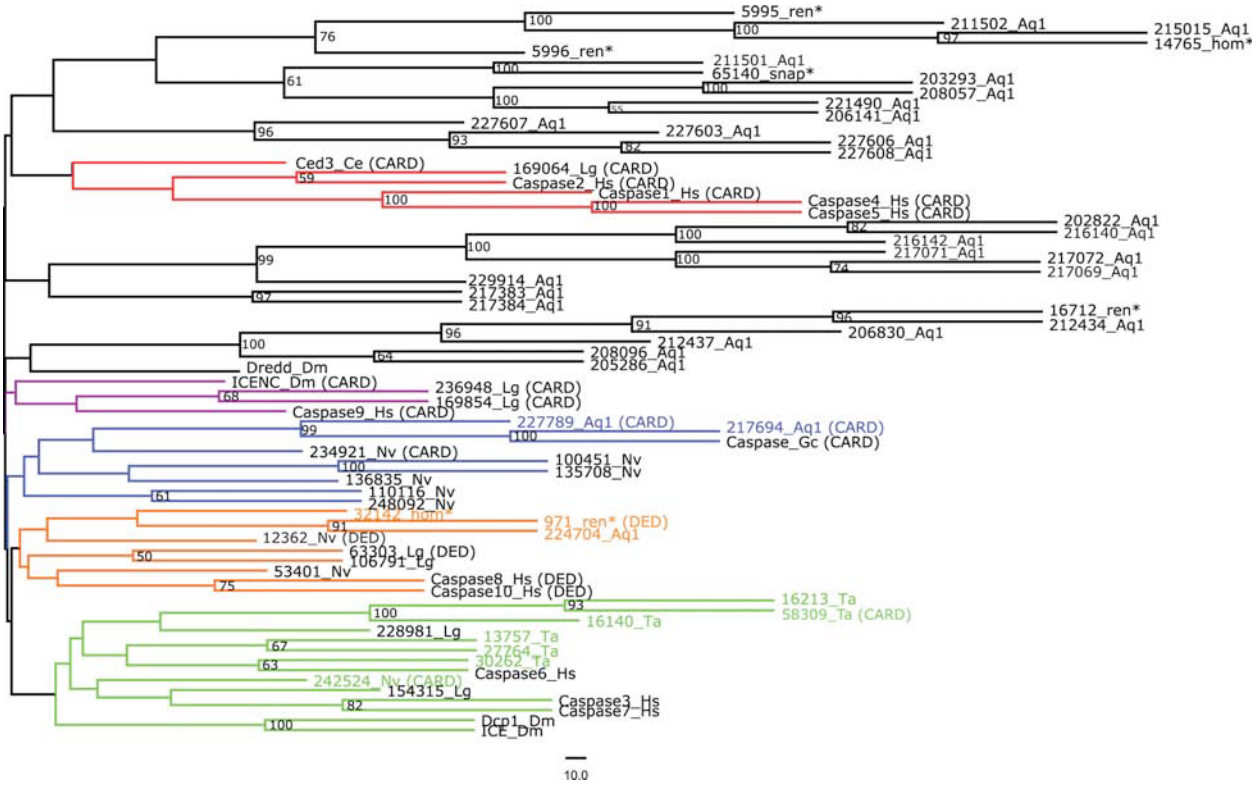


Figure S8.3.1: Rooted neighbor-joining tree for the conserved region of the caspase domain. Mid-point rooting is used. An expansion of the caspase gene family can be observed in the sponge lineage (gene models that clearly correspond to alleles of the same locus were not included, but some might have been overlooked). Only some of the sponge caspases predicted from the genome were used in the phylogenetic analyses. With the exception of three *Amphimedon* caspases (in orange) that group within the caspase 8/10 subtypes, all other sponge caspases could not be reliably assigned to other bilaterian subtypes (tree has poor branch support). *Trichoplax* does not appear to have caspases with prodomains but some models could be partial. All *Trichoplax* genes group together within the caspase 3/6/7 clade (in green); most occur in a cluster in the placozoan genome. *Nematostella* has candidate caspases that clade within the caspase 8/10 and the caspase 3/6/7 subtypes. Putative poriferan and cnidarian caspases with a CARD prodomain, which showed highest BLAST similarity to bilaterian caspase 9, form a separate clade represented in blue that was not recovered within bilaterian subfamilies (in purple and red). One *Nematostella* and *Trichoplax* CARD-containing caspase groups with the caspase 3/6/7 bilaterian subfamily (in green). Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Aq, *Amphimedon queenslandica;* Gc, *Geodia cydonium*; Lg *Lottia gigantea.* Sequences with an asterisk denote *Amphimedon* sequences that are derived from models other than Aqu1. The presence of prodomains (i.e. CARD and DED) is indicated in brackets after the gene identifier.

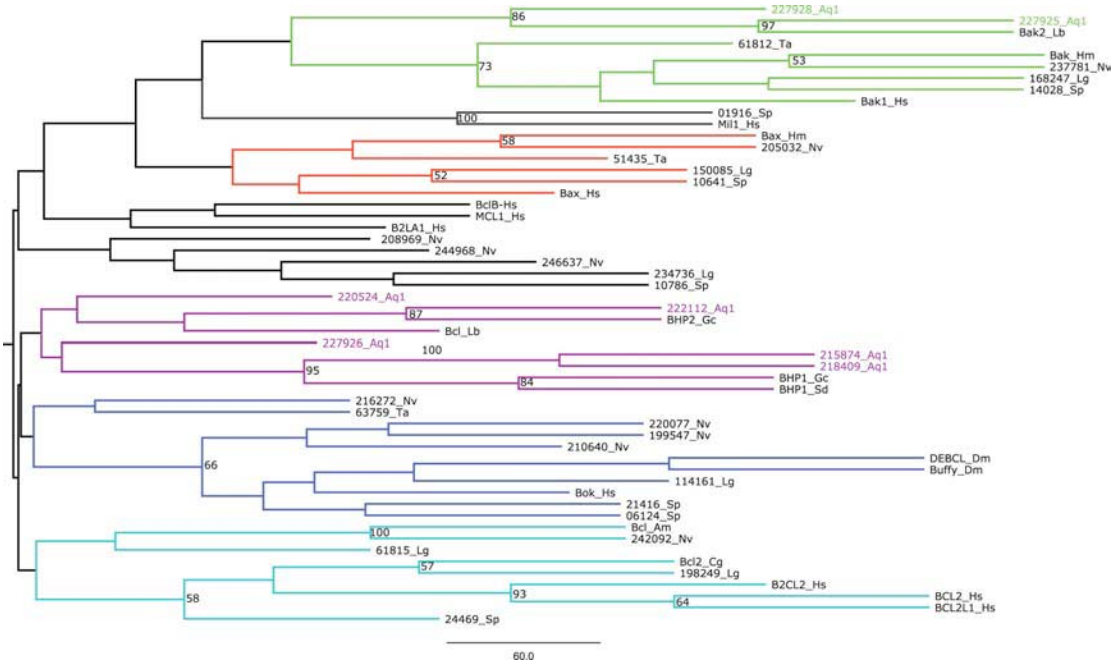


Figure S8.3.2: Rooted Neighbor-joining tree for Bcl2-related proteins. Mid-point rooting is used. The Bcl2-related proteins can be broadly divided into the pro-apoptotic groups (Bak, Bax and Bok) and the anti-apoptotic Bcl2 group. Although these groups were recovered in our analysis, we overall obtained poor branch support. *Amphimedon* has two Bak-like representatives (green), while the Bax-like and the Bok-like gene families are eumetazoan-specific (represented in red and blue respectively). A poriferan gene cluster (in purple), which is related to the anti-apoptotic Bcl2 group (in cyan) and includes five *Amphimedon* genes, forms a clade separate to other bilaterian subtypes. Various other animal Bcl2-related protein members are of uncertain affiliation. Nodes are labelled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Sp *Strongylocentrotus purpuratus,* Lg *Lottia gigantea,* Hm *Hydra magnipapillata*, Lb *Lubomirskia baicalensis*, Gc *Geodia cydonium,* Am *Acropora millepora,* Cg *Crassostrea gigas*, Sd *Suberites domuncula.*

Table S8.4.1: Classification of bilaterian germ-cell specification genes by origin.

Gene	Origin	Method used to determine origin
vasa	metazoan	Metazoan-specific. Present in *Nematostella*, *Amphimedon*, not in *Monosiga*. *Trichoplax* has a single gene in the vasa/PL10 family that is putatively assigned as PL10. Member of DEAD-box protein family, which has representatives throughout all domains of life.
PL10	metazoan	Metazoan-specific. Present in *Nematostella*, *Amphimedon*, not in *Monosiga*. *Trichoplax* has a single gene in the vasa/PL10 family that is putatively assigned as PL10. Member of DEAD-box protein family, which has representatives throughout all domains of life.
nanos	metazoan	Metazoan-specific. Present in in *Trichoplax*, *Nematostella*, *Amphimedon*, not in *Monosiga*. Member of zinc finger protein family that is found throughout eukaryotes.
piwi	eukaryotic	Eukaryote specific. Present in in *Nematostella*, *Amphimedon*, not in *Trichoplax* or *Monosiga*. See Grimson et al[109] for further details.
mago nashi	eukaryotic	Known to be ancient eukaryotic protein. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
tudor	eukaryotic	Known to be ancient eukaryotic protein. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
pumilio	eukaryotic	Known to be ancient eukaryotic protein. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
PAR-1	eukaryotic	Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*. Includes serine/threonine protein kinase catalytic domain that is found throughout all domains of life.
DMRT	eumetazoan	Eumetazoan-specific. Contains the metazoan-specific cysteine-rich DM DNA-binding domain first identified in doublesex (*D. melanogaster*) and mab-3 (*C.elegans*). Not in *Amphimedon* or *Monosiga*, putatively present in *Trichoplax* and *Nematostella*.

Table S8.5.1: Classification of Wnt signaling pathway genes by origin.

Gene	Origin	Method used to determine origin
WNT	metazoan	3 *Amphimedon Wnts* cannot be classified into particular family and do not look like lineage-specific duplicates - not possible to tell if ancestor had 1, 2, or 3 genes (Adamska et al., in preparation).
Frizzled (Fzd)	metazoan	2 *Amphimedon Fzd* cannot be classified into particular family and do not look like lineage-specific duplicates - not possible to tell if ancestor had 1 or 2. 2 Fzd-like GPCRs in *Dictyostelium* have SRC-like domains and KTXXXW motifs (Fzd-specific) according to Prabhu and Eichinger 2006[190] but their sequences are very divergent (*smo* sequences are more similar to Fzd than are these *Dictyostelium* GPCRs).
sfrp	eumetazoan *s.s.* or metazoan	Present in *Amphimedon* but may be convergent evolution as no netrin domain; presence in *Nematostella* with additional netrin domain indicates at least eumetazoan invention (Adamska et al., in preparation). Not found in *Trichoplax*.
LRP5/6	metazoan	Subfamily found in animals.
APC	bilaterian	*Amphimedon, Nematostella, Trichoplax* genes are most similar to bilaterian *APC*s but are missing certain domains. AmqAPC only has 7 armadillo repeats and is missing GSK3, Axin and β-catenin binding motifs and a PDZ-ligand motif at the C-terminal.
Axin	Animal	*Amphimedon* and *Nematostella* genes have all the axin domains except the β-catenin binding domain. Closest hits in *Dictyostelium* and *Monosiga* have different domain structures and are most similar to animal *RGS* genes.
GSK3	Eukaryotic	Ancient eukaryotic family.
dishevelled	metazoan	Multidomain protein; the domains have different origins: Dvl domain metazoan, PDZ eukaryotic, DEP eukaryotic, Dix/Dax metazoan. *Trichoplax* Dvl is not predicted accurately.
β-catenin	metazoan	Armadillo/β-catenin repeat is present outside of Metazoa but the specific domain combination of β-catenin is metazoan-specific. *Dictyostelium* Aardvark and plant Arabidillo proteins have different domain structures (additional F-box, missing C-terminal PDZ transactivation domain, less armadillo repeats).
TCF/LEF	metazoan	HMG domain is ancient eukaryotic but the TCF/LEF subfamily is metazoan-specific. The N-terminal β-catenin binding motif is metazoan-specific. See Figure S8.5.1. Closest *Monosiga* matches are capicua-like.
CBP/P300	metazoan	*CBP* genes are in *Amphimedon* and *Nematostella* (*CBP/P300* distinction only in vertebrates); *CBP*-like genes are found in *Monosiga* and *Trichoplax*, which appear to be missing the final CREB-binding domain; *CBP*-like genes are found in plants (missing multiple domains) but not in fungi[191].
CK1	eukaryotic	Ancient eukaryotic family.
groucho	metazoan	Multidomain protein; the domains have different origins: WD40 is ancient eukaryotic but other regions are metazoan-specific.
WIF	bilaterian	Absent from *Amphimedon, Nematostella*, and *Trichoplax.*

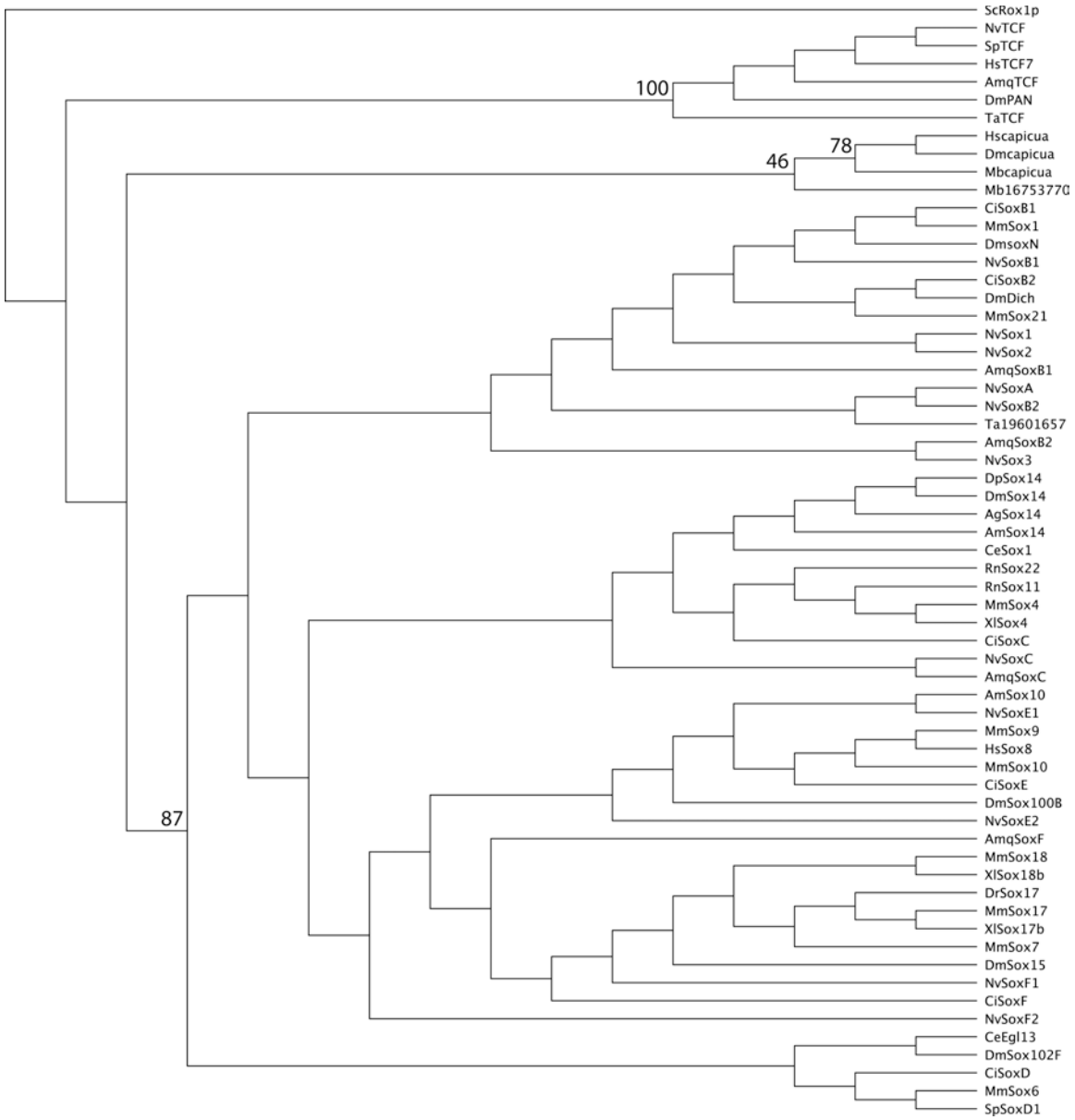


Figure S8.5.1: Rooted neighbor-joining tree of TCF, capicua, and Sox HMG genes. The tree is rooted with the fungal gene *ScRox1p*. *Amphimedon*, *Nematostella*, and *Trichoplax* have clear *TCF* genes while the closest BLAST matches to *TCF*s in the *M. brevicollis* genome are conclusively *capicua* genes. Nodes of interest are labeled with bootstrap values (100 replicates). Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Amq, *Amphimedon queenslandica*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*; Sc, *Saccharomyces cerevisiae*; Sp, *Strongylocentrotus purpuratus*; Hs, *Homo sapiens*; Mb, *Monosiga brevicollis*; Rn, *Rattus norvegicus*; Mm, *Mus musculus*; Dp, *Drosophila pseudoobscura*; Ag, *Anopheles gambiae*; Xl, *Xenopus laevis*; Am, *Apis mellifera*; Ce, *Caenorhabditis elegans*; Dr, *Danio rerio*.

Table S8.5.2 : Classification of TGF-β signaling pathway genes by origin.

Gene	Origin	Method used to determine origin
TGF-β pathway RI (STKR1)	metazoan	TGF-β Kinase, GS, and N-terminal ligand binding domain ("Activin/TGF-β domain" PF01064) present in *Nematostella* and Bilateria. In *Amphimedon*, TGF-β Kinase and GS domains are present, and although the N-terminal domain is not located by domain prediction software, a cysteine rich box (ligand binding motif) is present. Eumetazoan receptors group into 3 main subfamilies - ACTVR1; BMPR1; TGF-βR1 - *Amphimedon* and *Trichoplax* genes do not fall into these classes (Fig S8.5.2).
TGF-β pathway RII (STKR2)	metazoan	TGF-βKinase and N-terminal ligand binding domain ("Activin/TGF-β domain" PF01064) present in *Nematostella* and Bilateria. In *Amphimedon*, TGF-β Kinase domain is present, and although the N-terminal domain is not located by domain prediction software, a cysteine-rich box (ligand binding motif) is present. Eumetazoan receptors group into 3 main subfamilies - ACTVR2; BMPR2; TGF-βR2 (Vertebrates only) - *Amphimedon* and *Trichoplax* genes do not fall into these classes (Fig S8.5.2).
TGF- β pathway ligands	metazoan	Multiple genes with domain configurations diagnostic of TGF-β pathway ligands are present in *Amphimedon* and *Trichoplax*, but cannot be assigned to particular eumetazoan subfamilies with confidence. *Nematostella* possesses ligands related to: BMP2/4; BMP5/8; ACTV; GDF; ADMP; Myostatin (Fig S8.5.3).
Smad1/5	metazoan	Metazoan Smads have a MH1 domain and a MH2 domain. *Monosiga* has a Smad-like protein with a C2H2 zinc finger and an MH2 domain (Fig S8.5.4).
Smad2/3	metazoan	(Fig S8.5.4)
Smad4	metazoan	(Fig S8.5.4)
Smad6/7	eumetazoan	Not present in *Amphimedon* (Fig S8.5.4).
FOS	metazoan	bZIP are ancient eukaryotic genes but diversification occurred independently in the different major lineages. *Amphimedon* and *Nematostella* probably have a *Fos/ATF3* gene rather than a Fos gene. Not found in *Trichoplax* (Fig S8.5.5).
JUN	metazoan	bZIP are ancient eukaryotic genes but diversification occurred independently in the different major lineages. Not found in *Trichoplax* (Fig S8.5.5).
MYC	metazoan	bHLH are ancient eukaryotic genes but diversification occurred independently in the different major lineages. A *Myc*-like gene is present in Monosiga (see tree) but has no Myc domain.
MAX	holozoan	bHLH are ancient eukaryotic genes but diversification occurred independently in the different major lineages[192]. *Max* is present in *Monosiga* (Fig S8.5.6).
SMURF1	metazoan	Smurf is an animal-specific subfamily of HECT E3 ubiquitin ligases. Found in *Trichoplax*, *Nematostella* and *Amphimedon* (Fig S8.5.7).
TGF-β RIII	bilaterian	TGF-βR type III are bilaterian specific. Hits in *Nematostella* and *Trichoplax* are to other genes containing the ZP domain. No ZP domain present in *Amphimedon* or *Monosiga* or any non-holozoan.
ZFYVE9 (SARA)	eumetazoan	SARA is a eumetazoan subfamily of ancient FYVE-type zinc finger domain containing proteins. Present in *Trichoplax* and *Nematostella*. Absent from *Monosiga* and *Amphimedon*.
Noggin	metazoan	Present in *Trichoplax*, *Nematostella*. And *Amphimedon.* No hits in *Monosiga*.
Chordin	eumetazoan *s.s.*	Present in *Nematostella*. No hits to CHRD domain in *Amphimedon* or *Monosiga*.
Cerebrus/DAN (CAN family)	eumetazoan	Multiple CAN family members in *Hydra* and *Nematostella*; single ortholog in *Trichoplax*. No matches in *Amphimedon* or *Monosiga.*
Follistatin	eumetazoan	Present in *Nematostella* and *Trichoplax*. No matches in *Monosiga* or *Amphimedon*.
Ski/Sno	eumetazoan	Present in *Hydra*, *Nematostella* and *Trichoplax*. Not in *Monosiga*. *Amphimedon* top match contains the Smad4 binding domain only – it lacks the N-terminal DNA binding domain.
JNK	metazoan	No JNK-like kinases were found in non-animal genomes.
MEK7	metazoan	See Table S8.7.2.
MEKK2	holozoan	See Manning et al. 2008[127]
ASK	holozoan	See Manning et al. 2008[127]
P38 (MAPK11-14)	opisthokont	See Manning et al. 2008[127]

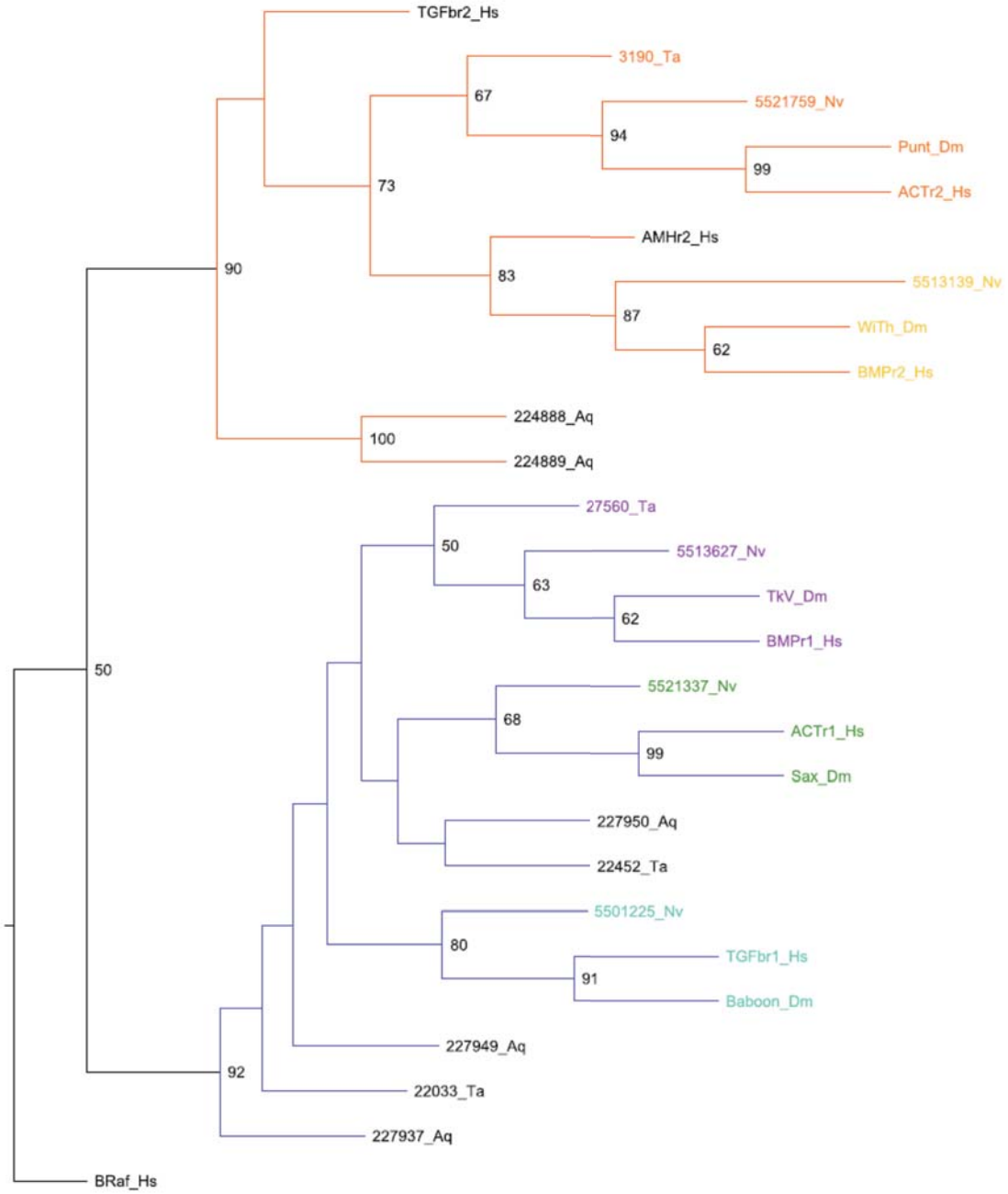


Figure S8.5.2: Unrooted neighbor-joining tree for TGF-β receptors. Representatives of Type I (blue) and Type II (red) receptors are common to all Metazoa. Receptor subtypes: Activin I, II (green, orange); BMP I, II (purple, yellow); TGF-β I (aqua); appear to be eumetazoan-specific subfamilies. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*.

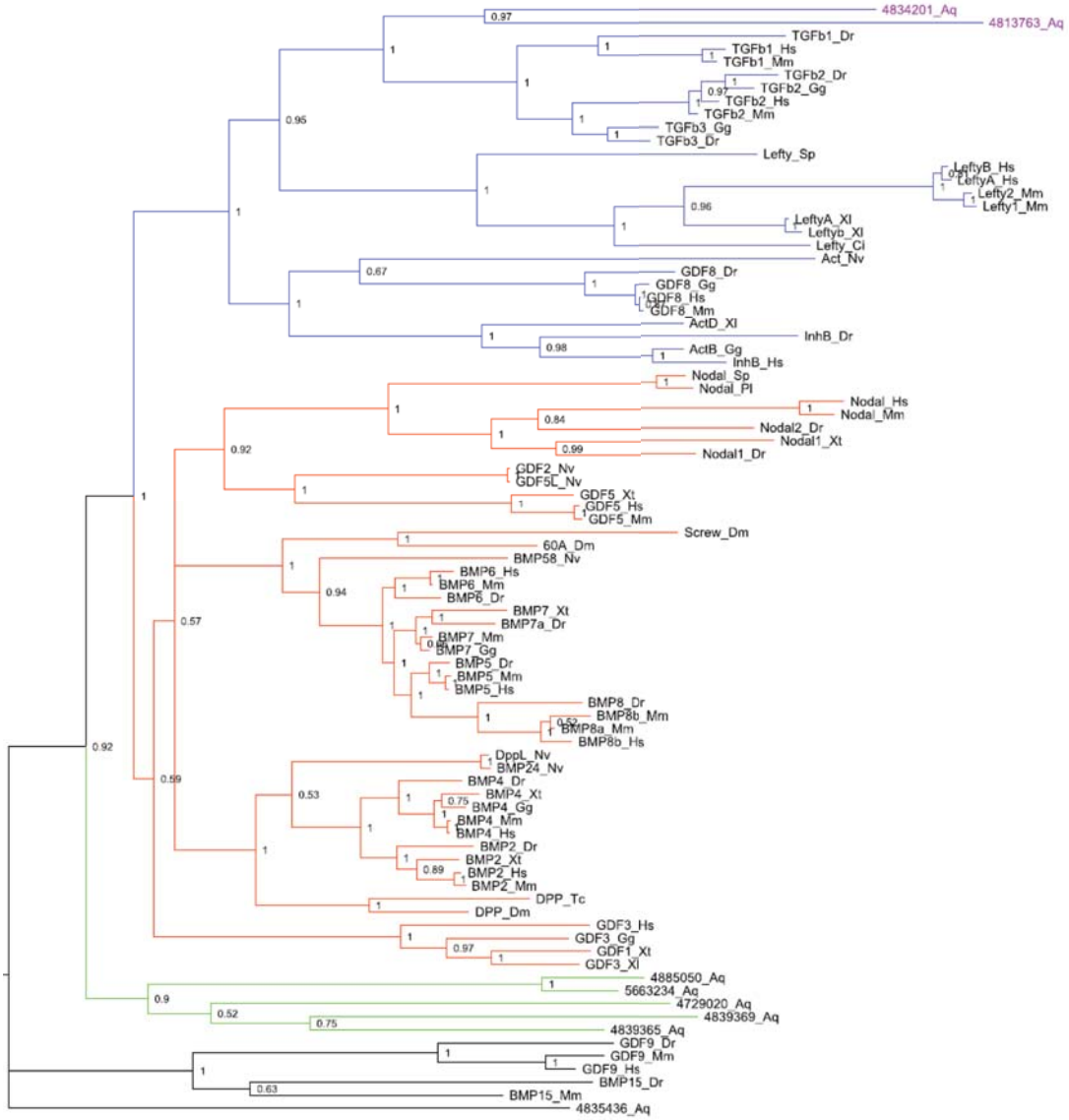


Figure S8.5.3: Bayesian tree for TGF-β ligands. Most of the TGF-β pathway ligands group into two major clades, TGF-β related (blue) or BMP related (red). 2 *Amphimedon* ligands clade within the TGF-β related group (purple). 5 other *Amphimedon* ligands appear to represent a lineage specific expansion (green), and are located outside the two major ligand clades, along with other divergent ligands such as GDF9 and BMP15. Nodes are labeled with posterior probability proportions, analysis was run for 4 million generations. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Mm, *Mus musculus*; Gg, Gallus gallus; Sp, *Strongylocentrotus purpuratus*; Nv, *Nematostella vectensis*; Aq, *Amphimedon queenslandica*; Xl, *Xenopus laevis*; Xt, *Xenopus tropicalis*; Ci, *Ciona intestinalis*; Pl, *Paracentrotus lividus*; Tc*, Tribolium castaneum*.

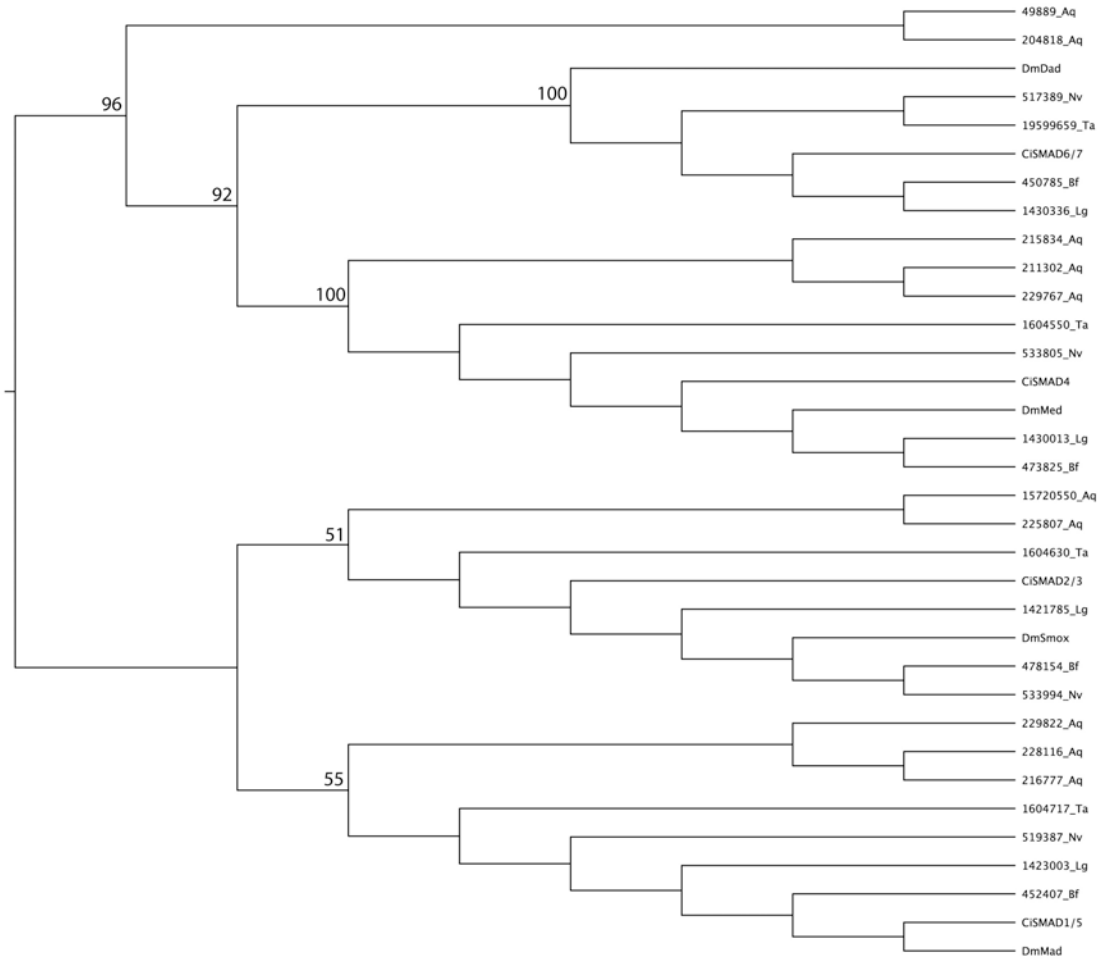


Figure S8.5.4: Unrooted neighbor-joining tree of Smad genes. *Amphimedon*, *Nematostella*, and *Trichoplax* have *Smad4*, *Smad2/3* and *Smad1/5* genes. Interestingly, the latter two taxa only have one representative of each, but *Amphimedon* has 2-3 representatives of each subfamily. *Nematostella* and *Trichoplax* also have *Smad6/7* genes (inhibitory Smads). Two *Amphimedon* genes are of uncertain affiliation; they are at the base of a clade of Smad4 and Smad6/7 subfamilies. Nodes are labeled with bootstrap values (100 replicates), only values >50 are shown. Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*; Bf, *Branchiostoma floridae*; Lg, *Lottia gigantea*.

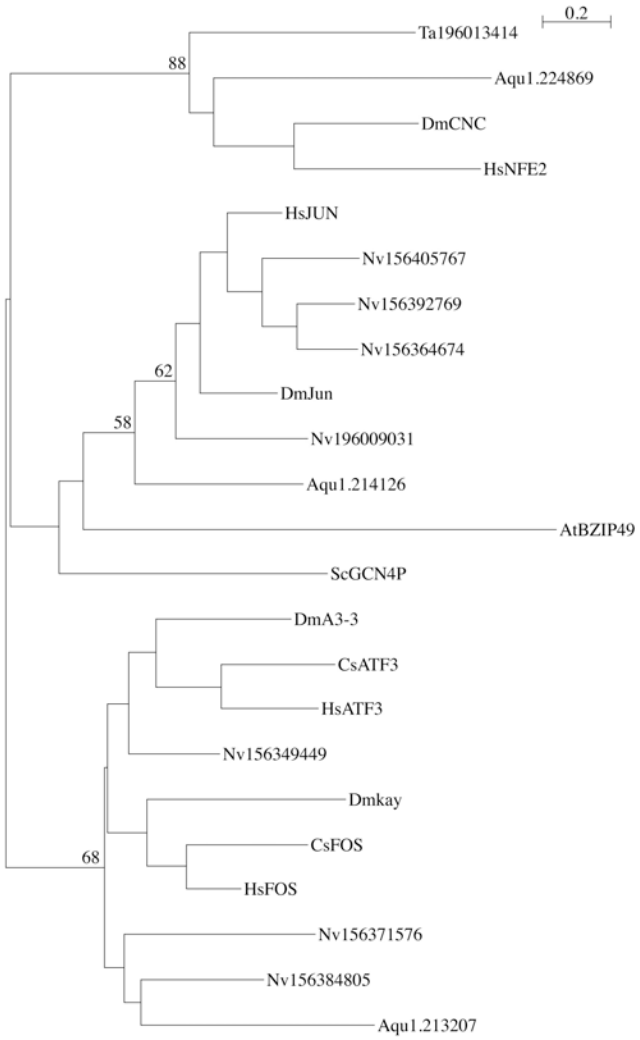


Figure S8.5.5: Rooted neighbor-joining tree of Fos and Jun bZIP genes. The tree was rooted with *NF-E2* genes. *Amphimedon* and *Nematostella Jun*-like genes belong to a metazoan family of *Jun* genes. *Amphimedon* and *Nematostella Fos*-like genes do not clearly belong to this family but rather are probably descendants of an ancestral *Fos/ATF3* gene that gave rise to the two families in bilaterians. The positions of the plant and fungi *bZIP* genes, which are most similar to metazoan *Jun* and *Fos* genes, are unresolved but they do not seem to belong to either metazoan clade. No *Trichoplax Fos*, *ATF3*, or *Jun* gene was detected; the closest match seems to be a *NF-E2* gene. Nodes are labeled with bootstrap values, only values >50 are shown. Mb, *Monosiga brevicollis*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Sp, *Strongylocentrotus purpuratus*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*, Cs, *Ciona savigny*; At, *Arabidopsis thaliana*; Sc *Saccharomyces cerevisiae*.

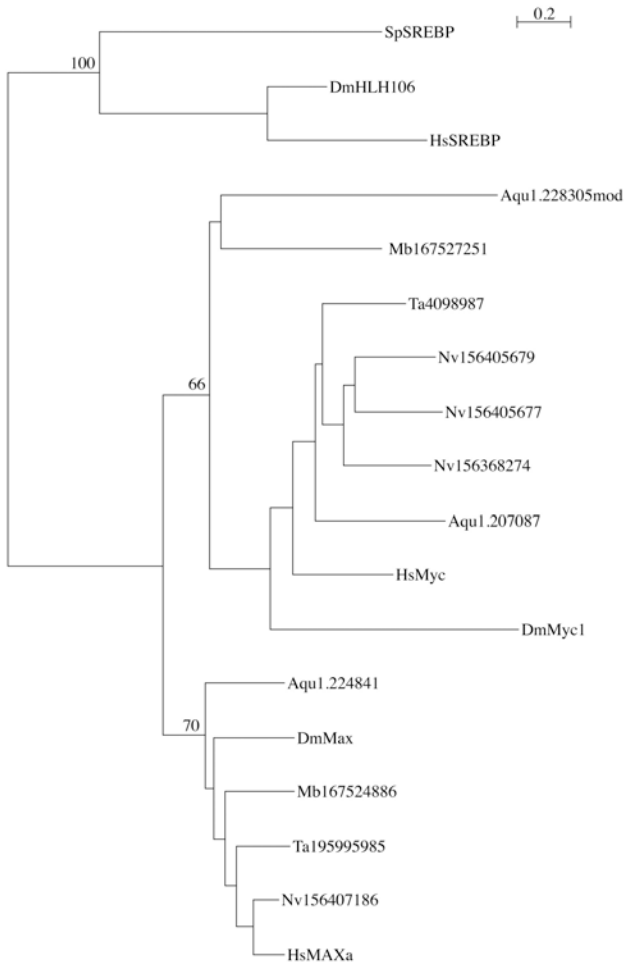


Figure S8.5.6: Rooted neighbor-joining tree of Myc and Max bHLH genes. As *Myc* and *Max* are known to be sister families, the tree was rooted with *SREBP* genes. *Monosiga*, *Amphimedon*, *Trichoplax*, and *Nematostella* genes fall into both the *Myc* and the *Max* clade, suggesting they all have these two genes. However, the *Monosiga* putative *Myc* gene does not have the Myc domain and does not fall into this family in other types of analyses (not shown). Hence, it is likely not a true *Myc* gene. Nodes are labeled with bootstrap values (100 replicates), only values >50 are shown. Mb, *Monosiga brevicollis*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Sp, *Strongylocentrotus purpuratus*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*.

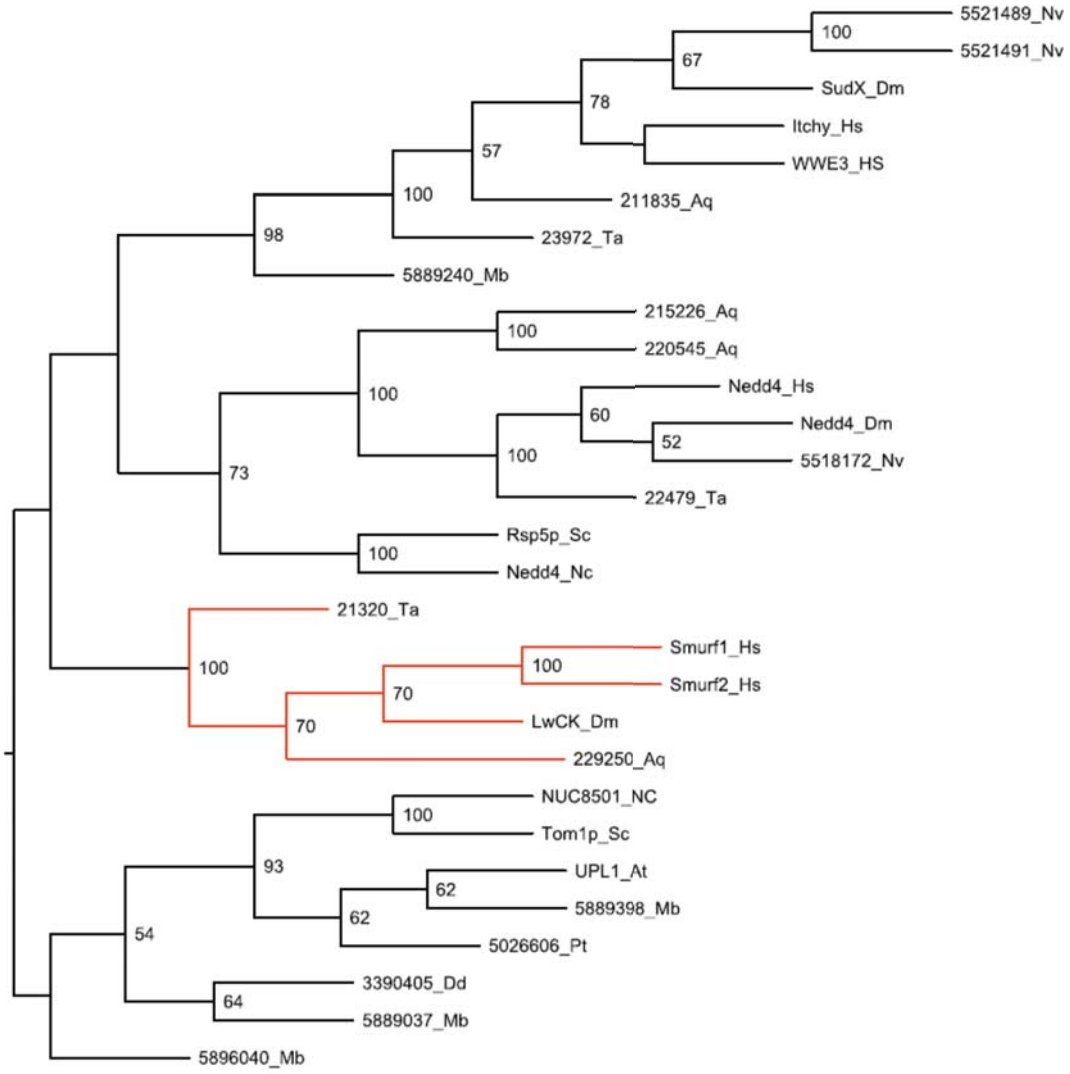


Figure S8.5.7: Unrooted neighbor-joining tree for E3 ubiquitin ligases with HECT domains. The Smurf subfamily (red) appears to be metazoan-specific grouping within the pan-eukaryotic family of E3 ubiquitin ligases. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*; At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Nc, *Neurospora crassa*; Pt, *Paramecium tetraurelia*; Sc *Saccharomyces cerevisiae*.

Table S8.5.3: Classification of Hedgehog signaling pathway genes by origin.

Gene	Origin	Method used to determine origin
Suppressor of Fused	metazoan	No hits outside animals, no hit in *Trichoplax*. Sufu domain found in single Bacteria - *Campylobacter curvus.*
Hedgehog	Eumetazoan *s.s*	Eumetazoan multidomain protein. Hh domain is holozoan-specific, but found in different configurations outside Eumetazoa – e.g. Hedgling proteins in *Amphimedon*, *Monosiga*. Not found in *Trichoplax*. Partial HintN domain in *Paramecium*.
Patched	holozoan subfamily	Patched genes are a holozoan subfamily of ancient sterol-sensing domain multi TM genes. Not in *Amphimedon*, putatively present in *Monosiga*. Sponge closest match more related to Niemann Pick-C subfamily (Fig S8.5.8).
Dispatched	holozoan subfamily	Dispatched genes are a holozoan subfamily of ancient sterol-sensing domain multi TM genes. Not in *Amphimedon*, putatively present in *Monosiga*. Sponge closest match more related to Niemann Pick-C subfamily (Fig S8.5.8).
Smoothened (Smo)	eumetazoan *s.s.* subfamily	Smo is a eumetazoan family of receptors, related to the FZD superfamily. Present in *Nematostella*, not in *Trichoplax* or *Amphimedon* (Fig S8.5.9).
Ihog/CDO	metazoan	Ig cell adhesion molecule. Models with similar domain arrangements (3-5 Ig; 2-3 FNIII) are present in *Amphimedon, Trichoplax* and *Nematostella* genomes, but these cannot be assigned to particular (IgCAM) subfamilies.
RPL29	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
HhIP	eumetazoan *s.s.*	Domains are found throughout the Eukaryota, but only deuterostomes have the HhIP configuration (Folate receptor-Gluc dehydrogenase-EGF). *Nematostella* has a model which lacks the EGF domains but otherwise is highly similar. EGF domains of HhIP are not involved in Hedgehog inhibition[193], so *Nematostella* Hhip-like protein may still function in signalling. No Hhip in *Drosophila* or *Caenorhabditis*.
Costal	metazoan subfamily	Kinesins are known to be ancient eukaryotic proteins. Human orthologs of *Drosophila* Costal = Kif27; Kif7. Present in *Nematostella*, *Trichoplax*, *Amphimedon* (Fig S8.5.10).
Gli	metazoan superfamily	Gli is an animal superfamily of ancient C2H2 zinc finger proteins. Present in *Nematostella* and *Amphimedon*, absent from *Trichoplax* (Fig S8.5.11).
Fused	eukaryotic subfamily	Fused belongs to a eukaryotic family of serine/threonine kinases. Present in *Nematostella*, *Trichoplax*, *Amphimedon, Monosiga.*
Rasp	eumetazoan *s.s.* subfamily	Member of the pan-eukaryotic MBOAT family, not found in *Amphimedon* or *Trichoplax*. Present in *Nematostella*.

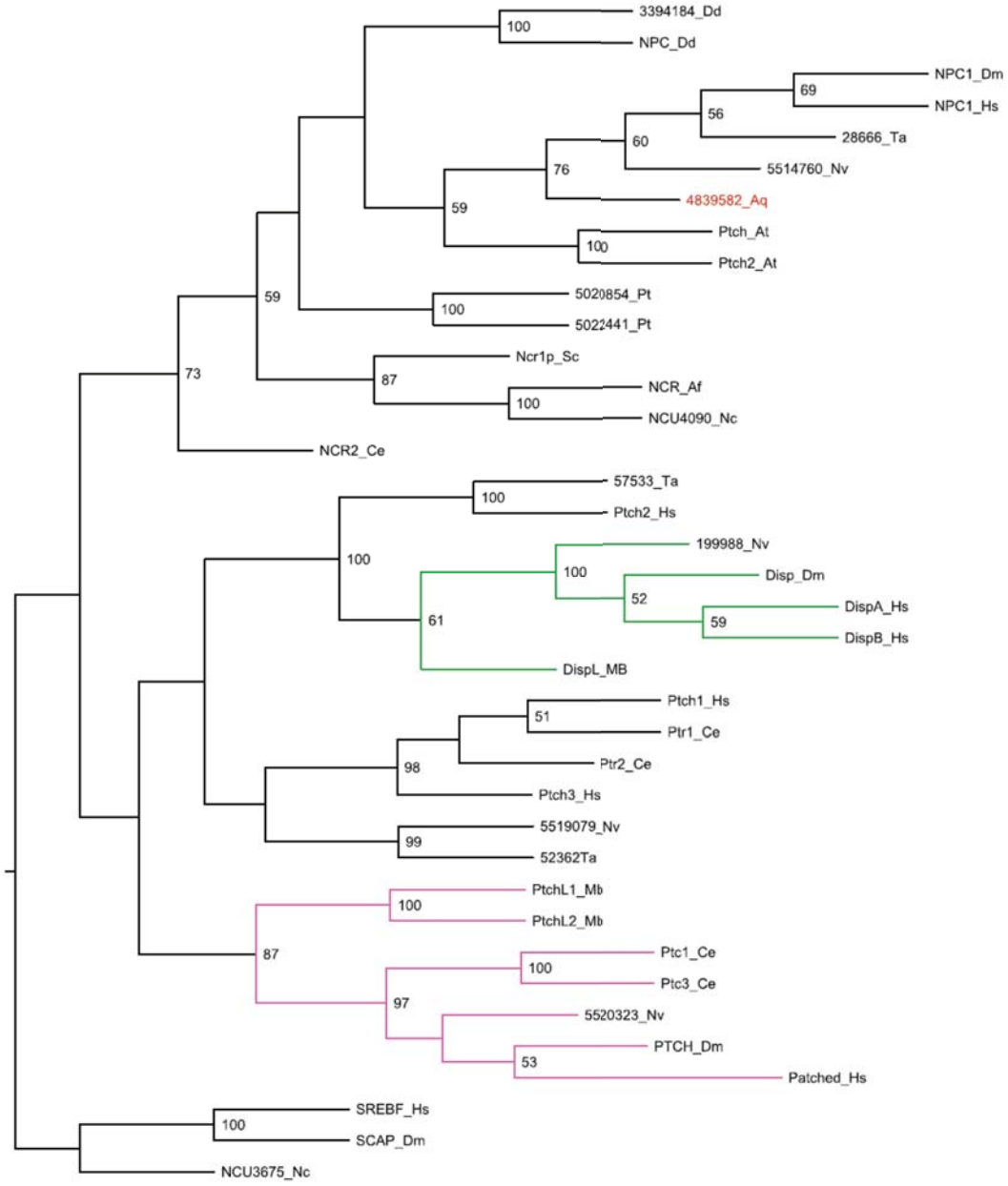


Figure S8.5.8: Unrooted neighbor-joining tree for Patched-related genes. Patched receptors are a holozoan subfamily within the ancient sterol-sensing domain (SSD) family of receptors. One *Amphimedon* gene (red) contains a SSD, but it is excluded from both the Patched (pink) and Dispatched (green) subfamilies, which are the only clades implicated in Hedgehog binding. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Hm, *Hydra magnipapillata*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*. At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Sc, *Saccharomyces cerevisiae*; Ns, *Neurospora crassa*; Pt, *Paramecium tetraurelia*; Af, *Aspergillus fumigatus*.

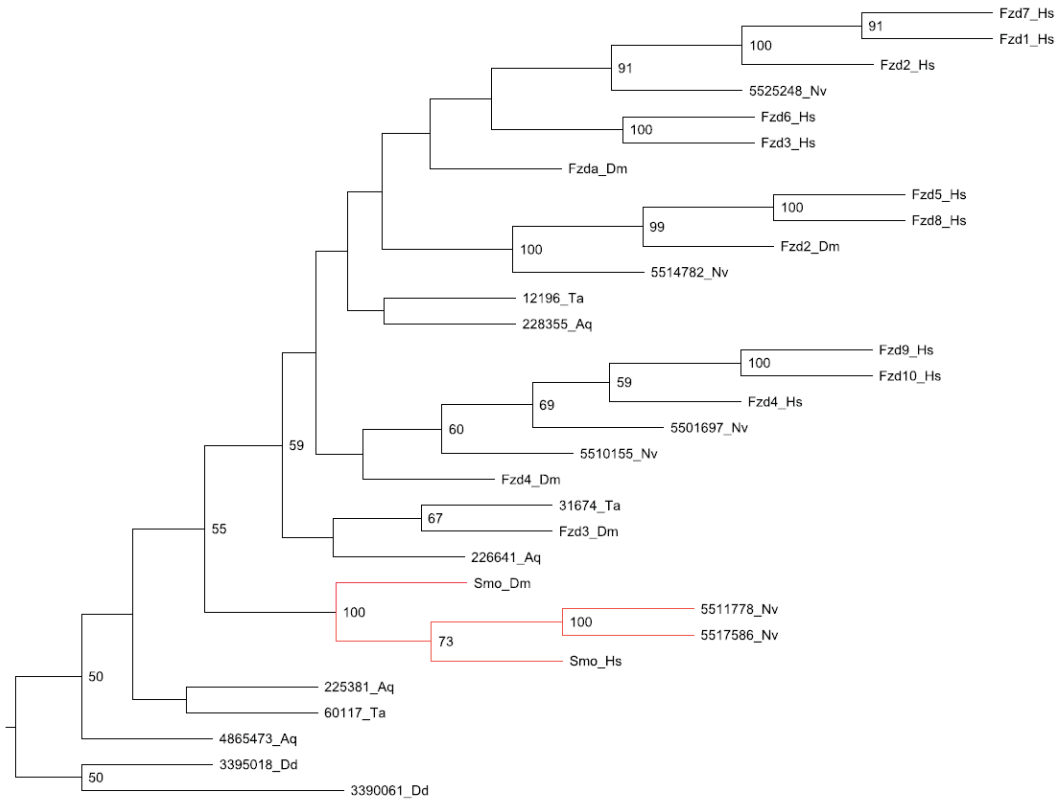


Figure S8.5.9: Unrooted neighbor-joining tree for Frizzled-related genes. Smoothened (red) appears to be a eumetazoan-specific gene family, related to the larger metazoan Frizzled family of ancient G-protein coupled receptors. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*. Dd, *Dictyostelium discoideum*.

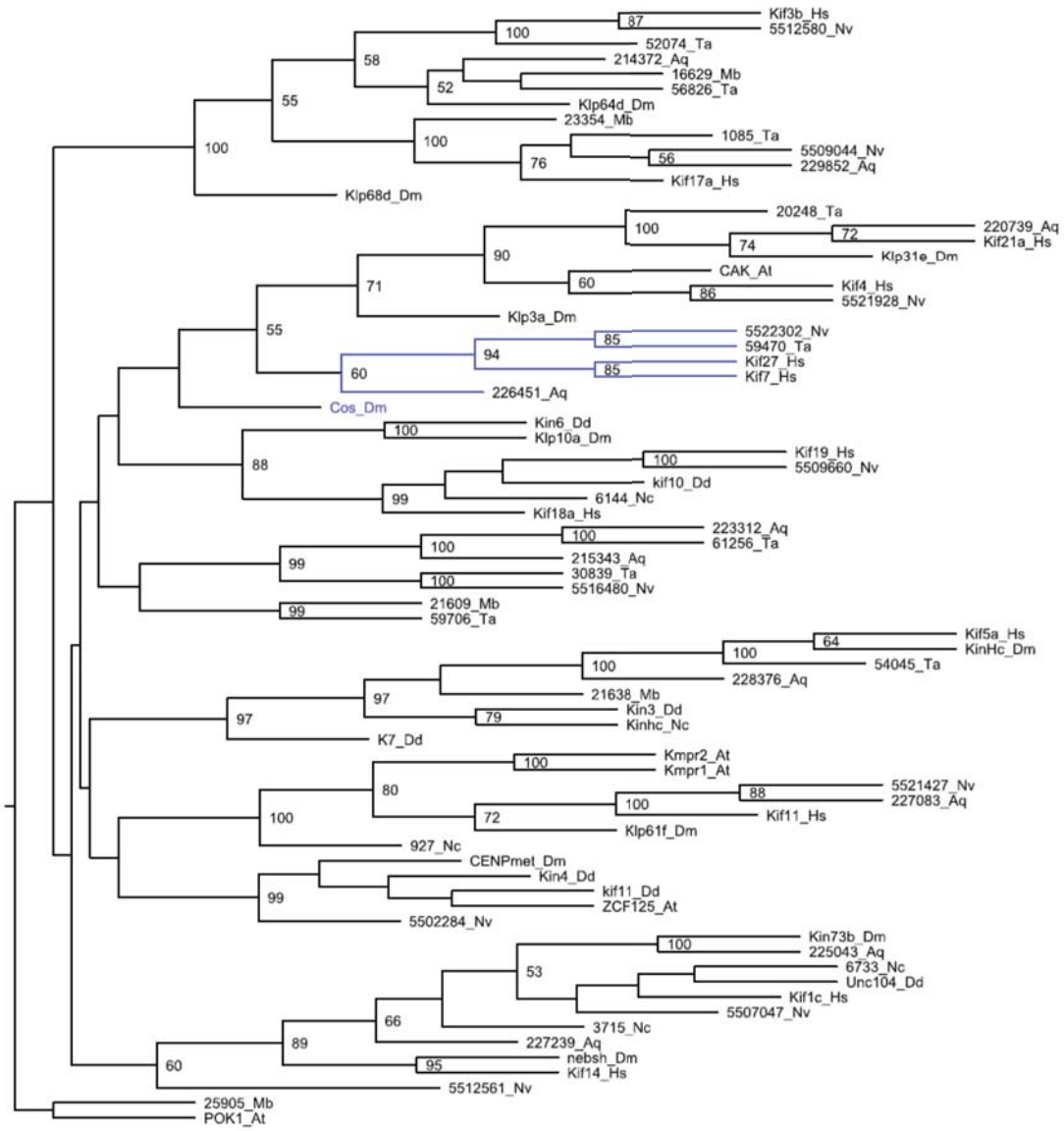


Figure S8.5.10: Unrooted neighbor-joining tree for the Kinesin family. Within the ancient kinesin family, Kif7/Kif27 belongs to an animal-specific subfamily (blue). The *Drosophila* ortholog *Costal* is a divergent ortholog of Kif7/Kif27, and is not recovered within the subfamily in this analysis. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*.;At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Nc, *Neurospora crassa*.

Figure S8.5.11: Unrooted neighbor-joining tree for C2H2 zinc fingers. The Gli transcription factor superfamily (blue) appears to be an animal-specific clade within the ancient family of C2H2 zinc finger binding proteins. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*; At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*.

Table S8.5.4: Classification of Notch signaling pathway genes by origin.

Gene	Origin	Method used to determine origin
Notch	metazoan	Metazoan multidomain protein comprised of ancient pan-eukaryotic domains (EGF, ANK, LNR) with additional eumetazoan-specific domains (Nod, Nodp). Present in *Amphimedon*, *Trichoplax*, *Nematostella*.
FURIN	holozoan subfamily	Furin is a subfamily of the pan-eukaryotic subtilisin family. Identified by the additional presence of Furin repeats at C-terminal of proteins. Present in *Monosiga*, *Trichoplax*, *Amphimedon*.
Fringe	metazoan	No hits outside of animals. No hit in *Trichoplax*. Related to the B3GLT superfamily of glycosyltransferases also found in plants (Fig S8.5.12).
Delta	metazoan	Metazoan multidomain protein comprised of ancient pan-eukaryotic domains (EGF) and metazoan-specific domains (DSL, MNLL). Present in *Amphimedon*, *Trichoplax*, *Nematostella*.
Jagged	eumetazoan *s.s.* subfamily	Eumetazoan multidomain protein comprised of ancient pan-eukaryotic domains (EGF) and metazoan-specific domains (DSL, VWC, MNLL). Present in *Nematostella*, not in *Amphimedon* or *Trichoplax*. May represent a subfamily of Delta ligands.
Presenilin	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
ADAM10/17	opisthokont	Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*. ADAM genes identified in Fungi are most similar to ADAMs10/17s.
Numb	bilaterian	Comprised of holozoan PTB domain and bilaterian specific Numb domain. No hits to Numb domain in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*. Partial Numb domain in *Hydra*.
HDAC	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
SIN3A	eukaryotic	Known to be ancient eukaryotic proteins. Present in *Trichoplax*, *Nematostella*, *Monosiga*, *Amphimedon*.
CSL	metazoan	Present in *Amphimedon*, *Trichoplax*, *Nematostella*. Lacks the β-trefoil domain (Notch binding) in *Monosiga*.
Nicastrin	eukaryotic	Present in *Nematostella*, *Trichoplax* and *Amphimedon*, *Arabidopsis*, *Dictyostelium*.
CBFβ	metazoan	Present in *Nematostella*, *Trichoplax* and *Amphimedon*. No hits outside of animals.
o-fucosyltransferase	holozoan	Present in *Monosiga*, *Trichoplax* and *Amphimedon*. No hits outside of holozoans.
Mastermind	eumetazoan *s.s.*	Present in *Nematostella*. No hits to Maml domain in *Amphimedon*, *Trichoplax*, or outside the Metazoa.

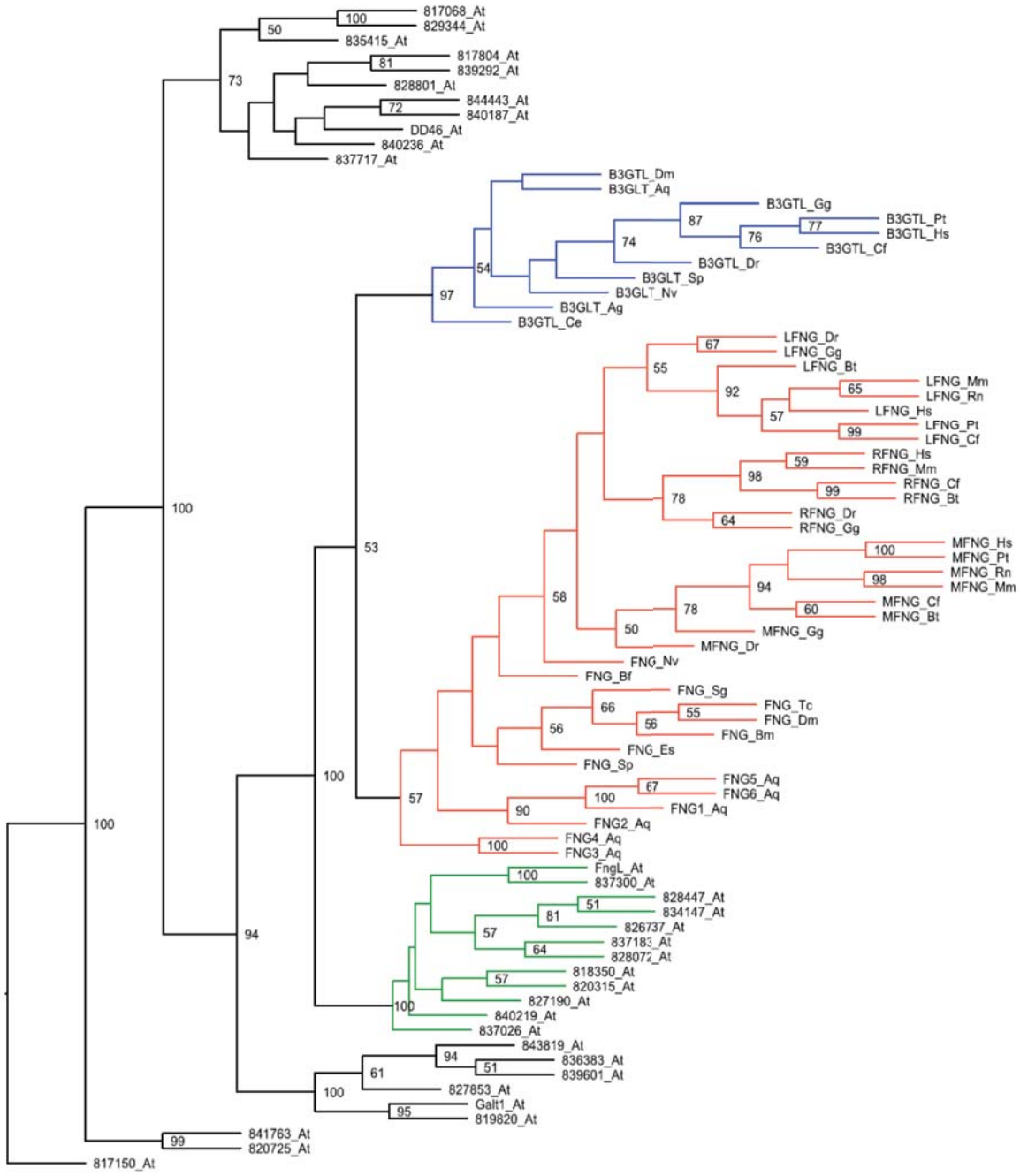


Figure S8.5.12: Unrooted neighbor-joining tree for Fringe and β3GLT. Metazoan Fringe (red) and β3GLT (blue) proteins form a monophyletic clade to the exclusion of plant Fringe-related (green) and other members of the B3GLT family. Nodes are labeled with bootstrap values, only values >50 are shown. Ag, *Anopheles gambiae*; Aq, *Amphimedon queenslandica*; At, *Arabidopsis thaliana*; Bf, *Branchiostoma floridae*; Bm, *Bombyx mori*; Bt, *Bos Taurus*; Ce, *Caenorhabditis elegans*; Cf, *Canis familiaris*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Es, *Euprymna scolopes*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Nv, *Nematostella vectensis*; Pt, *Pan troglodytes*; Rn, *Rattus norvegicus*; Sp, *Strongylocentrotus purpurtus*; Tc, *Tribolium castaneum*.

Table S8.5.5: Classification of growth factor, GPCR, and Ras signaling genes by origin.

Gene/Family	Origin	Methods used to determine origin
EGF	metazoan	Present in *Amphimedon*, no clear ortholog in *Monosiga*.
FGF	eumetazoan *s.s.*	Multiple copies present in *Nematostella*, no hits in *Trichoplax*, *Amphimedon* or *Monosiga*.
PDGF	eumetazoan *s.s.*	Present in *Nematostella*, not found in *Amphimedon*, *Trichoplax* or *Monosiga*.
GPCRs		See supplemental section S8.9
Integrin α	metazoan	N-terminal integrin α repeats are found outside Metazoa but the integrin α 2 domain and cytoplasmic conserved motif/pattern are metazoan-specific.
Integrin β	metazoan	Hits to the three integrin β-specific domains are not found outside Metazoa. *Amphimedon*, *Trichoplax* and *Nematostella* candidates contain most or all of these domains.
Discoidin domain receptor (DDR)	eumetazoan or metazoan	Proteins with a domain architecture resembling bilaterian DDRs were only found in *Nematostella* and *Trichoplax* genomes. Putative transmembrane tyrosine kinase receptors with intracellular DDR-like domains were found in the *Amphimedon* genome but they did not contain extracellular discoidin domains.
Guanine nucleotide binding protein (G protein) α	eukaryotic	G protein α subunit proteins were found in all surveyed eukaryote genomes except *Paramecium* (Figure S8.5.13).
Guanine nucleotide binding protein (G protein) β	eukaryotic	G protein β proteins belong to an ancient eukaryotic family [194]
Guanine nucleotide binding protein (G protein) γ	eukaryotic	G protein γ subunit-like proteins were found in animal, *Monosiga* and fungal genomes. G protein γ subunits have also been reported in *Dictyostelium* and *Arabidopsis*, confirming that this is an ancient eukaryotic family[195,196]
Focal adhesion kinase	metazoan	The domain architecture for focal adhesion kinase (B41 + TyrKinaseCatalytic + FAK_AT) is specific to metazoans.
Adenylate cyclase	eukaryotic	Adenylate cyclase proteins belong to an ancient eukaryotic family.
Phospholipase C β	holozoan	Phospholipase C β proteins from animals and *Monosiga* form a well supported clade (Figure S8.5.14)
Phosphoinositide-3-kinase (PI3K)	eukaryotic	PI3K-like proteins were found in all surveyed eukaryote genomes.
Growth factor receptor bound protein 2 (Grb2)	holozoan	Proteins with a Grb2-like domain structure and displaying sequence homology to Grb2 family proteins were only found in animal and *Monosiga* genomes.
Son of sevenless (SOS)	holozoan	Proteins sharing sequence similarity and domain architecture with vertebrate SOS1 were found in *Monosiga* and animal genomes but not in the other eukaryote genomes surveyed.
RhoA/B/C	opisthokont	RhoA/B/C-like Ras family proteins were found in opisthokont genomes only.
Graf family RhoGAP	holozoan	Clear orthologs of the Graf family of RhoGAPs were not found outside *Monosiga* and animal genomes.
Integrin-linked kinase (ILK)	metazoan	Proteins displaying sequence homology to bilaterian ILK proteins were not found in non-animal genomes.
Crk	holozoan	Proteins with the domain architecture characteristic of bilaterian Crk family proteins were only found in metazoan and *Monosiga* genomes.
Ras	eukaryotic	Ras superfamily small GTP binding proteins were found in all eukaryote genomes, while Ras family proteins (a subdivision of the Ras superfamily) were found in all eukaryote genomes except plants.
RAS p21 protein activator 1	holozoan	RAS p21 protein activator 1-like RasGAP proteins were only found in animal and *Monosiga* genomes.

Neurofibromin (NF1)	opisthokont or eukaryotic	Neurofibromin-like proteins were found in animal, *Monosiga* and fungi genomes. A divergent neurofibromin-like RasGAP was found in *Dictyostelium* but not in *Paramecium* or *Arabidopsis*.
RasGEF family	eukaryotic	Present in *Dictyostelium*, Fungi, *Monosiga*, Metazoa and Euglenozoa. Absent from plants.
RAS guanyl releasing protein	metazoan	Present in *Nematostella*. The best *Amphimedon* match is missing the RasGEFN domain but this is likely to be located within a gap in the assembly.
Synaptic Ras GTPase activating protein 1 (SynGAP)	metazoan	Proteins with homology to SynGAP and its close relatives are present in *Nematostella* and *Amphimedon* but may be absent from *Trichoplax*.
Kinase suppressor of ras (KSR)	metazoan	Proteins sharing sequence homology with human KSR proteins and possessing the domain structure characteristic of the bilaterian KSR family were not found outside the Metazoa.
Ral	opisthokont	Putative orthologs of bilaterian proteins belonging to the Ral group of the Ras family were found in *Monosiga* and animals but not in other analyzed genomes. However, phylogenetic analysis of Ral-like proteins from early diverging fungal genomes indicates that they are orthologous to holozoan Ral proteins, placing the origin of Ral in the opisthokont lineage rather than in the holozoan lineage[197]
Ral guanine nucleotide dissociation stimulator (RalGDS)	metazoan	The domain architecture for RalGDS was not found outside the Metazoa. The *Nematostella* representative lacks an RA domain but this might be a problem with the gene model. No better BLASTp hit was found for *Hydra* but a tBLASTn search was not conducted.
RapGEF (C3G)	holozoan	Clear orthologs of the RapGEF1 (C3G) family of RasGEFs were not found outside *Monosiga* and animals. The top *Dictyostelium* hit appears to be equally related to proteins within both the RapGEF1 and SOS families of RasGEFs.
Akt	eukaryotic	See Table S8.2.2
Protein kinase C α/β/γ family (cPKC)	holozoan	Protein kinase C proteins from the α/β/γ (conventional) family appear to be restricted to holozoan representatives (on the basis of conserved domain architecture).
CAMKII	holozoan	See Manning et al. 2008[127]
Cdc42	opisthokont	Rho-like proteins from non-opsithokont genomes have higher sequence homology to non-Cdc42 Rho proteins than to Cdc42 proteins.
Rac	eukaryotic	Rac-like proteins were found in all surveyed eukaryote genomes.
Rap1	eukaryotic	Proteins with similarity to Rap1 subfamily members were found in all surveyed genomes except plants.
Epac family RapGEFs	metazoan	*Amphimedon* and *Nematostella* Epac-like RapGEFs share a domain structure similar to human Epac2 (RapGEF4). *Trichoplax* does not appear to possess a gene with equivalent domains.
Raf	metazoan	Proteins sharing sequence homology with human B-Raf and possessing the domain structure characteristic of the bilaterian A-Raf/B-Raf/C-Raf family were not found outside the Metazoa.
14-3-3	eukaryotic	The 14-3-3 domain and proteins with sequence homology to animal 14-3-3 proteins were found in all surveyed genomes (Figure S8.5.15)
Mitogen-activated protein kinase kinase Mek1/2	opisthokont	MAP2K (or MAPKK or MEK) like proteins were found in all surveyed eukaryote genomes, however published phylogenetic analyses suggest that the Mek1/2 group (ie MAP2K1/MAP2K2) may be restricted to opisthokonts[198]
JNK	Metazoan	See Tables S8.5.2 and S8.7.2.
Mitogen-activated protein kinase ERK1/2	eukaryotic	See Manning et al. 2002[164]
p90 ribosomal protein S6 kinase (RSK-p90)	holozoan	RSK-p90 proteins were only found in animal and *Monosiga* genomes.

PKA	eukaryotic	See Manning et al. 2002[164]
CREB family bZIP protein	metazoan	Proteins resembling bilaterian CREB family bZIP proteins were only found in animal genomes.
FOS	metazoan	See Table S8.5.2.
JUN	metazoan	See Table S8.5.2.

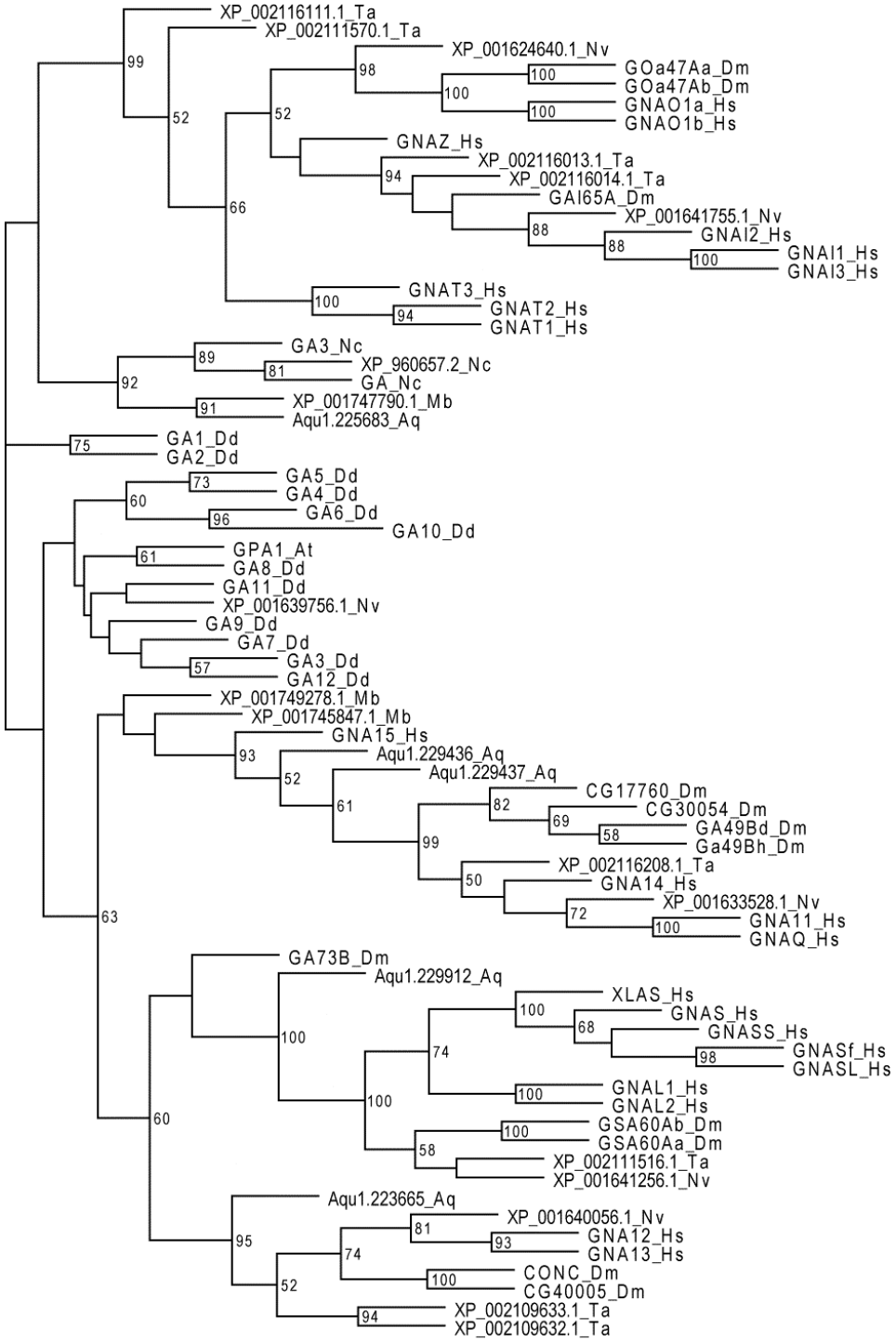


Figure S8.5.13: Rooted neighbour joining tree for G α proteins. Midpoint rooting has been used. Trees were built using the neighbor-joining method in Phylip[39] using default settings with all programs except Neighbour, for which the input order of species was randomized. *Amphimedon* G protein α-like proteins fall into several well-supported clades in a phylogenetic tree created for the eukaryotic G protein α family. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*; At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Nc, *Neurospora crassa*.

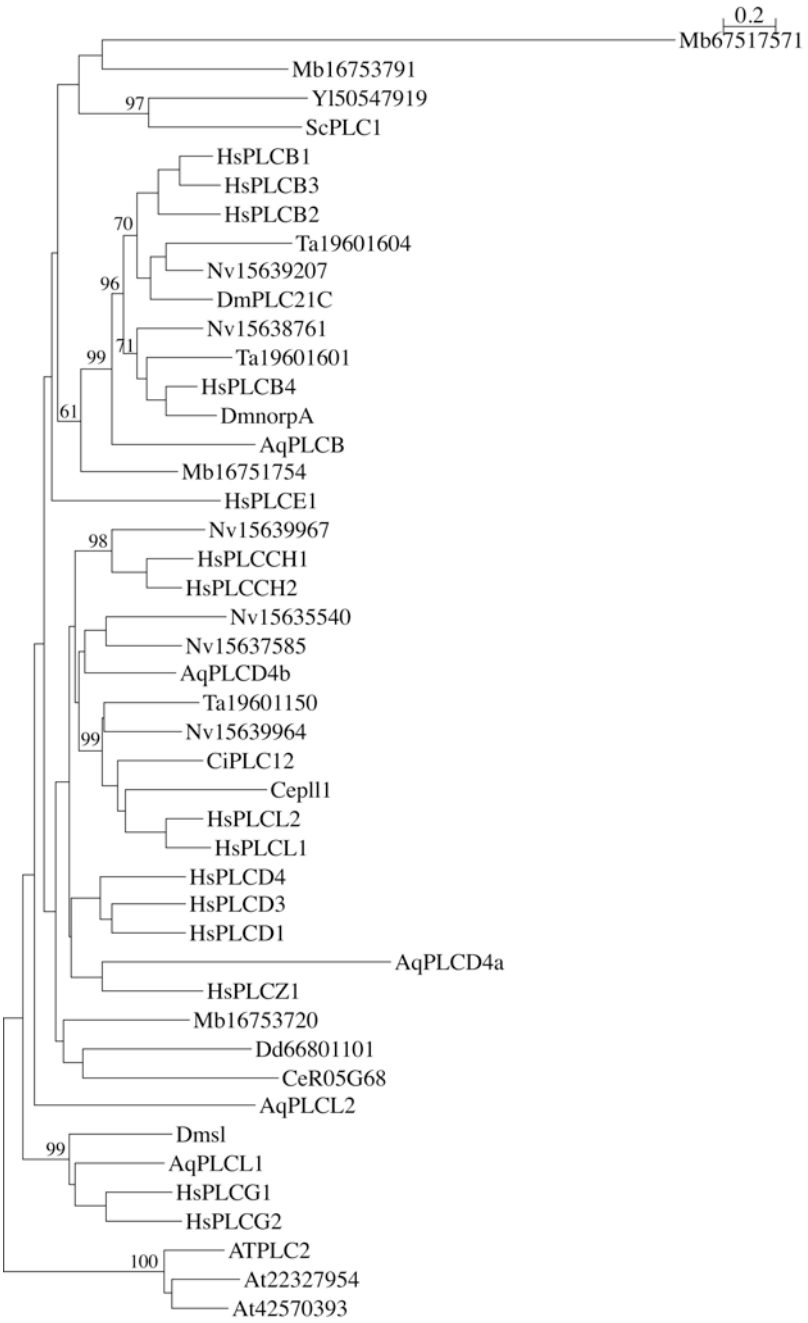


Figure S8.5.14: Rooted neighbor-joining tree of PLC genes. The tree is rooted with plant *PLC*s. The best match in the *M. brevicollis* genome for *PLCB* genes clusters alongside these genes, suggesting that it is orthologous to metazoan *PLCB* genes and that this gene originated in the holozoan stem. Nodes of interest are labeled with bootstrap values (100 replicates), only values >50 are shown. Mb, *Monosiga brevicollis*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*, At, *Arabidopsis thaliana*; Sc *Saccharomyces cerevisiae*; Ci, *Ciona intestinalis*; Dd, *Dictyostelium discoideum*; Ce, *Caenorhabditis elegans*; Yl, *Yarrowia lipolytica*.

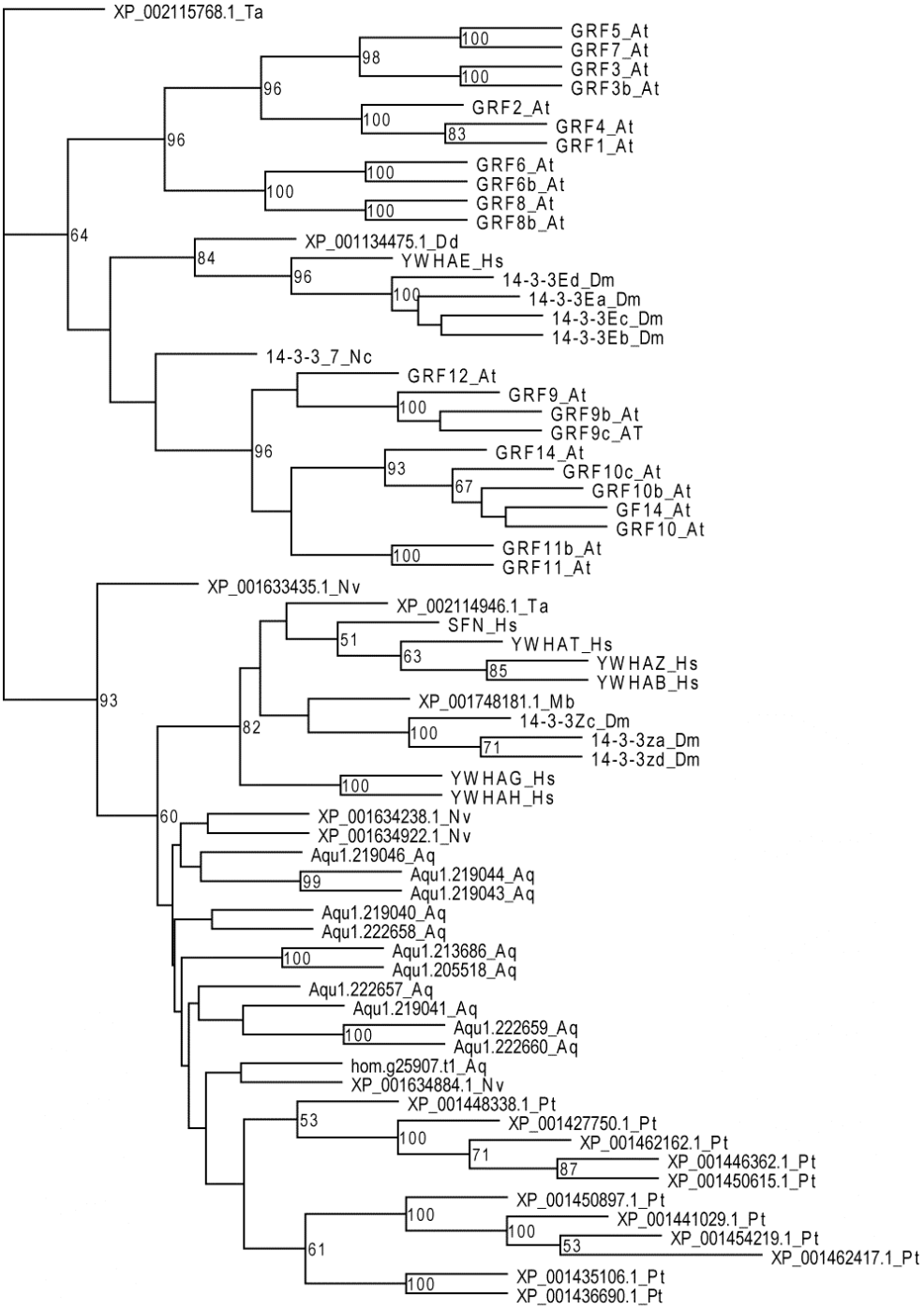


Figure S8.5.15: Rooted neighbour joining tree of 14-3-3 family genes. Midpoint rooting has been used. Trees were built using the neighbor-joining method in Phylip[39] using default settings with all programs except Neighbour, for which the input order of species was randomized. *Amphimedon* 14-3-3-like proteins fall into a well-supported clade along with several metazoan and *Paramecium* proteins and may represent descendants of a lineage specific expansion. Nodes are labeled with bootstrap values, only values >50 are shown. Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*; At, *Arabidopsis thaliana*; Dd, *Dictyostelium discoideum*; Nc, *Neurospora crassa*; Pt, *Paramecium tetraurelia*.

Table S8.6.1: Transcription factor genes in the *Amphimedon* genome.

	Protein	Gene model	Bilaterian orthologs	*Amphimedon* gene name	Functional domains present // absent	Protein domain configuration conserved?
BZIP	CREB/ATFI-IIa (CREBZF-like)	Aqu1.209649	CREB/ATFI-II (CREBZF-like)	AmqCREB/ATFI-IIa (CREBZF-like)	bzip	Y
BZIP	Fos	Aqu1.213207	Fos	AmqFos	bzip	Y
BZIP	Jun (AP1)	Aqu1.214126	Jun (AP1)	AmqJun (AP1)	bzip	Y
BZIP	CREB/ATFI-IIb (CREB3-like)	Aqu1.216147	CREB/ATFI-II (CREB3-like)	AmqCREB/ATFI-IIb (CREB3-like)	bzip	Y
BZIP	CREB/ATFI-IIc (CREB3-like)	Aqu1.219683	CREB/ATFI-II (CREB3-like)	AmqCREB/ATFI-IIc (CREB3-like)	bzip	Y
BZIP	NF-E2a	Aqu1.224064	NF-E2	AmqNF-E2a	bzip	Y
BZIP	XBP	Aqu1.224228	XBP	AmqXBP	bzip	Y
BZIP	ATP2/7a	Aqu1.224615	ATP2/7	AmqATP2/7a	bzip	Y
BZIP	CREB/ATFI-IId (CREB3-like)	Aqu1.224656	CREB/ATFI-II (CREB3-like)	AmqCREB/ATFI-IId (CREB3-like)	bzip	Y
BZIP	similar to CG17836 (also Nv protein)	Aqu1.224866	similar to CG17836 (also Nv protein)	similar to CG17836 (also Nv protein)	bzip	Y
BZIP	NF-E2b	Aqu1.224869	NF-E2	AmqNF-E2b	bzip	Y
BZIP	ATP2/7b	Aqu1.224873	ATP2/7	AmqATP2/7b	bzip	Y
BZIP	C/EBP	Aqu1.224916	C/EBP	AmqC/EBP	bzip	Y
BZIP	ATF4/5	Aqu1.225687	ATF4/5	AmqATF4/5	bzip	Y
BZIP	CREBP/ATFII Ia	Aqu1.229614	CREBP/ATFIII	AmqCREBP/ATFII Ia	bzip	Y
BZIP	CREBP/ATFII Ib	Aqu1.229932	CREBP/ATFIII	AmqCREBP/ATFII Ib	bzip	Y
BZIP	Maf	Aqu1.212336	Maf	AmqMaf	bzip	Y
ETS		Aqu1.215662	Elf/E74		ets	no SAM
ETS		Aqu1.220463	Fli/ERG/E26/v-ets		ets, Sterile α motif (SAM)//Pointed domain	
ETS		Aqu1.220473	ets/pointed		ets, Sterile α motif (SAM)//Pointed domain	
ETS		Aqu1.220474	ets/pointed		ets	
ETS		Aqu1.222248			ets	
ETS		aq_ka13438x00310			ets	
ETS		Aqu1.222249	etv5/erm/et1		ets	
ETS		hom.g28614.t1	ERM/ER81		ets	
ETS		Aqu1.224919	ERM/ER81		ets	
HMG	SSRP	Aqu1.209664	SSRP (HMG/UBF group)	AmqSSRP	SSRecognition//HMG	Y
HMG	TOX/LCPa	Aqu1.209717	CAGF9/TOX/LCP	AmqTOX/LCPa	1 HMG	Y
HMG	TOX/LCP2	Aqu1.214792	CAGF9/TOX/LCP	AmqTOX/LCPb	1 HMG	Y

HMG	HMG1	Aqu1.216195	HMG1 (HMG/UBF group)	AmqHMG1	2 HMG	Y
HMG	mtTFA	Aqu1.218967	mtTFA (HMG/UBF group)	AmqmtTFA	2 HMG	Y
HMG	Baf57	Aqu1.224100	Baf57 (HMG/UBF group)	AmqBaf57	1 HMG	Y
HMG	Maelstrom	Aqu1.227127	Maelstrom	AmqMaelstrom	1 HMG	Y
HMG	HMG2a	Aqu1.227714	HMG2 (HMG/UBF group)	AmqHMG2a	2 HMG	Y
HMG	HMG2b	Aqu1.227743	HMG2 (HMG/UBF group)	AmqHMG2b	3 HMG	Y
HMG	HMG20	Aqu1.229431	HMG20 (HMG/UBF group)	AmqHMG20	1 HMG	Y
HMG	TCF/LEF	Aqu1.229819	TCF/LEF	AmqTCF/LEF	CTNNB1-binding, HMG	Y
HMG	SoxB2-like	Aqu1.217790	SoxB2	AmqSoxB2-like	Sox	Y
HMG	SoxC-like	Aqu1.219828	SoxC	AmqSoxC-like	Sox	Y
HMG	SoxB1-like	Aqu1.223121	SoxB1	AmqSoxB1-like	Sox	Y
HMG	SoxF-like	Aqu1.223925	SoxF	AmqSoxF-like	Sox	Y
HMG	polybromo-like	Aqu1.226981	Polybromo (HMG/UBF group)	Amqpolybromo-like	3Bromo,HMG(below threshold), 2BAH, 2 C2H2 zinc fingers	Y
HMG	ALR-like	Aqu1.213265	ALR (HMG/UBF group)	AmqALR-like	HMG(below threshold), 5 PHD, FYRN, FYRC, SET, PostSET	Y
GATA TF (2 GATA ZF)	GATA	Aqu1.223181	GATA TF	AmqGATA	BOM, 2 GATA ZF, DUF1518	extra domains present
Gli C2H2 ZF	Gli2/3a	Aqu1.217717	Gli2/3	AmqGli2/3a	5 C2H2 zinc fingers	Y
Gli C2H2 ZF	Glis1/3	Aqu1.213405	GLIS1/3(Gli-similar)	AmqGlis1/3	5 C2H2 zinc fingers	Y
Gli C2H2 ZF	Gli2/3b	Aqu1.219964	Gli2/3	AmqGli2/3b	5 C2H2 zinc fingers	Y
Snail C2H2 ZF		none				
Zic C2H2 ZF		none				

Table S8.7.2: Origins of kinase classes

Metazoan-specific		
Group	**Family**	**Subfamily**
AGC	DMPK	CRIK
AGC	GRK	GRK
AGC	PKC	PKCi
AGC	PKG	
AGC	PKN	
AGC	RSKL	
CAMK	CAMKL	HUNK
CAMK	DAPK	DAPK
CAMK	MAPKAPK	MK2
CAMK	MAPKAPK	MK5
CAMK	MAPKAPK	MNK
CAMK	SgK495	
CAMK	Trbl	
CAMK	TSSK	
CMGC	CDK	PFTAIRE
CMGC	DYRK	HIPK
CMGC	MAPK	ERK5
CMGC	MAPK	JNK
CMGC	MAPK	nmo
CMGC	RCK	MOK
Other	Dusty	
Other	IKK	
Other	NEK	NEK11
Other	NEK	NEK6
Other	NKF2	
Other	NKF3	
STE	STE11	MEKK15
STE	STE20	KHS
STE	STE20	PAKB
STE	STE7	MEK7
TK	CCK4	
TK	DDR	
TK	EGFR	
TK	Eph	
TK	FAK	
TK	Fer	
TK	Met	
TK	Ret	
TK	Ryk	
TK	Sev	
TK	Src	Frk
TK	Syk	
TKL	LISK	TESK
TKL	RAF	KSR
TKL	RAF	RAF

TKL	STKR	STKR1
TKL	STKR	STKR2
***Amphimedon* Only**		
Group	**Family**	**Subfamily**
Atypical	Alpha	Alpha-sponge
TK	REF	
TK	Src	Src-Aque1
TK	TMR	
TKL	MLK	HH-a1
TKL	MLK	HH-a2
TKL	MLK	HH-B
TKL	MLK	HH-C
Lost from *Amphimedon*		
Group	**Family**	**Subfamily**
Atypical	G11	
Atypical	PDHK	BCKDK
Atypical	PDHK	PDHK
Atypical	TIF1	
CAMK	DCAMKL	
Other	NEK	NEK4
Other	NEK	NEK9
Other	NKF1	
Other	SgK071	
Other	TBCK	
Other	TOPK	
Other	TTK	
Eumetazoan-specific		
Group	**Family**	**Subfamily**
AGC	MAST	MASTL
Atypical	BCR	
CAMK	CAMKL	NIM1
CAMK	CASK	
CAMK	DAPK	DRAK
CAMK	PIM	
CMGC	CDK	CDK4
CMGC	MAPK	ERK3
Other	MOS	
Other	PLK	PLK2
Other	SgK493	
STE	STE20	STLK
STE	STE7	MEK3
TK	Ack	
TK	FGFR	
TK	InsR	
TK	Ror	

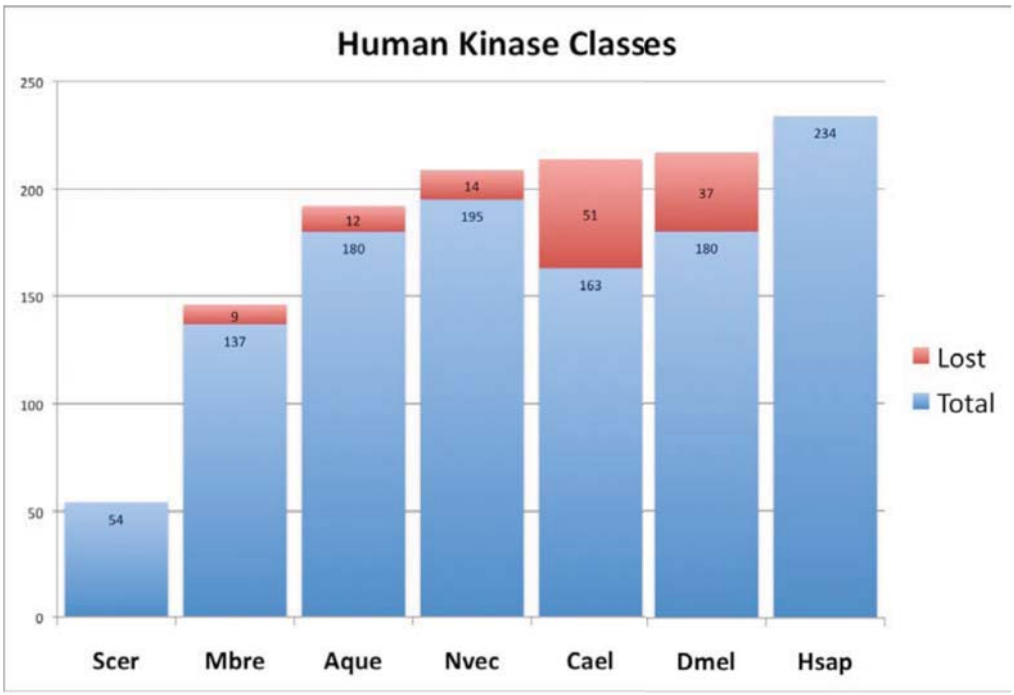


Figure S8.7.1: Loss and retention of human kinase classes in various species. Scer, *Saccharomyces cerevisiae*; Mbre, *Monosiga brevicollis*; Aque, *Amphimedon queenslandica*; Nvec, *Nematostella vectensis*; Cael, *Caenorhabditis elegans*; Dmel, *Drosophila melanogaster*; Hsap, *Homo sapiens*.

Table S8.8.1: Classification of adhesion genes by origin.

Gene/Family	Origin	Methods used to determine origin
Cadherin (for AJ cadherins see Table S8.8.2)	holozoan	The cadherin domain appears to be specific to *Monosiga* and animals. Even in *Monosiga* these domains are associated with the extracellular region of transmembrane proteins.
IgCAM	holozoan	Ig domains appear to be specific to *Monosiga* and animals although there are far greater numbers of Ig containing proteins in animal genomes than in the choanoflagellate genome. At least one of the Ig domains found in the choanoflagellate genome is associated with the putative extracellular region of an adhesion related transmembrane protein.
Ig + FN3 CAM	metazoan	Although one *Monosiga* protein contains both Ig and FN3 domains, the domain architecture consisting of a stretch of Ig repeats followed by a stretch of FN3 repeats in the putative extracellular region of a transmembrane protein appears to be metazoan-specific.
LRRCAM	eukaryotic	LRR is a taxonomically widespread domain found in both intracellular and extracellular proteins. Putative transmembrane proteins with extracellular LRR repeats were found in animal, *Monosiga*, *Arabidopsis*, and *Dictyostelium* genomes but not in fungal or *Paramecium* genomes.
Neurexin I/II/III	eumetazoan s.s.	Genes with a domain structure similar to bilaterian full length Neurexin I/II/III proteins were not found in *Nematostella, Trichoplax, Amphimedon* or non-animal genomes. Several shorter predicted transmembrane proteins with LamG and EGF domains were found in the *Nematostella* genome and these displayed similarity to bilaterian proteins within the Neurexin I/II/III familiy.
Collagen	holozoan/opisthokont	Collagen repeat domains were found in putative secreted proteins (containing signal peptides) in animal and choanoflagellate genomes. Only a few convincing collagen repeat domains were found in fungal proteins and only some of these were associated with a signal peptide. Therefore it is unclear whether collagen repeat containing extracellular proteins are opisthokont- or holozoan-specific.
Fibrillar collagen	metazoan	C-terminal COLFI domains were found in collagen repeat containing proteins in *Nematostella* and *Amphimedon*. Although proteins with COLFI domains are encoded by the *Monosiga* genome, none contain collagen repeats.
Agrin	eumetazoan	No genes with similar domain structure were found in sponges or non-animals. An agrin-like gene is present in *Nematostella* but contains an unrelated domain (VWA) at the N- and C-termini. The *Trichoplax* agrin-like protein lacks the LamEGF repeats and has a shortened FOLN/Kazal repeat region.
Netrin	eumetazoan	No netrin C-terminal domains were detected in sponge or non-animal genomes. *Trichoplax* has two netrin-like proteins but both lack the LamNT domain.
Thrombospondin	metazoan	No thrombospondin-like genes were found outside the Metazoa. The only identified *Trichoplax* representative is aberrant, containing a thrombospondin-like C-terminal region (EGF, TSP_3 and TSP C-terminal domains) but with LRR repeats and a Cadherin domain at the N-terminus.
Integrin α	metazoan	N-terminal integrin α repeats are found outside Metazoa but the integrin α 2 domain and cytoplasmic conserved motif/pattern are metazoan-specific.
Integrin β	metazoan	Hits to the three integrin β-specific domains are not found outside Metazoa. *Amphimedon*, *Trichoplax* and *Nematostella* candidates contain most or all of these domains.
Dystroglycan (DG)	metazoan	The dystroglycan-like cadherin domain (CADG) predates the Metazoa (hits in fungal proteins) but proteins with sequence similarity to bilaterian dystroglycan proteins were only found in metazoan genomes. Dystroglycan-like proteins found in *Nematostella* and *Trichoplax* seem to have increased numbers of CADG domains but still demonstrate clear homology to bilaterian dystroglycan proteins.

Table S8.8.2: Origins of genes associated with polarized epithelia

Gene/Family	Origin	Methods used to determine origin
Par-3	metazoan	No fungal or choanoflagellate PDZ proteins were found to contain the conserved N-terminal domain that is characteristic of animal Par-3 proteins.
Par-6	metazoan	Proteins with the domain architecture (PB1 + PDZ) specific to animal Par-6 proteins were not found in choanoflagellate or fungal genomes.
aPKC	metazoan	Proteins with the domain architecture (PB1 + C1 + Kinase + Kinase C-term) specific to animal aPKC proteins were not found in choanoflagellate or fungal genomes.
Par-1	holozoan	Holozoan Par-1 orthologs form a well-supported clade in a phylogeny based on the kinase domain (Fahey et al., in prep). Related fungal kinases do not group with or near this clade.
Crumbs	metazoan	Putative transmembrane proteins containing a combination of LamG and EGF domains as well as the conserved Crumbs cytoplasmic motif were found only in animal genomes.
Stardust (MPP5)	eumetazoan *s.s.*	Stardust/MPP5 orthologs from *Nematostella* and bilaterians form a well supported clade in a phylogeny based on the SH3 and GUK domain of holozoan MPP-related MAGUK proteins (Fahey et al. in prep) . Several *Amphimedon* and *Trichoplax* proteins are placed within the MPP family but with no clear relationships to Stardust/MPP5 or any of the other bilaterian orthology groups.
PatJ	metazoan	The domain architecture characteristic of animal PatJ proteins (N-terminal L27 followed by multiple PDZ) appears to be specific to proteins from metazoan genomes.
Scribble	metazoan	Proteins with the domain architecture (LRR + PDZ) specific to animal Scribble proteins were not found in choanoflagellate or fungal genomes.
Discs large	holozoan	Proteins sharing a domain architecture and sequence similarity with bilaterian Discs large orthologs are present in animal and choanoflagellate genomes.
Lethal giant larvae	metazoan	The metazoan lethal giant larvae and tomosyn (Syntaxin binding protein 5) gene families are believed to have arisen by duplication from a single orthology group present in plants, fungi and choanoflagellates. [199]The *Amphimedon* genome contains at least one representative for each family, suggesting that this duplication occurred in the lineage leading to the Metazoa.
Cadherin (Classic/AJ type)	metazoan	Choanoflagellate cadherin proteins lack the catenin-binding cytoplasmic domain which characterizes metazoan adherens junction forming cadherins.
α-catenin	metazoan	The metazoan α-catenin family probably arose through gene duplication and divergence from a vinculin-like precursor. Whereas choanoflagellates possess a single vinculin-like gene, *Amphimedon* and other animals possess both vinculin- and α-catenin-like genes.
β-catenin	metazoan	Armadillo/β-catenin repeat is present outside of Metazoa but the specific domain combination of β-catenin is metazoan-specific. *Dictyostelium* Aardvark and plant Arabidillo proteins have different domain structures (additional F-box, missing C-terminal PDZ transactivation domain, less armadillo repeats).
δ-catenin	metazoan	Armadillo repeat proteins from fungi and choanoflagellates do not display convincing sequence similarity to metazoan δ-catenin orthologs.
Claudin	bilaterian or vertebrate	Vertebrate tight-junction forming claudins belong to a wider claudin-related family which includes proteins found in non-vertebrate bilaterians like *Drosophila* and *C elegans*. The family consists of short four-transmembrane proteins which display little sequence similarity to each other outside of the transmembrane helices. For this reason, it is difficult to determine whether any of the *Drosophila* claudin-like proteins are more closely related to vertebrate tight junction claudins than to other claudin family proteins as has been claimed.[200,201]
Occludin	vertebrate	The domain architecture (MARVEL + Occludin_ELL) characteristic of vertebrate occludin and related proteins (tricellulin) appears to be specific to proteins from vertebrate genomes.
Neurexin IV/CASPR	eumetazoan	No sponge or choanoflagellate LamG domain containing transmembrane proteins possess the domain architecture characteristic of bilaterian Neurexin IV proteins. *Trichoplax* and *Nematostella* Neurexin IV-like proteins lack the N-terminal discoidin domain.
Contactin	bilaterian	Bilaterian Contactin orthologs are GPI-linked extracellular proteins containing 6 Ig and 4 FN3 domains. *Amphimedon*, *Trichoplax* and *Nematostella* genomes contain many Ig and FN3 containing putative transmembrane proteins but none display a GPI linkage, the specified number of domains and/or convincing sequence similarity to

		Contactin proteins outside of the Ig/FN3 repeat regions.
Neuroglian	eumetazoan	Bilaterian neuroglian orthologs are transmembrane proteins containing 6 Ig and 5 FN3 domains. None of the putative transmembrane Ig and FN3 containing proteins encoded by the *Amphimedon* genome were considered similar enough in terms of domain architecture or sequence similarity to qualify as convincing neuroglian orthologs.
Collagen IV	metazoan	Choanoflagellate collagen repeat encoding proteins lack the C-terminal C4 repeats which are characteristic of metazoan Collagen IV proteins. The *Amphimedon* genome lacks an ortholog of Collagen IV but a Collagen IV-like protein has been cloned from the homoscleromorph sponge, *Pseudocorticium jarrei.* [202]
Collagen XV/XVIII	eumetazoan *s.s.*	None of the collagen repeat encoding proteins encoded by the *Monosiga*, *Amphimedon* and *Trichoplax* genomes contain the C-terminal endostatin domain that is characteristic of bilaterian Collagen XV/XVIII proteins.
Laminin α	eumetazoan or animal	*Amphimedon* encodes a single laminin α 3/5-like protein but this lacks a well-defined LamNT and α 3/5 domain (a few conserved amino acids in comparable locations only) and completely lacks a IVA domain.
Laminin β	eumetazoan	Laminin-like proteins with a domain architecture resembling bilaterian laminin β were found in eumetazoan genomes only.
Laminin γ	metazoan	*Amphimedon* encodes a laminin-like protein with a domain architecture comparable to bilaterian laminin γ proteins. However, this protein does contain a short region in the middle of the coiled coil stretch which resembles the laminin β knob motif specific to bilaterian laminin β proteins.
Nidogen	eumetazoan	Proteins with a domain architecture characteristic of bilaterian Nidogen proteins appear to be restricted to eumetazoan genomes.
Perlecan	eumetazoan	Perlecan-like proteins were found in eumetazoan genomes only. In the *Nematostella* genome gene fragments resembling the N-terminus and C-terminus of Perlecan were found at the edges of two different contigs and no additional complete genes were found.

Table S8.9.1. Origins of regulatory genes involved in bilaterian neurogenesis.

Class	Gene/Family	Origin	Explanation
Sox	SoxB1	metazoan	Larroux et al 2008[123], Magie et al 2005[203]
	SoxB2	metazoan	Larroux et al 2008[123], Magie et al 2005[203]
LIM-HD	lim3	metazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
	lin11	metazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
	islet	metazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
	apterous	eumetazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
	Lmx	eumetazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
	Lhx6/7	eumetazoan	Larroux et al 2008[123], and Srivastava et al 2010[129]
Pax	PaxB	metazoan	Larroux et al 2008[123], Matus et al 2007[204], Hadrys et al 2005[205]
	PaxD	eumetazoan	Larroux et al 2008[123], Matus et al 2007[204]
	PaxA/C/pox neuro	eumetazoan	Larroux et al 2008[123], Matus et al 2007[204]
	Pax6	bilaterian	Larroux et al 2008[123], Matus et al 2007[204]
	Pax1/9	eumetazoan	Larroux et al 2008[123], Matus et al 2007[204]
prd-like	Rx	metazoan	Larroux et al 2008[123], Matus et al 2007[204],
	al	metazoan	Larroux et al 2008[123], Ryan et al 2006[206]
	otx	eumetazoan	Larroux et al 2008, Matus et al 2006[207], Ryan et al 2006[206]
POU	PouI	metazoan	Larroux et al 2008, Ryan et al 2006
	PouII	bilaterian	Larroux et al 2008, Ryan et al 2006; PouII/III/IV gene in Amphimedon, PouII/III gene in *Nematostella*
	PouIII	bilaterian	Larroux et al 2008, Ryan et al 2006; PouII/III/IV gene in Amphimedon, PouII/III gene in *Nematostella*
	PouIV	eumetazoan	Larroux et al 2008, Ryan et al 2006
	PouVI	metazoan	Larroux et al 2008, Ryan et al 2006
Six	Six1/2	metazoan	Larroux et al 2008
	Six3/6	eumetazoan	Larroux et al 2008
	Six4/5	eumetazoan	Larroux et al 2008
bHLH	COE	metazoan	Simionato et al 2007[125], Pang et al 2004[208]
	atonal	bilaterian	Simionato et al 2007
	neurogenin	eumetazoan	Simionato et al 2007, Degnan et al submitted
	NeuroD	eumetazoan	Simionato et al 2007
	β3	eumetazoan	Simionato et al 2007
	Oligo	eumetazoan	Simionato et al 2007
	ASH	metazoan	Simionato et al 2007
Tbx	Tbx2/3	eumetazoan	Larroux et al 2008, Yamada et al 2007[209], Martinelli and Spring 2003[210]
ANTP	Msx	metazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006[211], Monteiro et al 2006[212]
	Bsh	metazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	Gbx	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	Hox	eumetazoan	Larroux et al 2007, Finnerty and Martindale 1997, Ryan et al 2006, Chourrout et al 2006, Kamm et al 2006, Monteiro et al 2006
	Gsx	eumetazoan	Larroux et al 2007, Finnerty et al 2003, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006, Jakob et al 2004
	NK2	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	NK6	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	Dlx	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	Dbx	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
	Mnx	eumetazoan	Larroux et al 2007, Ryan et al 2006, Chourrout et al 2006, Monteiro et al 2006
NR	COUP-TF	eumetazoan	Brigham et al., unpublished

Table S8.9.2. Classification of pre-synaptic genes by origin

Gene	Origin	Description	Domains
CELL ADHESION/ CELL-MATRIX INTERACTIONS			
CDH	Holozoa	Cadherin	Cadherin repeat, cadherin C-term (C-term domain only in Metazoa)
CNTN	Bilateria	Contactin	Ig, FN3
CNTNAP	Eumetazoa	Contactin-associated protein; Neurexin 4	FA58C, Lam, EGF
CTNNA	Metazoa	Alpha-catenin	Vinculin
CTNNB	Metazoa	Beta-catenin	ARM
CTNND	Metazoa	Delta-catenin	ARM
CTTN	Holozoa	Cortactin	HS1 repeat, SH3
EFNB	Eumetazoa s.s.	Ephrin B	Ephrin
PSCD	Eukaryota	Cytohesin	Sec7, PH (*Amphimedon* gene has extra domains)
PTPRF	Holozoa	Protein tyrosine phosphatase, receptor type, F (LAR)	Ig, FN3, PTPc (*Monosiga* gene has FN3 and PTPc domains only)
LSAMP/HNT/ OPCML	Bilateria	Limbic system-associated membrane protein/ Neurotrimin/ Opioid binding protein/cell adhesion molecule-like isoform a preproprotein	Ig
NCAM/NFASC	Eumetazoa	Neural cell adhesion molecule/ Neurofascin	Ig, FN3
NRCAM	Eumetazoa s.s.	Neuronal cell adhesion molecule	Ig, FN3
NRXN	Eumetazoa s.s.	Neurexin I/II/III	LamG, EGF (*Nematostella* gene has LamG domain only)
PPFIA	Metazoa	Liprin alpha	SMC N-term, SAM
SDK	Eumetazoa	Sidekick	Ig, FN3
CADM1	Vertebrate	Syncam1, immunoglobulin superfamily, member 4D isoform 1	Ig
SIGNALING			
ABL	Holozoa	c-abl oncogene 1, receptor tyrosine kinase	SH3, SH2, Pkinase, actin binding
CACNA1	Holozoa	Calcium channel, voltage-dependent	Ion transport domain
CAMK	Holozoa	Calcium/calmodulin-dependent protein kinase	Pkinase, NTF2
CIB	Holozoa	Calcium and integrin binding	EF hand
FMR	Metazoa	Fragile X mental retardation	Agenet, KH
GIT	Holozoa	G protein-coupled receptor kinase interactor	ArfGAP, Ankyrin, SPA2 (*Monosiga* gene lacks the SPA domain; *Amphimedon* gene has extra domains)
YWHAQ	Eukaryota	14-3-3	14-3-3
SMALL GTPases AND RELATED PROTEINS			
CHM	Eukaryota	Choroideremia isoform a	GDI
RABAC	Eukaryota	Rab acceptor 1	PRA1
RAB11B	Eukaryota	Rab GTPase/member RAS oncogene family 11B	Ras
RAB3A	Eukaryota	Rab GTPase/member RAS oncogene family 3A	Ras
RAB5A	Eukaryota	Rab GTPase/member RAS oncogene family 5A	Ras
RAB7A	Eukaryota	Rab GTPase/member RAS oncogene family 7A	Ras

RAB3IP/RAB3IL1	Holozoa	RAB3A interacting protein	Sec2p
RAB3GAP1	Eukaryota	RAB3 GTPase-activating protein	
RABGGTB	Eukaryota	Rab geranylgeranyltransferase, beta subunit	GGTase II
RABIF	Eukaryota	RAB-interacting factor	RabGEF
RAPGEF4	Holozoa	Rap guanine nucleotide exchange factor (GEF) 4	CAP ED, DEP, REM, RasGEF
RASA1	Eukaryota	RAS p21 protein activator 1 isoform 1	SH3, SH2, PH, C2, RasGAP
RPH3A/RPH3AL	Eumetazoa	Rabphilin 3A	RPH3 effector, C2
SCAFFOLDING			
BSN	Vertebrate	Bassoon; neuronal double zinc finger protein	ZnF
CASK	Eumetazoa	LIN2; Calcium/calmodulin-dependent serine protein kinase (MAGUK family)	Pkinase, L27, PDZ, SH3, P-loop NTPase
ERC2	Eumetazoa s.s.	ELKS/RAB6-interacting/CAST family member	Cast
APBA	Metazoa	MINT; Lin-10; Munc18-interacting protein	PH-like, PDZ
PCLO	Vertebrate	Piccolo	ZnF, PDZ, C2
PFN	Eukaryota	Profilin; actin-binding	Prof
RIMBP2	Metazoa	RIMS binding protein 2	SH3, FN3 (*Amphimedon* gene has SH3 domain only)
RIMS	Eumetazoa s.s.	Regulating synaptic membrane exocytosis	RPH3A effector, PDZ, C2 (*Trichoplax* gene has C2 domain only)
UNC13	Metazoa	Unc-13 homolog	C2, C1, DUF1041, Membr_Traf_MHD
LIN7	Metazoa	VELI; Lin-7 homolog	L27, PDZ
SNARES			
SEC22B	Eukaryota	SEC22 vesicle trafficking protein homolog B	SNC1
SNAP25/SNAP23	Eukaryota	Synaptosomal-associated protein 23/25	t-SNARE, SNAP25
SNAP29	Metazoa	Synaptosomal-associated protein 29	v-SNARE, SNAP25, t-SNARE
SNAP47	Eumetazoa	C1orf142; Synaptosomal-associated protein 47	t-SNARE
STX1A/2/3	Eukaryota	Syntaxin 1A	SynN, t-SNARE
STX6	Eukaryota	Syntaxin 6	Syntaxin-6 N, t-SNARE
STX7/12	Eukaryota	Syntaxin 7/12	t-SNARE
STX16	Eukaryota	Syntaxin 16	SynN, t-SNARE
STXBP1	Eukaryota	Munc18/syntaxin binding protein 1	Sec1
STXBP4	Holozoa	Munc18/syntaxin binding protein 4	PDZ, WW
STXBP5/STXBP5L	Eukaryota	Syntaxin binding protein 5 (tomosyn)	WD40, LLGL (pre-Metazoan genes have incomplete domains)
STXBP6	Vertebrate	Amisyn	
VAMP	Eukaryota	Vesicle-associated membrane protein	Synaptobrevin
VTI1A	Eukaryota	SNARE Vti1a-beta protein	v-SNARE
TRAFFICKING REGULATORY PROTEINS			
BAIAP3	Holozoa	BAI1-associated protein 3	C2, DUF1041, Membr_Traf_MHD (*Monosiga* gene only has the C2 domain)
CPLX	Eumetazoa	Complexin	Synaphin
NSF	Eukaryota	NSF	CDC48, P-loop NTPase
NAPA/NAPB	Eukaryota	N-ethylmaleimide-sensitive factor	

		attachment protein	
SNAPIN	Eukaryota	SNAP-associated protein	
SNIP	Vertebrate	SNAP25-interacting protein; AC115090.8	AIP3, SMC
SNPH	Vertebrate	Syntaphilin	BAR
SYNGR	Holozoa	Synaptogyrin	MARVEL domain
SYNPR/SYP	Metazoa	Synaptoporin/synaptophysin	MARVEL domain
SYT1/2/5	Metazoa	Synaptotagmin 1/2/5	C2 (*Amphimedon* gene has extra domains)
SYT7	Eumetazoa	Synaptotagmin 7	C2
SYT12/17	Metazoa	Synaptotagmin 12/17	C2
SYTL	Eumetazoa s.s.	Synaptotagmin-like	C2
VAPA	Eukaryota	VAMP (vesicle-associated membrane protein)-associated protein A, 33kDa	MSP
OTHER SYNAPTIC VESICLE PROTEINS			
DNAJC5	Eukaryota	CSP	DNAJ
AC079061.8	Vertebrate	GOLSYN; Syntabulin; AC079061.8	
SCAMP	Eukaryota	Secretory carrier membrane protein	SCAMP
SV2	Metazoa	Synaptic vesicle glycoprotein 2	MFS
SVOP/SVOPL	Eukaryota	SV2 related protein	MFS
SYN	Eumetazoa s.s.	Synapsin	Synapsin, D-ala lig C-term
TMEM163	Eumetazoa	Transmembrane protein 163	cation efflux
VPS18	Eukaryota	Vacuolar protein sorting 18	Pep3_Vps18, clathrin
VPS33B	Eukaryota	Vacuolar protein sorting 33B	Sec1
VPS45	Eukaryota	Vacuolar protein sorting 45A	Sec1
ENDOCYTOSIS			
AMPH/BIN	Eukaryota	Amphiphysin	BAR, MARCKS, SH3
DNM	Eukaryota	Dynamin	Ras-like GTPase, dynamin, PH-like, GED
EPN	Eukaryota	Epsin	VHS_ENTH_ANTH
PACSIN	Eukaryota	Syndapin/protein kinase C and casein kinase substrate in neurons	FCH, SH3
SNAP91	Eukaryota	PICALM	VHS_ENTH_ANTH
SYNJ	Eukaryota	Synaptojanin	Syja N, Exo_endo_phos, DUF1866
TRANSPORTERS/ION CHANNELS			
ATP8A1	Eukaryota	ATPase, aminophospholipid transporter (APLT), class I, type 8A, member 1	E1-E2 ATPase
ATP6V0A1	Eukaryota	ATPase, H+ transporting, lysosomal V0 subunit A1	V-ATPase I
ATP6V0C	Eukaryota	ATPase, H+ transporting, lysosomal V0 subunit C	ATP synthase C
ATP6V0D1	Eukaryota	ATPase, H+ transporting, lysosomal V0 subunit D1	ATP synthase AC39
ATP6V1A	Eukaryota	ATPase, H+ transporting, lysosomal V1 subunit A	ATP synthase, P-loop NTPase
ATP6V1B1	Eukaryota	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B1	ATP synthase, P-loop NTPase
ATP6V1C1	Eukaryota	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	V-ATPase C
ATP6V1D	Eukaryota	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D	ATP synthase D

ATP6V1E1	Eukaryota	ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E1	V-ATPase E
ATP6V1F	Eukaryota	ATPase, H+ transporting, lysosomal 14kDa, V1 subunit F	ATP synthase F
ATP6V1G1	Eukaryota	ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G1	V-ATPase G
ATP6V1H	Eukaryota	ATPase, H+ transporting, lysosomal 50/57kDa, V1 subunit H	V-ATPase H
CLCN	Eukaryota	Chloride channel	Voltage CLC, CBS
VACHT	Bilateria	Vesicular acetylcholine transporter (SLC18A3)	MFS
VAT1	Eumetazoa	Vesicle amine transport protein 1	ADH N, NADB Rossman
VGAT	Eukaryota	Solute carrier family 32, member 1 (SLC32A1)	Amino acid transporter (*Monosiga* and *Amphimedon* proteins are more similar to SLC36 family)
VGLUT	Eukaryota	Solute carrier family 17, member 7 (SLC17A7/6/8)	MFS
VMAT2	Bilateria	Solute carrier family 18 (vesicular monoamine), member 2 (SLC18A2)	MFS
ZNT3	Eukaryota	Solute carrier family 30 (zinc transporter), member 3 (SLC30A3)	Cation efflux

Table S8.9.3. Classification of post-synaptic genes by origin

Gene	Origin	Description	Domains
SCAFFOLDING PROTEINS			
DLG	holozoan	Discs large homolog	L27, PDZ, SH3, Guanylate kinase
ERBIN	vertebrate	Erbb2 interacting protein	LRR, PDZ
GRASP	metazoan	Tamalin; GRP1 (general receptor for phosphoinositides 1)-associated scaffold protein	PDZ (*Amphimedon* gene has additional domains)
GRIP	metazoan	Glutamate receptor interacting protein 1	PDZ
HOMER	holozoan	Homer homolog	PH-like
LRRC7	vertebrate	Densin180; leucine rich repeat containing 7	LRR, PDZ
MAGI	metazoan	Membrane associated guanylate kinase, WW and PDZ domain containing	Guanylate kinase, WW, PDZ
PICK	holozoan	Protein interacting with PRKCA	PDZ, Arfaptin
SCRIB	metazoan	Scribbled homolog	LRR, PDZ
SHANK	holozoan	SH3 and multiple ankyrin repeat domains	ANK, SH3, PDZ, SAM (*Monosiga* gene has only one ANK and inverted SH3 and PDZ domains)
MEMBRANE PROTEINS/ION CHANNELS/RECEPTORS			
ATP2B	eukaryotic	PMCA; ATPase, Ca++ transporting, plasma membrane	Cation transporting ATPase, E1-E2 ATPase
CACNG2	bilateria	Stargazin; calcium channel, voltage-dependent, γ subunit 2	PMP22 claudin
CHRNA	eumetazoan s.s.	Cholinergic receptor, nicotinic, α (neuronal)	Neurotransmitter-gated ion channel LBD, TM domain
DRD	metazoan	Dopamine receptor	7TM-1
EPHB	metazoan	Ephrin receptor B	Ephrin LBD, FN3, PTKc, SAM
ERBB	metazoan	Epidermal growth factor receptor	Receptor L domain, PTKc
GABBR1	eumetazoan	Γ-aminobutyric acid (GABA) B receptor, 1	CCP, PBP1 GABAb receptor, 7TM-3
GABBR2	eukaryotic	Γ-aminobutyric acid (GABA) B receptor, 2	PBP1 GABAb receptor, 7TM-3
GRIA	eumetazoan	Glutamate receptor, ionotropic, AMPA	PBP1 iGluR AMPA, PBPb, PBPe
GRIN1	eumetazoan s.s.	Glutamate receptor, ionotropic, N-methyl D-aspartate 1	PBP1 iGluR NMDA, PBPb, CaM
GRIN2	bilaterian	Glutamate receptor, ionotropic, N-methyl D-aspartate 2	PBP1 iGluR NMDA, PBPb, NMDAR C-term (only in vertebrate)
GRM	metazoan	Glutamate receptor, metabotropic	PBP1 mGluR group I, NCD3G, 7TM-3, Homer-binding domain
HTR	metazoan	5-hydroxytryptamine (serotonin) receptor	7TM-1
KCNA	holozoan	Potassium voltage-gated channel, shaker-related subfamily	BTB, ion transporter (not found in *Amphimedon*)
NLGN	eumetazoan s.s.	Neuroligin	Esterase
SIGNALING PROTEINS			
AKAP79	vertebrate	A kinase (PRKA) anchor protein 5	WSK
CIT	metazoan	Citron (rho-interacting, serine/threonine kinase 21)	STKc CRIK, CNH (*Amphimedon* gene has kinase domain only)
CRIPT	holozoan	Cysteine-rich PDZ-binding protein	Cript
GKAP	holozoan	Discs, large (Drosophila) homolog-associated protein 1	GKAP
KALRN	metazoan	Kalirin, RhoGEF kinase	Spectrin, RhoGEF, PH, Ig, FN3, STKc (*Amphimedon* gene lacks spectrin domains)
NOS	metazoan	Nitric oxide synthase 1 (neuronal)	PDZ (appears in vertebrate), NOS oxygenase, FMN reductase, NO synthase
SPAR	holozoan	SIPA1L1; signal-induced proliferation-associated 1 like 1	RapGAP, PDZ (*Monosiga*, *Amphimedon*, and *Trichoplax* genes lack the PDZ domain; *Drosophila* gene more similar to RapGAP1)
SYNGAP	metazoan	Synaptic Ras GTPase activating protein 1 homolog	PH-like, C2, RasGAP

Table S8.9.4: Classification of neurosecretory genes by origin.

Gene	Origin	Methods used to determine origin
proprotein convertase 2 (PC2)	holozoan	*PC2* genes found in *Amphimedon*, *Monosiga,* and eumetazoans *s.s.* - no *PC2* in *Trichoplax* - not found in other eukaryotes - Figure S8.9.2
proprotein convertase 1/3 (PC1/3)	eumetazoan	*PC1/3* found in eumetazoans including *Trichoplax*, but not in *Amphimedon* and *Monosiga* - Figure S8.9.2
arginyl aminopeptidase B (AP-B)	bilaterian	*AP-B* only found in chordates, *Strongylocentrotus*, and *Lottia* - Figure S8.9.3
arginyl aminopeptidase O (AP-O)	eumetazoan	*AP-O* found in eumetazoans including *Trichoplax*, but not in *Amphimedon* nor in *Monosiga* - Figure S8.9.3
leukotriene A4 hydrolase (LTA4H)	eukaryotic	found in various eukaryotes - Figure S8.9.3
carboxypeptidase E (CP-E)	bilaterian	not found in *Amphimedon*, *Trichoplax*, *Nematostella*, and *Hydra*
carboxypeptidase D (CP-D)	holozoan	found in animals including *Amphimedon* and in *Monosiga* - not found in other eukaryotes - CP-D proteins in contrast to the other CP proteins, are characterised by the presence of several carboxypeptidase domains [135]
cathepsin-L (CTS-L)	holozoan	found in animals including *Amphimedon* and in *Monosiga* - not found in other eukaryotes - Figure S8.9.4
peptidylglycine a-amidating monooxygenase (PAM)	metazoan	found in animals, but neither in *Monosiga* nor in other eukaryotes - Figure S8.9.5
glutaminyl-peptide cyclotransferase (GC)	opisthokont	found in animals, *Monosiga*, and many fungi - Figure S8.9.6
calcium activated protein for secretion (caps)	metazoan	found in animals, but neither in *Monosiga* nor in other eukaryotes - Figure S8.9.7
Protein tyrosine phosphatase receptor type N (ptprn)	metazoan	found in animals, but neither in *Monosiga* nor in other eukaryotes - Figure S8.9.8

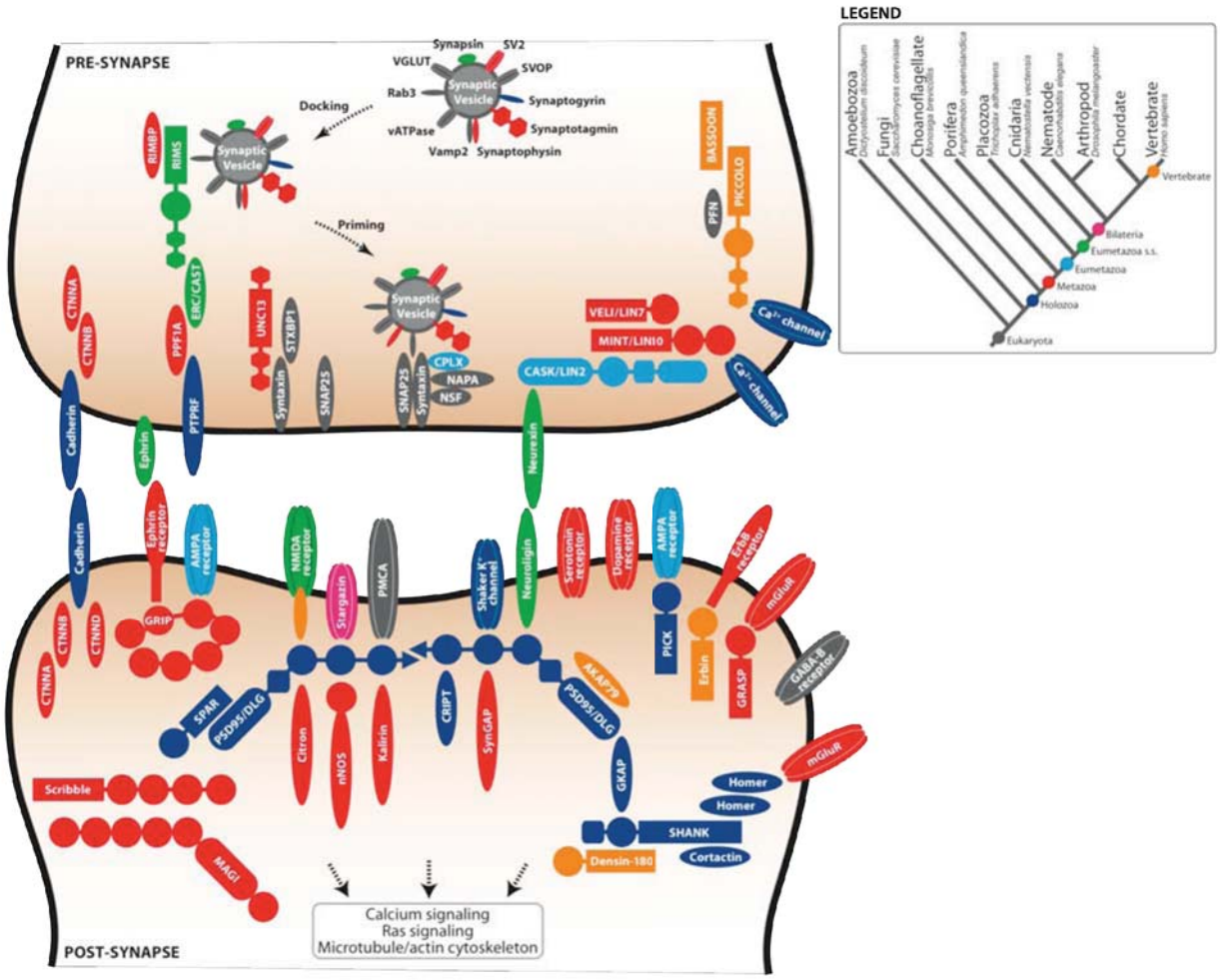


Fig. S8.9.1. Evolution of the synaptic scaffold. Orthologs of synaptic proteins were identified in animal genomes. Colors indicate the common ancestor in which a gene ortholog most likely emerged (refer to tree at right). The inclusion of a gene in a particular species is based upon reciprocal best hit with the human gene/gene family and results were filtered by looking at conservation of domain architecture whenever possible.


PC2-like1 Braflo
PC2-like2 Braflo
PC2-like3 Braflo
PC2-like1 Ampque
PC2-like2 Ampque
100/100
PC2-like3 Ampque
PC2-like4 Ampque
PC2-like5 Ampque
PC2-like Monbre
PC2-like Capsp1
PC2-like Lotgig
PC2-like4 Braflo
PC2-like5 Braflo
PC2-like1 Nemvec
PC2-like1 Hydmag
PC2-like2 Nemvec
PC2-like2 Hydmag
PC2-like3 Nemvec
PC2-like4 Nemvec
PC2-like3 Hydmag
PC2-like5 Nemvec
PC2-like4 Hydmag
PC2-like6 Nemvec
PC2 Apimel
PC2 Nasvit
PC2 Pedhumcor
PC2 Tricas
PC2 Dromel
PC2 Dappul
PC2 Caeele
PC2 Brumal
PC2 Lotgig
PC2 Lymsta
PC2 Aplcal
PC2 Halasi
PC2 Pladum
PC2 Capsp1
PC2 Cioint
PC2 Homsap
PC2 Musmus
PC2 Galgal
PC2 Danrer
PC2 Braflo
PC2 Strpur
PC2
PC5/6-1 Braflo
PC5/6-2 Braflo
PC5/6 Cioint
PC5a Danrer
PC5 Homsap
PC5 Musmus
PC5 Galgal
PC5b Danrer
100/100
PC5/6-1 Ampque
PC5/6-2 Ampque
PC5/6-3 Ampque
PC5/6
PC6 Danrer
PC6 Homsap
PC6 Musmus
PC6 Galgal
PC5/6 Strpur
furin Dromel
furin Pedhumcor
furin Apimel
furin-1 Dappul
furin-2 Dappul
furin Tricas
furin Lymsta
furin Lotgig
furin Capsp1
furin Nemvec
furin Caeele
furin Brumal
furin-1 Cioint
furin-2 Cioint
furin Homsap
furin Musmus
furin Galgal
furinC Danrer
furinB Danrer
furinA Danrer
PC4 Homsap
PC4 Musmus
furin/PC4
PC-A Aplcal
PC-A Lotgig
PC-A Capsp1
PC-A Caeele
PC-A Brumal
PC-A Tricas
PC-A Nasvit
PC-A Pedhumcor
PC-A Dromel
PC-A Dappul
PC1 Homsap
PC1 Musmus
PC1 Danrer
PC1 Braflo
PC1 Brumal
PC1 Apimel
PC1 Nasvit
PC1 Tricas
PC1 Pedhumcor
PC1 Dappul
PC1 Capsp1
PC1 Lotgig
PC1 Triadh
PC1 Strpur
PC1 Nemvec
PC1 Hydmag
PC1
PC-B Ampque
PC-B Triadh
PC-B Nemvec
PC-B Hydmag
kexin Monbre
kexin Schpom
kexin Saccer
PC7 Ampque
PC7 Nemvec
PC7 Hydmag
PC7 Triadh
PC7 Homsap
PC7 Musmus
PC7 Galgal
PC7a Danrer
PC7b Danrer
PC7-1 Strpur
PC7-1 Braflo
PC7-2 Braflo
PC7-2 Strpur
PC7 Cioint
PC7-3 Braflo
PC7 Capsp1
PC7 Lotgig
PC7/kexin
PC-C Dappul
PC-C Hydmag

Figure S8.9.2: Rooted phylogenetic tree of the prohormone convertases (PC). A Maximum-likelihood (ML) tree produced with PHYML is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Midpoint rooting has been used. Nodes that define the different subfamilies are indicated by red circles and are supported by bootstrap values superior to 90% (150 replicates). There are 5 *PC2* genes in *Amphimedon* but no *PC1* gene. *Ampque = Amphimedon queenslandica*; *Apimel = Apis mellifera*; *Aplcal = Aplysia californica*; *Braflo = Branchiostoma floridae*; *Brumal = Brugia malayi*; *Danrer = Danio rerio*; *Caeele = Caenorhabditis elegans*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Dappul = Daphnia pulex*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Halasi = Haliotis asinina*; *Homsap = Homo sapiens*; *Hydmag = Hydra magnipapillata*; *Lotgig = Lottia gigantea*; *Lymsta = Lymnaea stagnalis*; *Monbre = Monosiga brevicollis*; *Musmus = Mus musculus*; *Nasvit = Nasonia vitripennis*; *Nemvec = Nematostella vectensis*; *Pedhumcor = Pediculus humanus corporis*; *Pladum = Platynereis dumerilii*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Tricas = Tribolium castaneum*.

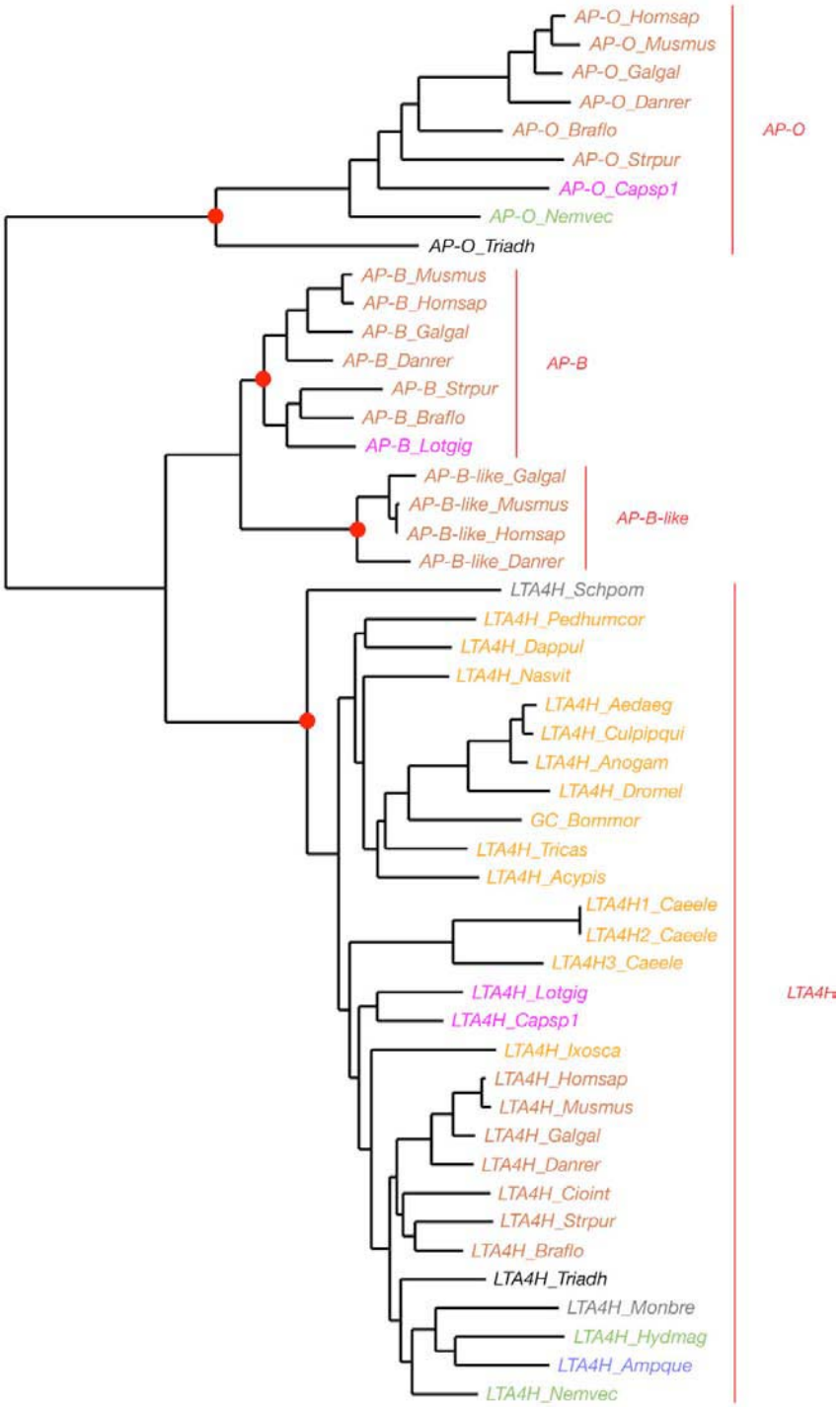


Figure S8.9.3: Rooted phylogenetic tree of the arginyl aminopeptidase B (AP-B), arginyl aminopeptidase O (AP-O), and leukotriene A4 hydrolase (LTA4H) proteins. A Maximum-likelihood (ML) tree produced with PHYML is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Midpoint rooting has been used. Nodes that define the different subfamilies are indicated by red circles and are supported by bootstrap values superior to 80% (150 replicates). There is a single *LTA4H* gene in *Amphimedon* but neither *AP-B* nor *AP-O* genes. *Acypis = Acyrthosiphon pisum*; *Aedaeg = Aedes aegypti*; *Ampque = Amphimedon queenslandica*; *Anogam = Anopheles gambiae*; *Bommor =*

Bombyx mori; *Braflo = Branchiostoma floridae*; *Caeele = Caenorhabditis elegans*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Culqui = Culex quinquefasciatus*; *Danrer = Danio rerio*; *Dappul = Daphnia pulex*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Hydmag = Hydra magnipapillata*; *Ixosca = Ixodes scapularis*; *Lotgig = Lottia gigantea*; *Monbre = Monosiga brevicollis*; *Musmus = Mus musculus*; *Nasvit = Nasonia vitripennis*; *Nemvec = Nematostella vectensis*; *Pedhumcor = Pediculus humanus corporis*; *Schpom = Schizosaccharomyces pombe*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Tricas = Tribolium castaneum*.

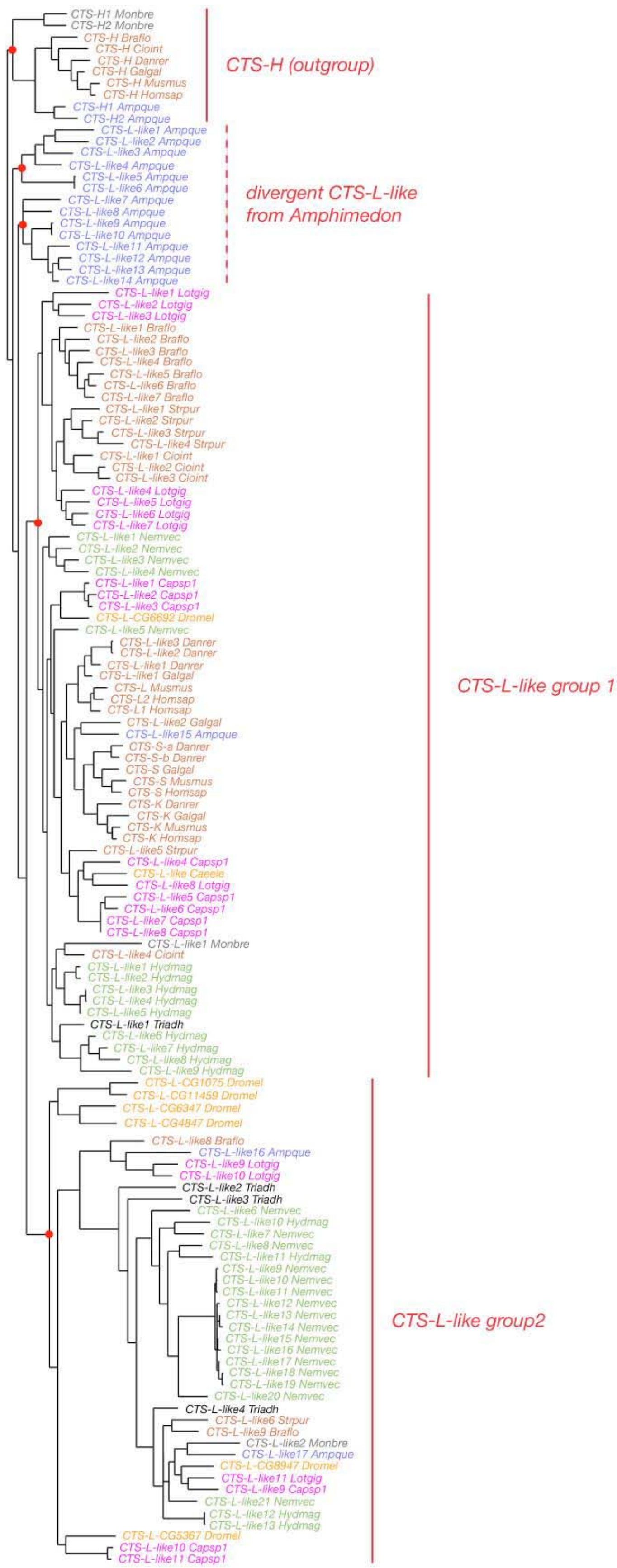

CTS-H1 Monbre
CTS-H2 Monbre
CTS-H Braflo
CTS-H Cioint
CTS-H Danrer
CTS-H Galgal
CTS-H Musmus
CTS-H Homsap
CTS-H1 Ampque
CTS-H2 Ampque
CTS-H (outgroup)
CTS-L-like1 Ampque
CTS-L-like2 Ampque
CTS-L-like3 Ampque
CTS-L-like4 Ampque
CTS-L-like5 Ampque
CTS-L-like6 Ampque
CTS-L-like7 Ampque
CTS-L-like8 Ampque
CTS-L-like9 Ampque
CTS-L-like10 Ampque
CTS-L-like11 Ampque
CTS-L-like12 Ampque
CTS-L-like13 Ampque
CTS-L-like14 Ampque
divergent CTS-L-like
from Amphimedon
CTS-L-like1 Lotgig
CTS-L-like2 Lotgig
CTS-L-like3 Lotgig
CTS-L-like1 Braflo
CTS-L-like2 Braflo
CTS-L-like3 Braflo
CTS-L-like4 Braflo
CTS-L-like5 Braflo
CTS-L-like6 Braflo
CTS-L-like7 Braflo
CTS-L-like1 Strpur
CTS-L-like2 Strpur
CTS-L-like3 Strpur
CTS-L-like4 Strpur
CTS-L-like1 Cioint
CTS-L-like2 Cioint
CTS-L-like3 Cioint
CTS-L-like4 Lotgig
CTS-L-like5 Lotgig
CTS-L-like6 Lotgig
CTS-L-like7 Lotgig
CTS-L-like1 Nemvec
CTS-L-like2 Nemvec
CTS-L-like3 Nemvec
CTS-L-like4 Nemvec
CTS-L-like1 Capsp1
CTS-L-like2 Capsp1
CTS-L-like3 Capsp1
CTS-L-CG6692 Dromel
CTS-L-like5 Nemvec
CTS-L-like3 Danrer
CTS-L-like2 Danrer
CTS-L-like1 Danrer
CTS-L-like1 Galgal
CTS-L Musmus
CTS-L2 Homsap
CTS-L1 Homsap
CTS-L-like2 Galgal
CTS-L-like15 Ampque
CTS-S-a Danrer
CTS-S-b Danrer
CTS-S Galgal
CTS-S Musmus
CTS-S Homsap
CTS-K Danrer
CTS-K Galgal
CTS-K Musmus
CTS-K Homsap
CTS-L-like5 Strpur
CTS-L-like4 Capsp1
CTS-L-like Caeele
CTS-L-like8 Lotgig
CTS-L-like5 Capsp1
CTS-L-like6 Capsp1
CTS-L-like7 Capsp1
CTS-L-like8 Capsp1
CTS-L-like1 Monbre
CTS-L-like4 Cioint
CTS-L-like1 Hydmag
CTS-L-like2 Hydmag
CTS-L-like3 Hydmag
CTS-L-like4 Hydmag
CTS-L-like5 Hydmag
CTS-L-like1 Triadh
CTS-L-like6 Hydmag
CTS-L-like7 Hydmag
CTS-L-like8 Hydmag
CTS-L-like9 Hydmag
CTS-L-like group 1
CTS-L-CG1075 Dromel
CTS-L-CG11459 Dromel
CTS-L-CG6347 Dromel
CTS-L-CG4847 Dromel
CTS-L-like8 Braflo
CTS-L-like16 Ampque
CTS-L-like9 Lotgig
CTS-L-like10 Lotgig
CTS-L-like2 Triadh
CTS-L-like3 Triadh
CTS-L-like6 Nemvec
CTS-L-like10 Hydmag
CTS-L-like7 Nemvec
CTS-L-like8 Nemvec
CTS-L-like11 Hydmag
CTS-L-like9 Nemvec
CTS-L-like10 Nemvec
CTS-L-like11 Nemvec
CTS-L-like12 Nemvec
CTS-L-like13 Nemvec
CTS-L-like14 Nemvec
CTS-L-like15 Nemvec
CTS-L-like16 Nemvec
CTS-L-like17 Nemvec
CTS-L-like18 Nemvec
CTS-L-like19 Nemvec
CTS-L-like20 Nemvec
CTS-L-like group2
CTS-L-like4 Triadh
CTS-L-like6 Strpur
CTS-L-like9 Braflo
CTS-L-like2 Monbre
CTS-L-like17 Ampque
CTS-L-CG8947 Dromel
CTS-L-like11 Lotgig
CTS-L-like9 Capsp1
CTS-L-like21 Nemvec
CTS-L-like12 Hydmag
CTS-L-like13 Hydmag
CTS-L-CG5367 Dromel
CTS-L-like10 Capsp1
CTS-L-like11 Capsp1

Figure S8.9.4: Rooted phylogenetic tree of the cathepsin-L (CTS-L) proteins. A Maximum-likelihood (ML) tree produced with PHYML[40] is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). The tree has been rooted using the closest cathepsin subfamily, CTS-H. Nodes that define the different subfamilies are indicated by red circles and are supported by bootstrap values superior to 90% (150 replicates). There are 17 *Amphimedon* genes that group with *CTS-L* genes from eumetazoans: 14 of them form 2 *Amphimedon*-specific groups of divergent *CTS-L* genes; the 3 other ones are included into 2 groups of *CTS-L-like* genes that include all the *CTS*-L genes from eumetazoans. *Ampque = Amphimedon queenslandica*; *Braflo = Branchiostoma floridae*; *Caeele = Caenorhabditis elegans*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Danrer = Danio rerio*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Hydmag = Hydra magnipapillata*; *Lotgig = Lottia gigantea*; *Monbre = Monosiga brevicollis*; *Musmus = Mus musculus*; *Nemvec = Nematostella vectensis*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*.

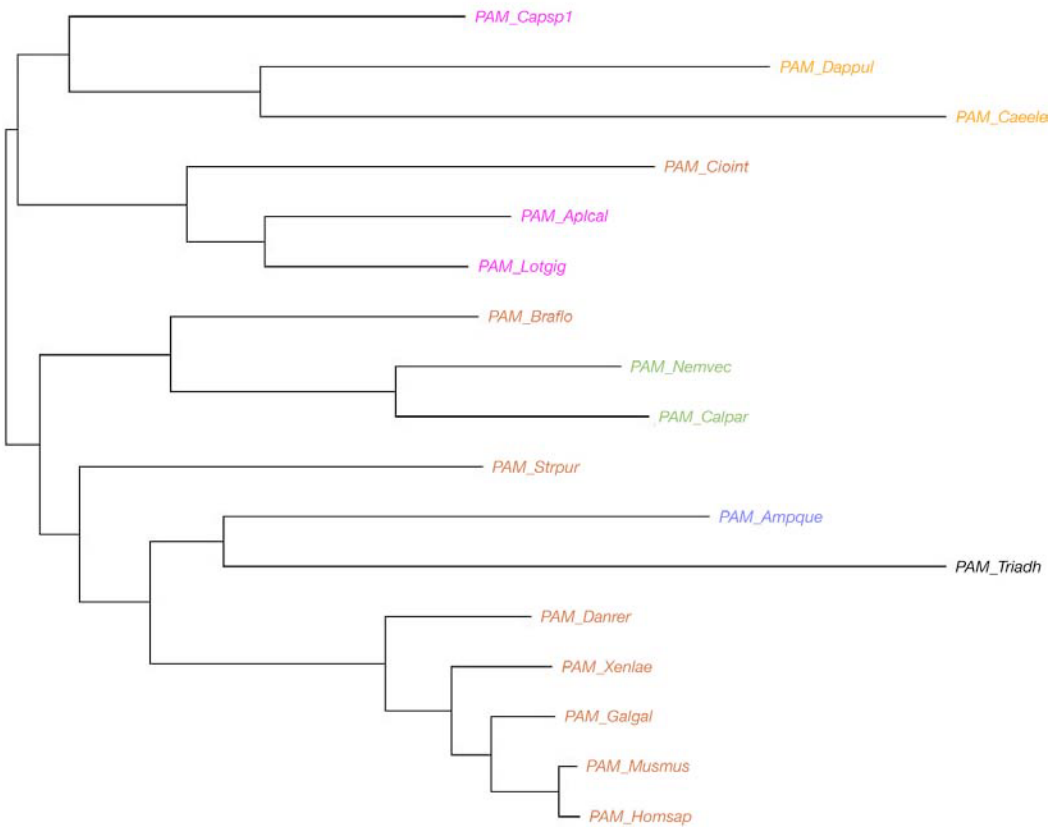


Figure S8.9.5: Rooted phylogenetic tree of the peptidylglycine α-amidating monooxygenase (PAM) proteins. A Maximum-likelihood (ML) tree produced with PHYML[40] is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Midpoint rooting has been used. All the nodes of the tree are supported by bootstrap values superior to 95% (150 replicates). There is a single *PAM* gene in *Amphimedon*. *Ampque = Amphimedon queenslandica*; *Braflo = Branchiostoma floridae*; *Caeele = Caenorhabditis elegans*; *Calpar = Calliactis parasitica*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Danrer = Danio rerio*; *Dappul = Daphnia pulex*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Lotgig = Lottia gigantea*; *Musmus = Mus musculus*; *Nemvec = Nematostella vectensis*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Xenlae = Xenopus laevis*.

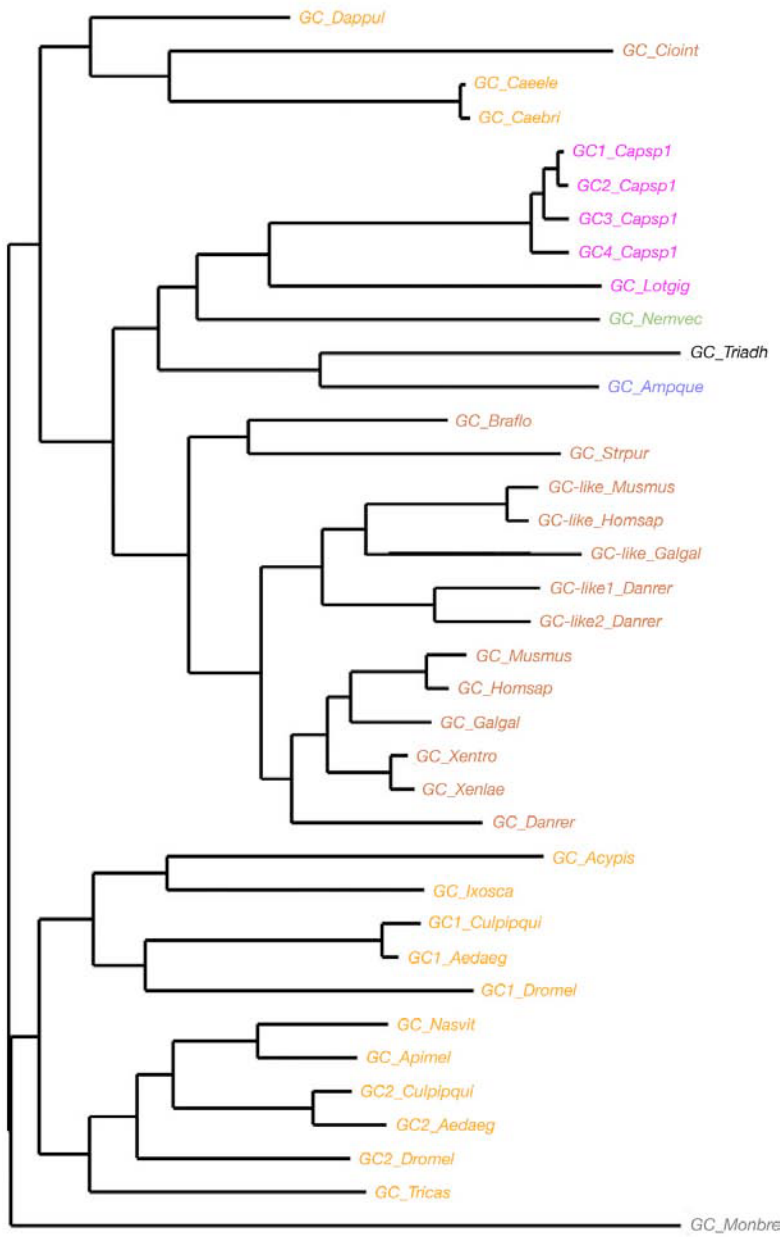


Figure S8.9.6: Rooted phylogenetic tree of the glutaminyl-peptide cyclotransferase (GC) proteins. A Maximum-likelihood (ML) tree produced with PHYML[40] is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Midpoint rooting has been used. Most of the nodes of the tree are supported by bootstrap values superior to 70% (150 replicates). There is a single *GC* gene in *Amphimedon*. *Acypis = Acyrthosiphon pisum*; *Aedaeg = Aedes aegypti*; *Ampque = Amphimedon queenslandica*; *Apimel = Apis mellifera*; *Aplcal = Aplysia californica*; *Braflo = Branchiostoma floridae*; *Caebri = Caenorhabditis briggsae*; *Caeele = Caenorhabditis elegans*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Culpipqui = Culex quinquefasciatus*; *Danrer = Danio rerio*; *Dappul = Daphnia pulex*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Ixosca = Ixodes scapularis*; *Lotgig = Lottia gigantea*; *Monbre = Monosiga brevicollis*; *Musmus = Mus musculus*; *Nasvit = Nasonia vitripennis*; *Nemvec = Nematostella vectensis*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Tricas = Tribolium castaneum*; *Xenlae = Xenopus laevis*; *Xentro = Xenopus tropicalis.*

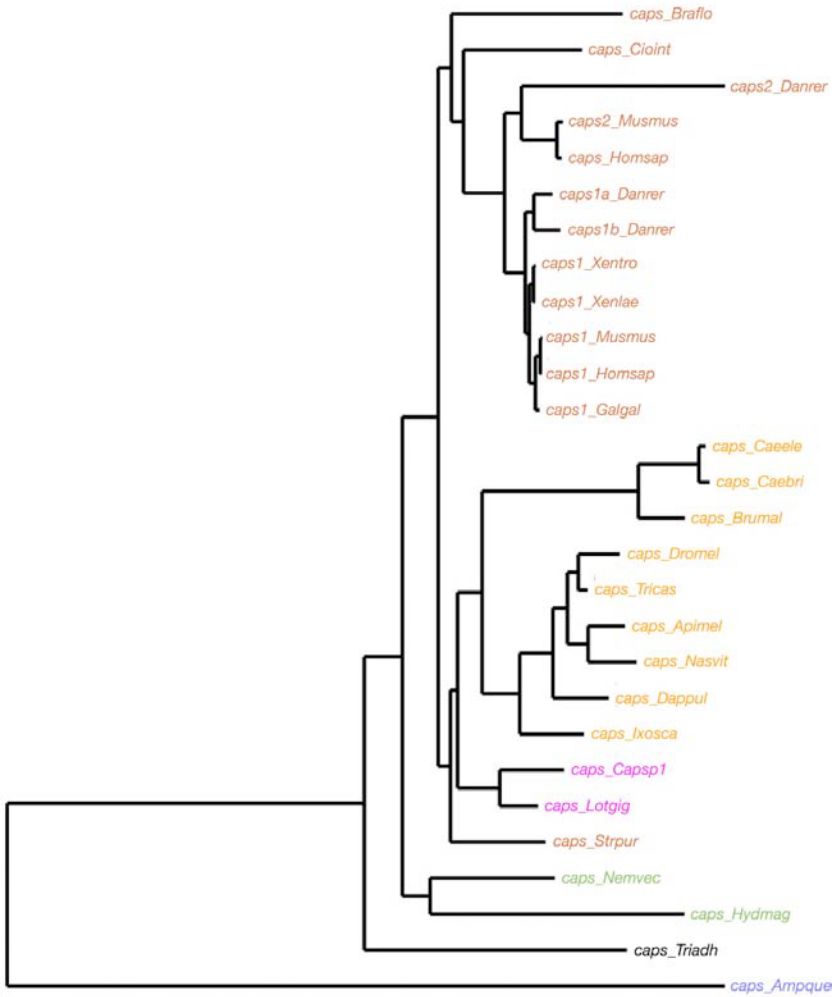


Figure S8.9.7: Rooted phylogenetic tree of the calcium activated protein for secretion (caps) proteins. A Maximum-likelihood (ML) tree produced with PHYML[40] is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). Midpoint rooting has been used. Most of the nodes of the tree are supported by bootstrap values superior to 75% (150 replicates). There is a single *caps* gene in *Amphimedon*. *Acypis = Acyrthosiphon pisum*; *Ampque = Amphimedon queenslandica*; *Anogam = Anopheles gambiae*; *Apimel = Apis mellifera*; *Braflo = Branchiostoma floridae*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Danrer = Danio rerio*; *Dappul = Daphnia pulex*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Ixosca = Ixodes scapularis*; *Lotgig = Lottia gigantea*; *Musmus = Mus musculus*; *Nasvit = Nasonia vitripennis*; *Nemvec = Nematostella vectensis*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Tricas = Tribolium castaneum*; *Xentro = Xenopus tropicalis*.

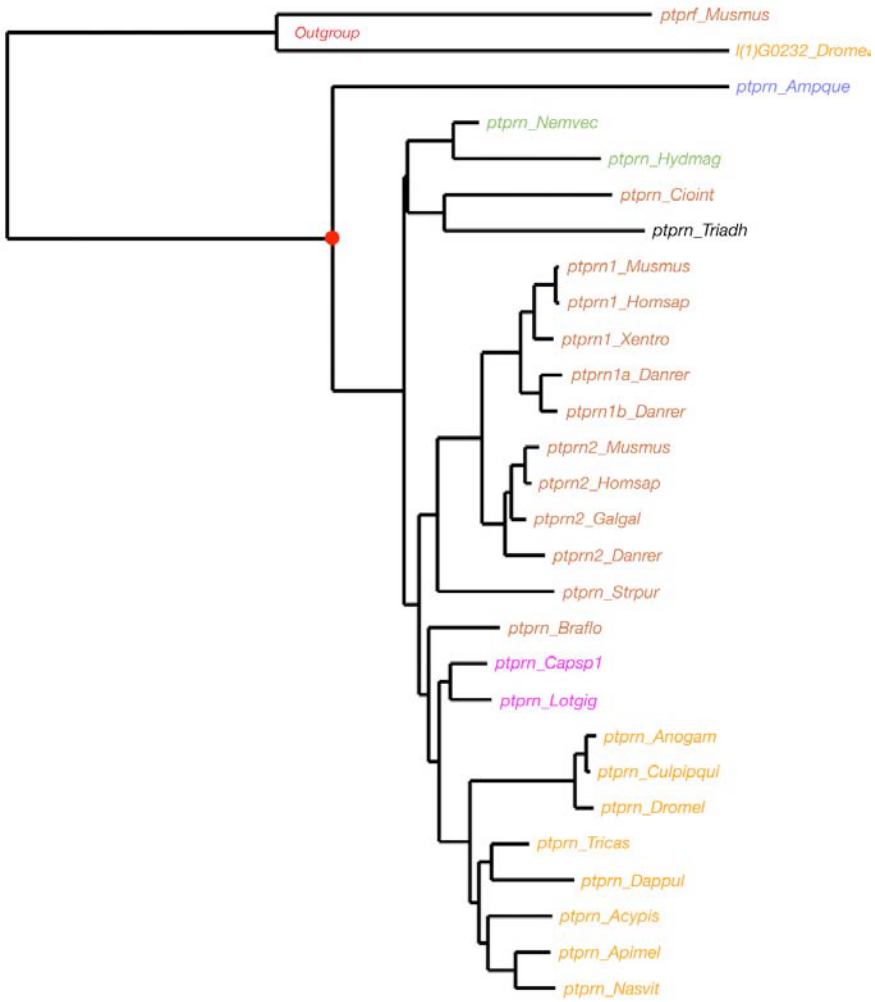


Figure S8.9.8: Rooted phylogenetic tree of the protein tyrosine phosphatase receptor type N (ptprn) proteins. A Maximum-likelihood (ML) tree produced with PHYML[40] is shown. PHYML analyses were performed using the WAG amino-acid substitution model, the frequencies of amino acids being estimated from the data set, and rate heterogeneity across sites being modelled by two rate categories (one constant and eight γ-rates). The tree has been rooted using the closest protein tyrosine phosphatase receptors from mouse and *Drosophila*. The node defining the *ptprn* subfamily is indicated by a red circle and is supported by a bootstrap value of 100% (150 replicates). There is a single *ptprn* gene in *Amphimedon*. *Ampque = Amphimedon queenslandica*; *Apimel = Apis mellifera*; *Braflo = Branchiostoma floridae*; *Brumal = Brugia malayi*; *Caebri = Caenorhabditis briggsae*; *Caeele = Caenorhabditis elegans*; *Capsp1 = Capitella sp I*; *Cioint = Ciona intestinalis*; *Culpipqui = Culex quinquefasciatus*; *Danrer = Danio rerio*; *Dappul = Daphnia pulex*; *Dromel = Drosophila melanogaster*; *Galgal = Gallus gallus*; *Homsap = Homo sapiens*; *Hydmag = Hydra magnipapillata*; *Lotgig = Lottia gigantea*; *Musmus = Mus musculus*; *Nasvit = Nasonia vitripennis*; *Nemvec = Nematostella vectensis*; *Strpur = Strongylocentrotus purpuratus*; *Triadh = Trichoplax adhaerens*; *Tricas = Tribolium castaneum*; *Xenlae = Xenopus laevis*; *Xentro = Xenopus tropicalis.*

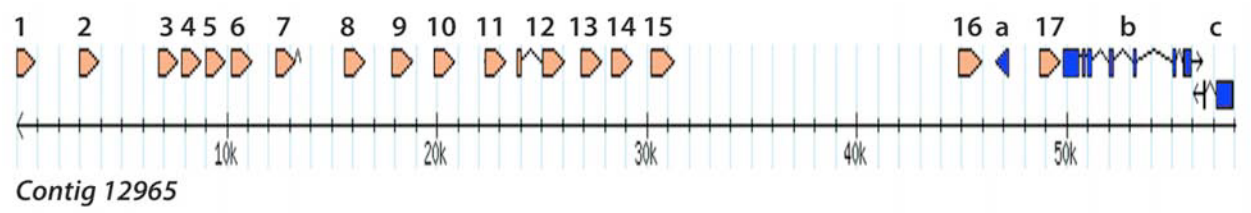


Fig. S8.9.9. An example tandem array of Rhodopsin GPCR genes in *Amphimedon*. 16 single exon genes and one 2-exon gene arranged in a head–to-tail manner; 3 non-GPCR genes are labeled a-c are in blue. Contig 12965 is only 60 kb suggesting this array may be larger. Gene model identifiers: Aqu1.212154-212167 (excluding 212161) and g.13855.t1-13859.t1.

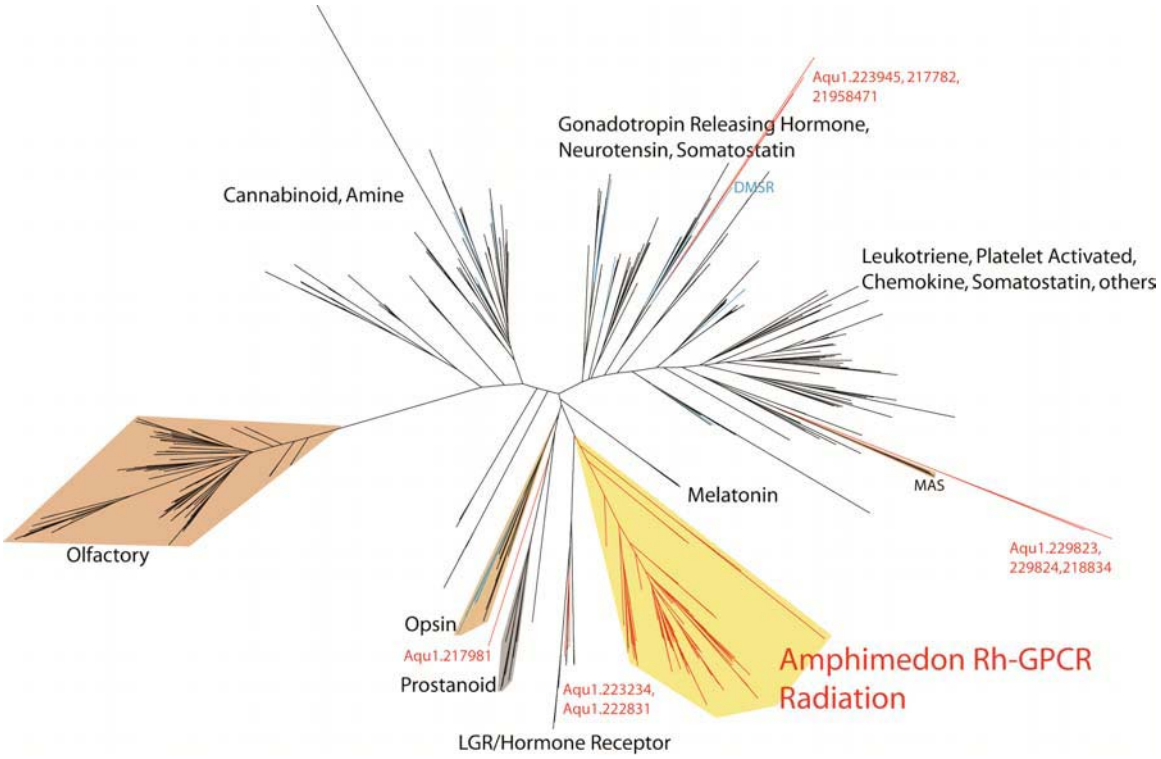


Fig. S8.9.10. Estimated phylogenetic position of 138 Rhodopsin-class GPCR genes of *Amphimedon* using Maximum Likelihood. Most (129) *Amphimedon* GPCR genes (red branches), including those of the tandem array illustrated in Figure S8.9.9, form a distinct cluster. All human and fly genes were downloaded from the curated database GPCRDB version 8.10.6 (Horn et al 2003). GPCR sequences were obtained from the Aqu1 filtered gene models using blastp searches (retaining the 100 best hits) with the following human GPCRs as queries: mtr1a, cxcr7,ta2r;o43898, aa2ar, tshr, ptafr, hrh1, q8nh44, ffar3, gpbar, q5juh7, q59er8, mrgx2, cltr1, qrfpr, opsx, q5ku28, gnrr2, gpr85, q5jrh7. Only non-redundant sequences longer than 100 amino acids were included in the phylogenetic analysis. Sequences from all 3 species were then aligned to a Hidden Markov Model trained on the rhodopsin class GPCR alignment (4,993 rhodopsin class GPCRs from various animal phyla) curated at GPCRDB (Horn et al 2003) using HMMR3 (Eddy 1998). This alignment, which includes 918 sequences and 245 positions, was used to estimate the illustrated unrooted gene tree. Tree construction was done under maximum likelihood, implemented in RaxML (Stamatakis, 2006) and assuming the WAG+G+F model. Other phylogenetic analyses (not shown) using selected non-rhodopsin GPCRs as outgroups supported the assignment of sponge genes shown here to the rhodopsin class of GPCR's.

Table S8.10.1: Classification of allorecognition and innate immunity signaling pathway genes by origin.

Gene/Family	Origin	Methods used to determine origin
Nod-like receptor (NLR)	metazoan	NLR candidates are not found outside of animals. These genes have a tripartite composition consisting of an NACHT and LRR generally associated with different N-terminal death-fold domains. Vertebrates utilize CARD or PYD. *Amphimedon* has a large repertoire of NLRs and uses the death domain. NLRs were not detected in *Trichoplax*. *Nematostella* has NLR candidates that consists of NACHT-LRR, some of which might be partial predictions, being located on the edge of assembled contigs), and possibly DED-NACHT-LRR (JGI gene model ID 214132-3). *Strongylocentrotus* has a large NLR repertoire and uses the death domain.[213] *Amphioxus* has CARD/DED/death domain-NACHT-LRR.[101] (Also see S9.3.1 and Fig 2a).
Multiple SRCR receptors	eukaryotic	Multiple SRCR receptors are ancient[148] but their association with domains such as the fibrinogen or the complement control protein domain is an animal novelty.[149]
TLR/IL1R-like	metazoa	Poriferans (*Amphimedon* and *Suberites*) encode putative receptors related to TLRIL1R superfamily[150](Gauthier, in prep). *Nematostella* has true TLRs with LRRs.[186] *Hydra* has LRR and TIR domain on separate proteins that are reported to interact on cell membrane.[214] TLR-like proteins not found in *Trichoplax*.
IL1R-like	deuterostome	(see Table S8.3.1)
IRAK	metazoan	The kinase domain found in IRAK proteins forms a clade within an ancient eukaryotic kinase family. Its association with an N-terminal death domain is an animal novelty.
MyD88	metazoan	The combination of a TIR and a death domain is an animal novelty. Missing in *Trichoplax.*
TRAF	eukaryotic	TRAF candidates are present in *Dictyostelium* with a complete RING-type zinc finger, a TRAF zinc finger and a MATH domain arrangement. The portion that contains the zinc fingers is homologous to TRAF6 but the MATH domain is homologous to Speckle/POZ.[215] TRAFs are otherwise found in all the animal lineages surveyed. A lineage-specific expansion of the TRAF family seems to have occurred in *Amphimedon.*
NIK	vertebrate	No hit outside of vertebrates.
IKK	metazoan	No hit found outside of metazoans.
NFkB	holozoan	An NFkB candidate detected in *Capsaspora owczarzaki* genomic traces[126] but not in *Monosiga brevicollis* and other non-animal groups surveyed. NFkB is otherwise found in *Amphimedon* and *Nematostella*[216,217]. *Trichoplax* has one putative protein with a highly divergent RHD.
Rel	bilaterian	No hit found outside of bilaterians.
JNK	metazoan	See Table S8.5.2 and Table S8.7.2
CD36/ LMPII-like	eukaryotic	CD36 domain-containing proteins detected in *Paramecium*, *Dictyostelium*, *Monosiga*, *Amphimedon*, *Nematostella* and *Trichoplax*. Not detected in plants or fungi.
C3-C5 like	eumetazoan s.s.	Found in *Nematostella* but not in *Trichoplax*; no hit in *Amphimedon*, *Monosiga* or other non-animal group. The *Nematostella* candidate shows weak homology to the Anato domain found in bilaterian proteins.
C6-9 like	vertebrate	Multidomain protein comprised of ancient domains and vertebrate-specific FIMAC domain.
Bf/ C2-like	eumetazoan s.s.	Eumetazoan multidomain protein comprised of ancient domains (CCP, VWA, serine protease domain). Found in *Nematostella* but not in *Trichoplax*.
C1q/ MBL	vertebrate	The Collagen triple helix repeat and the C-type lectin domain found outside of animal group but the C1q domain is deuterostome-specific. The combination of a Collagen triple helix repeat with a C1q domain (in C1q) or with a C-type lectin domain (in MBL) is a vertebrate innovation. Millectin, a derived MBL that lacks a collagen domain, is found in *Acropora*.[218]
A2M	eumetazoan	Eumetazoan-specific. Found in *Nematostella* and *Trichoplax* but absent from *Amphimedon* and non-animal groups.
MASP	eumetazoan s.s.	Present in *Nematostella* but appears to be absent in *Trichoplax*. Not found in *Amphimedon* and non-animal groups.
PGLYRP	bilaterian	PGRP domain-containing protein found in molluscs, echinoderms, arthropods and vertebrates. Not found in non-animal groups,

		Amphimedon, *Nematostella*, *Trichoplax*, or *Ciona*. Belongs to the type 2 amidase family that is also found in bacteriophage.
β propeller lectin	eukaryotic	Found in *Physarum* (*e.g.* tectonin), *Amphimedon*, *Nematostella*, *Hydractinia* and *Branchiostoma.* Vertebrate sequence equipped with additional domains[219]; no hit found in *Monosiga* or *Trichoplax*. The antimicrobial properties of β propeller lectins have only been observed in animals, the role of tectonin being restricted to phagocytocis in *Physarum.*[220]
C-type lectin domain containing receptors	eukaryotic	C-type lectin domain-containing proteins detected in all kingdoms. Present in *Monosiga*, *Amphimedon*, *Trichoplax* and *Nematostella* (the latter has an expanded repertoire; see Wood-Charlson 2009 [221].
BGBP	eukaryotic	Member of the glucoside hydrolase 16 superfamily that is present in protists, fungi, *Amphimedon*, *Suberites,*[222] molluscs, arthropods, deuterostomes, urochordates and cephalochordates but not in vertebrates. Not found in *Monosiga, Nematostella* or *Trichoplax* but detected in *Hydra*.
MPEG1	metazoan	Belongs to the MACPF superfamily that occurs across eukaryotes and in bacteria. The perforin subtype is animal-specific. Present in *Amphimedon*, *Suberites*[150] and *Nematostella.*[186] Not detected in *Trichoplax.*
ProPO/ Hemocyanin-like	bilaterian	ProPo is part of the invertebrate immune defence and belongs to the Type 3 copper family that is found across eukaryotes. Putative proteins with domains specific to hemocyanin/ prophenoloxidase are encoded in fungi (*Aspergillus*) and some animals (*Amphimedon*, arthropods and urochordates) but not in *Monosiga*, *Nematostella*, *Trichoplax* and vertebrates. However, the sponge and fungal candidates lack the C domain.
25OAS	eukaryotic	Putative proteins related to the 25OAS are predicted in non-animal genomes (*Monosiga* and *Chlamydomonas)*.[223] Present in sponges (*Amphimedon*, *Lubomirskia*, *Geodia* and *Suberites*[224]), and *Nematostella* but not in *Trichoplax*.
DICER	eukaryotic	Found in plants, fungi and animals (detected in *Amphimedon*, *Nematostella*, *Trichoplax*).[225] *Monosiga* appears to have other helicases but not DICER.
IFIH1/ DDX58-like	metazoan but CARD-helicase combo bilat specific	The combination of 1 or 2 CARD + helicase ATP binding domain + helicase C-terminal domain is bilaterian-specific. Sponge has death domain instead of CARD, *Nematostella* candidates appear to lack N-terminal death-fold domains but models might be fragmented. Not detected in *Trichoplax*. Not found in *Monosiga* or other non-animal groups.
IRF	metazoan	Metazoan-specific[226] but IRF-like domains present in viral proteins.
IL	vertebrate	Vertebrate-specific but IL-1 receptor antagonists found in bacteria and possibly viruses.
INF	vertebrate	Vertebrate-specific.
AF	sponge-specific	The aggregation factors are sponge-specific extracellular proteoglycans. In *Microciona prolifera,* the central core protein structure is sunburst-like and composed of multiple units of two proteins that form the central ring and the radiating arms, respectively. These two proteins are encoded by a single gene (*MpAF3c*).[227] *Amphimedon* has one homolog to *MpAF3c*, which clusters with other *AF*-like genes in the genome. No other animal representatives appear to encode a similar polypeptide. However, candidates with a Calx-β domain arrangement similar to that of the protein forming the radiating arm are predicted in bacteria and *Nematostella.*

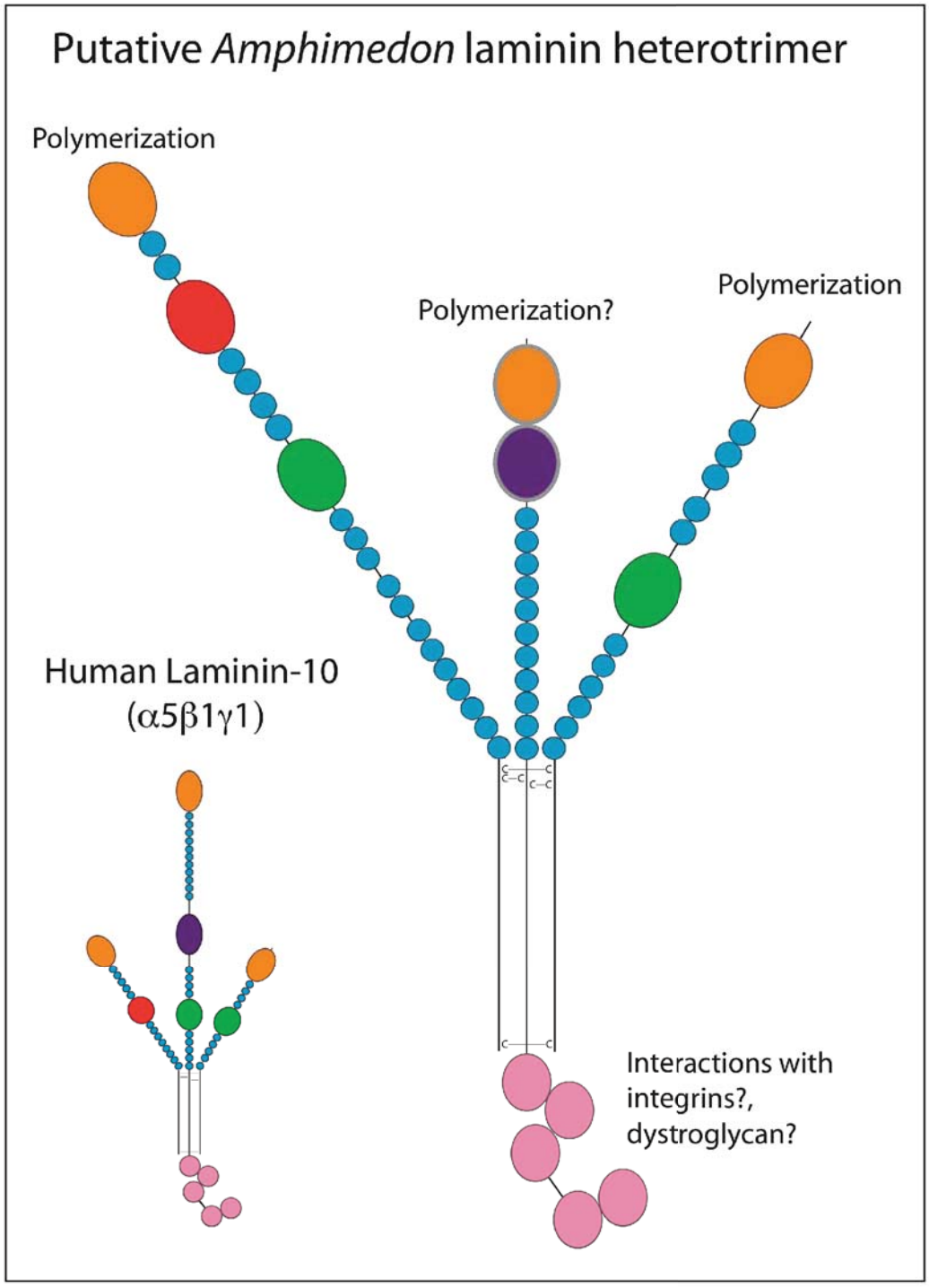


Figure S9.3.1: Hypothetical assembly of *Amphimedon* laminins. Schematic diagram of a putative *Amphimedon* laminin heterotrimer structure. The trimer contains the three *Amphimedon* laminin chains which display closest resemblance to the α, β, and γ chains of characterized bilaterian laminins. The most similar full length mammalian laminin is shown as an inset for comparison. Diagrams are drawn roughly to scale and depict the locations of domains on the primary sequence. The coiled coil regions of all *Amphimedon* laminin chains are approximately the same length and possess putative interchain disulfide forming cysteines at the N- and C-termini suggesting that they have the potential to form heterotrimers *in vivo*.

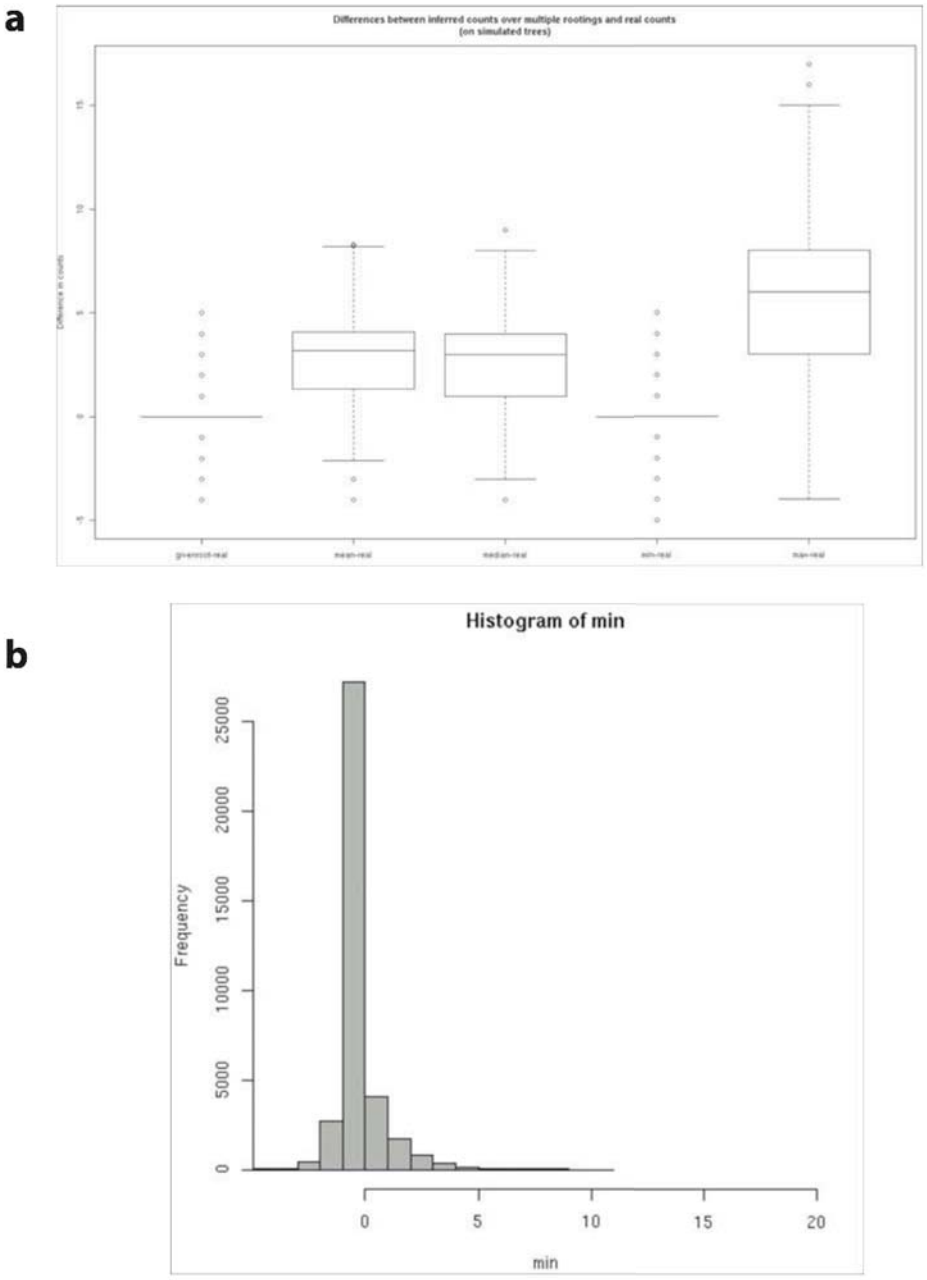


Figure S10.1.1: The rooting with the minimum count is the best approximation of the true count in simulation studies. (a) Boxplot showing the differences in numbers of subfamilies inferred using the assumed root of the tree (given), the rootings with the minimum, maximum, mean and median counts and the known real numbers of ancestral subfamilies. The data points for all the internal nodes of the species tree and simulations ranging over varied levels of duplication, pruning and deresolution rates were pooled together for this analysis. **(b)** A histogram of the differences between the minimum estimate and the known correct number of subfamilies (min) showing that only a small fraction deviate from a difference of zero.

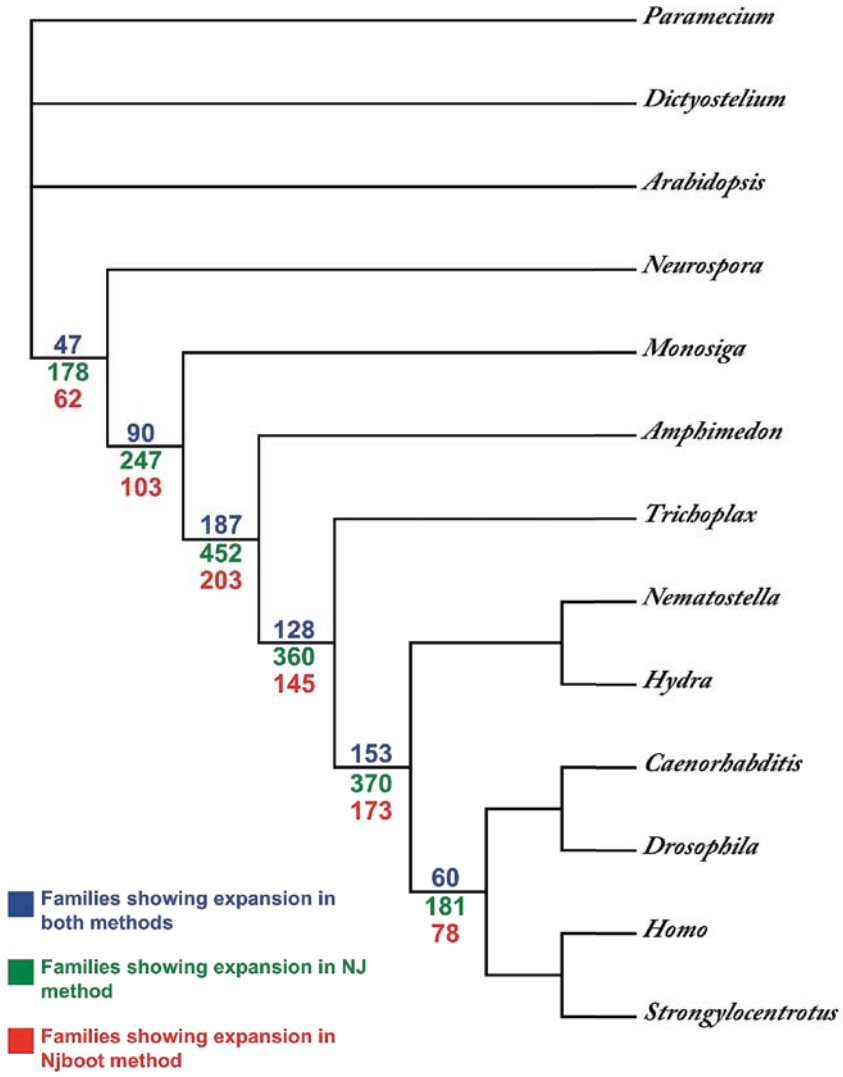


Figure 310.1.2: Subfamily expansions in animal evolution. A tree indicating relationships of species used in the subfamily analysis is shown. The stems are decorated with numbers of gene families (of the 725 families included in the analysis) that are inferred to have expanded using the neighbor joining (green) and the neighbor joining with bootstrap (red) methods, and using both methods (blue).

Table S10.2.1: p-values for significantly linked paralog pairs generated at different animal evolution nodes in 725 gene families.

stem	species	observed linked pairs (x)	total pairs	percent linked	# random experiments	# experiments with at least x pairs linked	p-value
holozoan	human	27	427	0.06323185	10000	1633	0.1633
holozoan	*Nematostella*	26	1516	0.017150396	10000	0	0
holozoan	*Trichoplax*	43	257	0.167315175	10000	0	0
holozoan	*Amphimedon*	8	572	0.013986014	10000	0	0
holozoan	*Monosiga*	11	268	0.041044776	10000	3059	0.3059
metazoan	human	76	877	0.086659065	10000	0	0
metazoan	*Nematostella*	29	1387	0.020908435	10000	0	0
metazoan	*Trichoplax*	228	998	0.228456914	10000	0	0
metazoan	*Amphimedon*	87	3832	0.022703549	10000	0	0
eumetazoan	human	79	1144	0.069055944	10000	32	0.0032
eumetazoan	*Nematostella*	53	2637	0.020098597	10000	0	0
eumetazoan	*Trichoplax*	247	834	0.29616307	10000	0	0
cnidarian-bilaterian	human	90	1377	0.065359477	10000	140	0.014
cnidarian-bilaterian	*Nematostella*	27	1164	0.023195876	10000	0	0
bilaterian	human	22	253	0.086956522	10000	115	0.0115

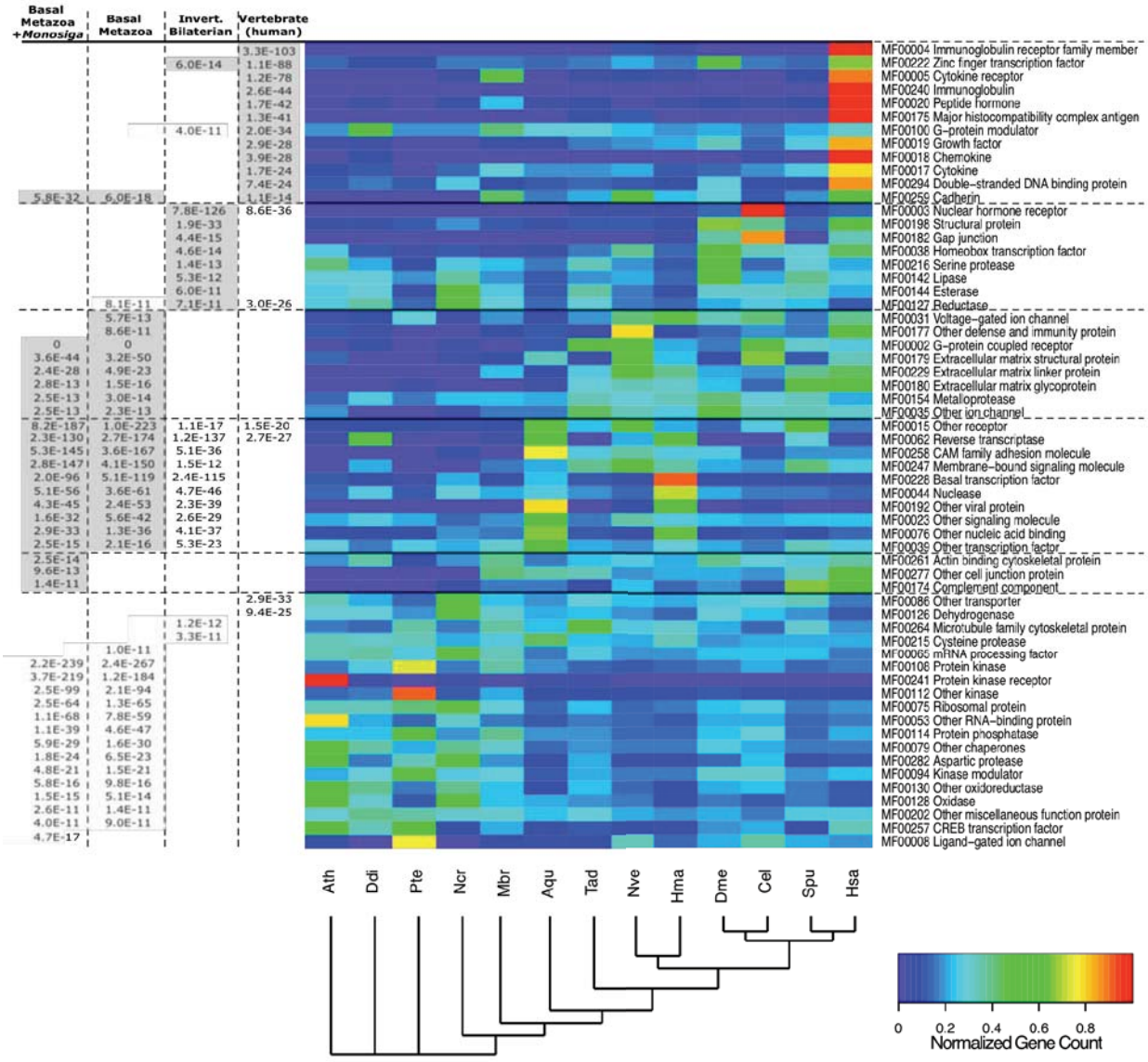


Figure S11.1.1: Heatmap representation of molecular functions enriched or depleted in various complexity groups. Molecular function categories that show significant (1e-10) enrichment or depletion in Fisher's exact tests were selected (Supplementary Note S11). Significance of enrichments (grey background) and depletions (white background) for the three animal complexity groups are indicated in the columns to the left of the heatmap. The heatmap shows normalized gene counts of PANTHER molecular function categories for the species in the analysis (color range: blue 0, red 1). Ath, *Arabidopsis thaliana*; Ddi, *Dictyostelium discoideum*; Pte, *Paramecium tetraurelia*; Ncr, *Neurospora crassa*; Mbr, *Monosiga brevicollis*; Aqu, *Amphimedon queenslandica*; Tad, *Trichoplax adhaerens*; Nve, *Nematostella vectensis*; Hma, *Hydra magnipapillata*; Dme, *Drosophila melanogaster*; Cel, *Caenorhabditis elegans*; Spu, *Strongylocentrotus purpuratus*; Hsa, *Homo sapiens*.

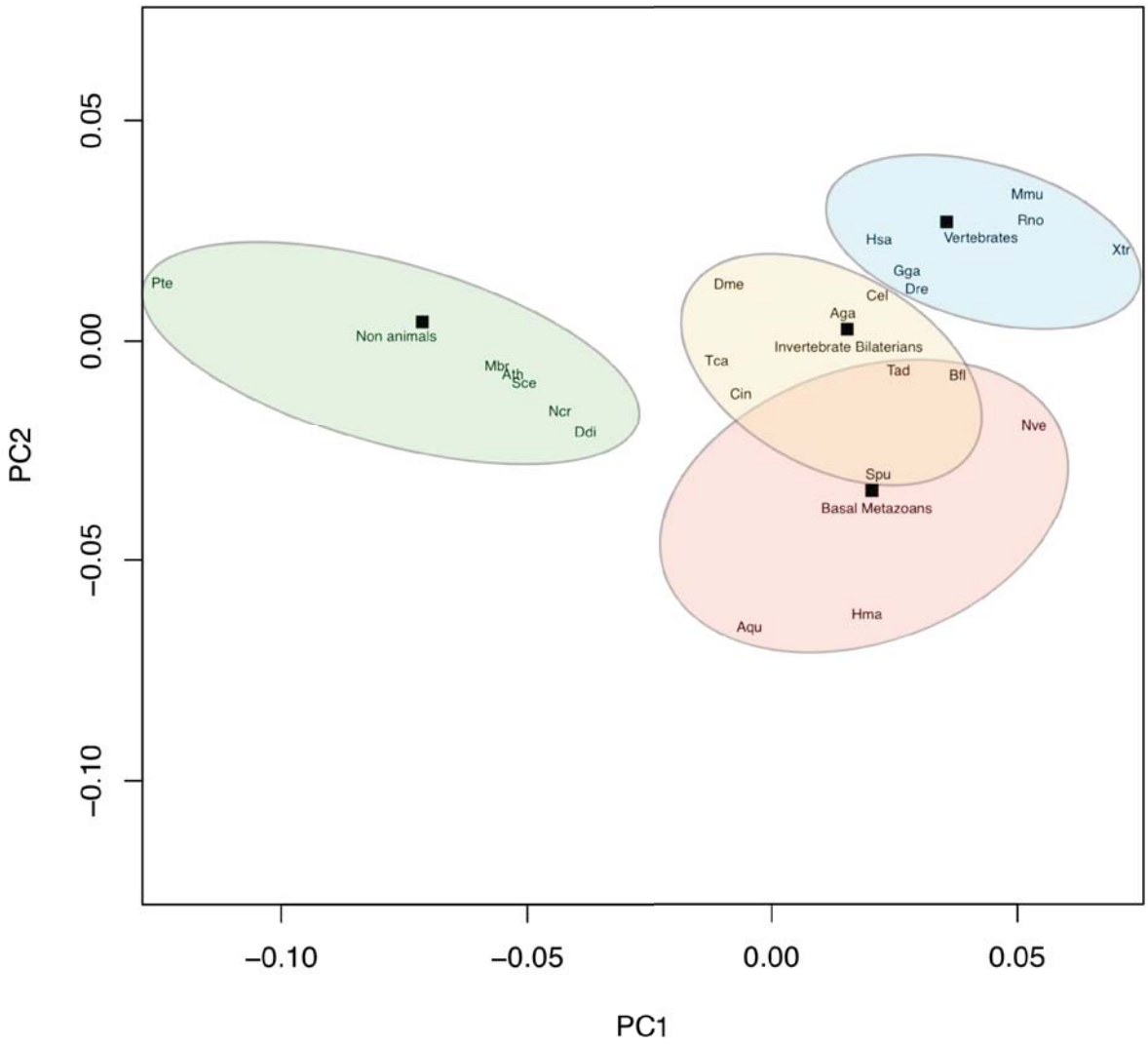


Figure S11.2.1: Projection of species and complexity groups on the first two principal components. Black squares represent the four morphological complexity groups used in the principle components analysis. Species are bound in colored ellipses for the complexity group they belong to. Non-animals, green; basal metazoans, pink; invertebrate bilaterians, yellow; vertebrates, blue. Ath, *Arabidopsis thaliana*; Ddi, *Dictyostelium discoideum*; Pte, *Paramecium tetraurelia*; Ncr, *Neurospora crassa*; Sce, *Saccharomyces cerevisiae;* Mbr, *Monosiga brevicollis*; Aqu, *Amphimedon queenslandica*; Tad, *Trichoplax adhaerens*; Nve, *Nematostella vectensis*; Hma, *Hydra magnipapillata*; Dme, *Drosophila melanogaster*; Cel, *Caenorhabditis elegans*; Tca, *Tribolium castaneum;* Aga, *Anopheles gambiae;* Cin, *Ciona intestinalis*; Bfl, *Branchipstoma floridae*; Spu, *Strongylocentrotus purpuratus*; Hsa, *Homo sapiens*; Dre, *Danio rerio*; Xtr, *Xenopus tropicalis*; Gga, *Gallus gallus*; Rno, *Rattus norvegicus*; Mmu, *Mus musculus*.

References

1. Hooper, J.N.A. & van Soest, R.W.M. A new species of *Amphimedon* (Porifera, Demospongiae, Haplosclerida, Niphatidae) from the Capricorn-Bunker Group of Islands, Great Barrier Reef, Australia: target species for the 'sponge genome project'. *Zootaxa* **1314**, 31-39 (2006).
2. Leys, S.P. & Degnan, B.M. The cytological basis of photoresponsive behavior in a sponge larva. *Biol Bull* **201**, 323-38 (2001).
3. Degnan, B. et al. The Demosponge *Amphimedon queenslandica*: Reconstructing the Ancestral Metazoan Genome and Deciphering the Origin of Animal Multicellularity. in *Emerging Model Organisms: A Laboratory Manual, Volume 1* (Cold Spring Harbor Laboratory Press, 2009).
4. Larroux, C. et al. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol Dev* **8**, 150-73 (2006).
5. Degnan, S.M. & Degnan, B.M. The initiation of metamorphosis as an ancient polyphenic trait and its role in metazoan life cycle evolution. *Philos Trans R Soc Lond B Biol Sci* **365**, 641-651 (2010).
6. Leys, S.P. & Degnan, B.M. Embryogenesis and metamorphosis in a haplosclerid demosponge: gastrulation and transdifferentiation of larval ciliated cells to choanocytes. *Invert. Biol.* **121**, 171-189 (2002).
7. Adamska, M. et al. Wnt and TGF-beta Expression in the Sponge *Amphimedon queenslandica* and the Origin of Metazoan Embryonic Patterning. *PLoS ONE* **2**, e1031 (2007).
8. Sambrook, J. & Russell, D. *Molecular Cloning: A Laboratory Manual (Third Edition)* (Cold Spring Harbor Press, New York, 2001).
9. Degnan, S.M., Craigie, A. & Degnan, B.M. Genotyping individual *Amphimedon* embryos, larvae, and adults. *Cold Spring Harb. Protoc.* (2008).
10. Chapman, J. University of California (2004).
11. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of F*ugu rubripes*. *Science* **297**, 1301-10 (2002).
12. Chapman, J., Simakov, O., Rokhsar, D., David, C.N. & Steele, R.E. The dynamic genome of *Hydra*. *Nature* **464**, 592-6(2010).
13. Small, K.S., Brudno, M., Hill, M.M. & Sidow, A. Extreme genomic variation in a natural population. *Proc Natl Acad Sci U S A* **104**, 5698-703 (2007).
14. Small, K.S., Brudno, M., Hill, M.M. & Sidow, A. A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome Biol* **8**, R41 (2007).
15. Velasco, R. et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
16. Putnam, N.H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).
17. Srivastava, M. et al. The *Trichoplax* genome and the nature of placozoans. *Nature* **454**, 955-60 (2008).
18. Haas, B.J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-66 (2003).
19. Chen, Y.A., Lin, C.C., Wang, C.D., Wu, H.B. & Hwang, P.I. An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics* **8**, 416 (2007).

20. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).
21. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**, 967-74 (1998).
22. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
23. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-86 (2007).
24. Darling, A.E., Carey, L. & Feng, W. The design, implementation, and evaluation of mpiBLAST. in *Proceedings of the 4th International Conference on Linux Clusters and ClusterWorld 2003* (2003).
25. Maldonado, M. Intergenerational transmission of symbiotic bacteria in oviparous and viviparous demosponges, with emphasis on intracytoplasmically-compartmented bacterial types. *J. Mar. Biol. Ass.* **87**, 1701-1713 (2007).
26. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7 Suppl 1**, S11 1-8 (2006).
27. Yeh, R.F., Lim, L.P. & Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res* **11**, 803-16 (2001).
28. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
29. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
30. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
31. Nielsen, C. Six major steps in animal evolution: are we derived sponge larvae? *Evol Dev* **10**, 241-57 (2008).
32. Borchiellini, C. et al. Molecular phylogeny of Demospongiae: implications for classification and scenarios of character evolution. *Mol Phylogenet Evol* **32**, 823-37 (2004).
33. Nichols, S.A. An evaluation of support for order-level monophyly and interrelationships within the class Demospongiae using partial data from the large subunit rDNA and cytochrome oxidase subunit I. *Mol Phylogenet Evol* **34**, 81-96 (2005).
34. Sperling, E.A., Pisani, D. & Peterson, K.J. Poriferan paraphyly and its implications for Precambrian palaeobiology. *Geological Society, London, Special Publications* **286**, 355-368 (2007).
35. Nichols, S.A., Dirks, W., Pearse, J.S. & King, N. Early evolution of animal cell signaling and adhesion genes. *Proc Natl Acad Sci U S A* **103**, 12451-6 (2006).
36. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
37. Higgins, D.G. CLUSTAL V: multiple alignment of DNA and protein sequences. *Methods Mol Biol* **25**, 307-18 (1994).
38. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
39. Felsenstein, J. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

40. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
41. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-9 (2001).
42. Schmidt, H.A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-4 (2002).
43. Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**, 170-9 (1989).
44. Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci* **269**, 137-42 (2002).
45. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* **16**, 1114-1116 (1999).
46. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-7 (2001).
47. Goldman, N., Anderson, J.P. & Rodrigo, A.G. Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**, 652-70 (2000).
48. Felsenstein, J. *Inferring Phylogenies*, (Sinauer Associates, Sunderlan, MA, 2004).
49. Wägele, J.W. & Mayer, C. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evolutionary Biology* **7**, 147 (2007).
50. Bergsten, J.A. A review of long-branch attraction. *Cladistics* **21**, 163-193 (2005).
51. Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F. & Douzery, E.J. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol* **20**, 248-54 (2003).
52. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4 (2003).
53. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5 (2001).
54. Huelsenbeck, J.P., Joyce, P., Lakner, C. & Ronquist, F. Bayesian analysis of amino acid substitution models. *Philos Trans R Soc Lond B Biol Sci* **363**, 3941-3953 (2008).
55. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* **21**, 1095-109 (2004).
56. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7 Suppl 1**, S4 (2007).
57. Lemmon, A.R. & Moriarty, E.C. The importance of proper model assumption in bayesian phylogenetics. *Syst Biol* **53**, 265-77 (2004).
58. Schierwater, B. et al. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol* **7**, e20 (2009).
59. Dunn, C.W. et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-9 (2008).
60. Dellaporta, S.L. et al. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci U S A* **103**, 8751-6 (2006).

61. Signorovitch, A.Y., Buss, L.W. & Dellaporta, S.L. Comparative genomics of large mitochondria in placozoans. *PLoS Genet* **3**, e13 (2007).
62. Haen, K.M., Lang, B.F., Pomponi, S.A. & Lavrov, D.V. Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? *Mol Biol Evol* **24**, 1518-27 (2007).
63. Lavrov, D.V., Forget, L., Kelly, M. & Lang, B.F. Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution. *Mol Biol Evol* **22**, 1231-9 (2005).
64. Philippe, H. et al. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**, 706-12 (2009).
65. Lecointre, G. & Deleporte, P. Total evidence requires exclusion of phylogenetically misleading data. *Zool. Scripta* **33**(2004).
66. Burki, F. et al. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE* **2**, e790 (2007).
67. Burki, F., Shalchian-Tabrizi, K. & Pawlowski, J. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol Lett* **4**, 366-9 (2008).
68. Philippe, H. et al. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* **21**, 1740-52 (2004).
69. Yoon, H.S. et al. Broadly sampled multigene trees of eukaryotes. *BMC Evol Biol* **8**, 14 (2008).
70. Roger, A.J. & Simpson, A.G. Evolution: revisiting the root of the eukaryote tree. *Curr Biol* **19**, R165-7 (2009).
71. Peterson, K.J., Cotton, J.A., Gehling, J.G. & Pisani, D. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc Lond B Biol Sci* **363**, 1435-43 (2008).
72. Peterson, K.J. et al. Estimating metazoan divergence times with a molecular clock. *Proc Natl Acad Sci U S A* **101**, 6536-41 (2004).
73. Sanderson, M.J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-2 (2003).
74. Douzery, E.J., Snell, E.A., Bapteste, E., Delsuc, F. & Philippe, H. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A* **101**, 15386-91 (2004).
75. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41 (2004).
76. Thomas, P.D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).
77. Thomas, P.D. et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* **31**, 334-41 (2003).
78. Schafer, K.A. The cell cycle: a review. *Vet Pathol* **35**, 461-78 (1998).
79. Morgan, D.O. SnapShot: Cell-cycle regulators II. *Cell* **135**, 974-974 e1 (2008).
80. Morgan, D.O. SnapShot: cell-cycle regulators I. *Cell* **135**, 764-764 e1 (2008).
81. King, N. et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**, 783-8 (2008).
82. Gallant, P., Shiio, Y., Cheng, P.F., Parkhurst, S.M. & Eisenman, R.N. Myc and Max homologs in *Drosophila*. *Science* **274**, 1523-7 (1996).

83. De Clercq, A. & Inze, D. Cyclin-dependent kinase inhibitors in yeast, animals, and plants: a functional comparison. *Crit Rev Biochem Mol Biol* **41**, 293-313 (2006).
84. Stocker, H. & Hafen, E. Genetic control of cell size. *Curr Opin Genet Dev* **10**, 529-35 (2000).
85. Tapon, N., Moberg, K.H. & Hariharan, I.K. The coupling of cell growth to the cell cycle. *Curr Opin Cell Biol* **13**, 731-7 (2001).
86. Hariharan, I.K. & Bilder, D. Regulation of imaginal disc growth by tumor-suppressor genes in Drosophila. *Annu Rev Genet* **40**, 335-61 (2006).
87. Rawlings, J.S., Rosler, K.M. & Harrison, D.A. The JAK/STAT signaling pathway. *J Cell Sci* **117**, 1281-3 (2004).
88. Heinrich, P.C. et al. Principles of interleukin (IL)-6-type cytokine signalling and its regulation. *Biochem J* **374**, 1-20 (2003).
89. Rane, S.G. & Reddy, E.P. Janus kinases: components of multiple signaling pathways. *Oncogene* **19**, 5662-79 (2000).
90. Shuai, K. Modulation of STAT signaling by STAT-interacting proteins. *Oncogene* **19**, 2638-44 (2000).
91. Foster, F.M., Traer, C.J., Abraham, S.M. & Fry, M.J. The phosphoinositide (PI) 3-kinase family. *J Cell Sci* **116**, 3037-40 (2003).
92. Brown, S., Hu, N. & Hombria, J.C. Identification of the first invertebrate interleukin JAK/STAT receptor, the *Drosophila* gene *domeless*. *Curr Biol* **11**, 1700-5 (2001).
93. Bardin, A.J. & Amon, A. Men and sin: what's the difference? *Nat Rev Mol Cell Biol* **2**, 815-26 (2001).
94. Bedhomme, M., Jouannic, S., Champion, A., Simanis, V. & Henry, Y. Plants, MEN and SIN. *Plant Physiol Biochem* **46**, 1-10 (2008).
95. Reddy, B.V. & Irvine, K.D. The Fat and Warts signaling pathways: new insights into their regulation, mechanism and conservation. *Development* **135**, 2827-38 (2008).
96. Harvey, K. & Tapon, N. The Salvador-Warts-Hippo pathway - an emerging tumour-suppressor network. *Nat Rev Cancer* **7**, 182-91 (2007).
97. Wiens, M., Krasko, A., Muller, C.I. & Muller, W.E. Molecular evolution of apoptotic pathways: cloning of key domains from sponges (Bcl-2 homology domains and death domains) and their phylogenetic relationships. *J Mol Evol* **50**, 520-31 (2000).
98. Wiens, M. et al. Axial (apical-basal) expression of pro-apoptotic and pro-survival genes in the lake baikal demosponge *Lubomirskia baicalensis. DNA Cell Biol* **25**, 152-64 (2006).
99. Robertson, A.J. et al. The genomic underpinnings of apoptosis in *Strongylocentrotus purpuratus*. *Dev Biol* **300**, 321-34 (2006).
100. Zmasek, C.M., Zhang, Q., Ye, Y. & Godzik, A. Surprising complexity of the ancestral apoptosis network. *Genome Biol* **8**, R226 (2007).
101. Huang, S. et al. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res* **18**, 1112-26 (2008).
102. Seydoux, G. & Braun, R.E. Pathway to totipotency: lessons from germ cells. *Cell* **127**, 891-904 (2006).
103. Extavour, C.G. & Akam, M. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* **130**, 5869-84 (2003).
104. Extavour, C. Evolution of the bilaterian germ line: lineage origin and modulation of specification mechanisms. *Integr Comp Biol* **47**, 770-785 (2007).

105. Extavour, C.G., Pang, K., Matus, D.Q. & Martindale, M.Q. *vasa* and *nanos* expression patterns in a sea anemone and the evolution of bilaterian germ cell specification mechanisms. *Evol Dev* **7**, 201-15 (2005).
106. Ruppert, E.E., Fox, R.S. & Barnes, R.D. *Invertebrate Zoology: a Functional Evolutionary Approach (7th ed).* (Brooks/Cole-Thomson Learning, Belmont, CA, 2004).
107. Gaino, E., Burlando, B., Zunino, L., Pansini, M. & Buffa, P. Origin of male gametes from choanocytes in *Spongia officinalis* (Porifera, Demospongiae). *Int. J. Invert. Repr. Dev.* **7**, 83-93 (1984).
108. Mochizuki, K., Nishimiya-Fujisawa, C. & Fujisawa, T. Universal occurrence of the vasa-related genes among metazoans and their germline expression in *Hydra*. *Dev Genes Evol* **211**, 299-308 (2001).
109. Grimson, A. et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**, 1193-7 (2008).
110. Wikramanayake, A.H. et al. An ancient role for nuclear beta-catenin in the evolution of axial polarity and germ layer segregation. *Nature* **426**, 446-50 (2003).
111. Broun, M., Gee, L., Reinhardt, B. & Bode, H.R. Formation of the head organizer in *Hydra* involves the canonical Wnt pathway. *Development* **132**, 2907-16 (2005).
112. Lee, P.N., Kumburegama, S., Marlow, H.Q., Martindale, M.Q. & Wikramanayake, A.H. Asymmetric developmental potential along the animal-vegetal axis in the anthozoan cnidarian, *Nematostella vectensis*, is mediated by Dishevelled. *Dev Biol* **310**, 169-86 (2007).
113. Momose, T., Derelle, R. & Houliston, E. A maternally localised Wnt ligand required for axial patterning in the cnidarian *Clytia hemisphaerica*. *Development* **135**, 2105-13 (2008).
114. Kusserow, A. et al. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156-60 (2005).
115. Grimson, M.J. et al. Adherens junctions and beta-catenin-mediated cell signalling in a non-metazoan organism. *Nature* **408**, 727-31 (2000).
116. Coates, J.C. Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol* **13**, 463-71 (2003).
117. Massague, J. How cells read TGF-beta signals. *Nat Rev Mol Cell Biol* **1**, 169-78 (2000).
118. Matus, D.Q., Thomsen, G.H. & Martindale, M.Q. Dorso/ventral genes are asymmetrically expressed and involved in germ-layer demarcation during cnidarian gastrulation. *Curr Biol* **16**, 499-505 (2006).
119. Ma, G., Xiao, Y. & He, L. Recent progress in the study of Hedgehog signaling. *J Genet Genomics* **35**, 129-37 (2008).
120. Adamska, M. et al. The evolutionary origin of hedgehog proteins. *Curr Biol* **17**, R836-7 (2007).
121. Bray, S.J. Notch signalling: a simple pathway becomes complex. *Nat Rev Mol Cell Biol* **7**, 678-89 (2006).
122. Rentzsch, F., Fritzenwanker, J.H., Scholz, C.B. & Technau, U. FGF signalling controls formation of the apical sensory organ in the cnidarian *Nematostella vectensis.* *Development* **135**, 1761-9 (2008).
123. Larroux, C. et al. Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* **25**, 980-96 (2008).

124. Larroux, C. et al. The NK homeobox gene cluster predates the origin of Hox genes. *Curr Biol* **17**, 706-10 (2007).
125. Simionato, E. et al. Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. *BMC Evol Biol* **7**, 33 (2007).
126. Mikhailov, K.V. et al. The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* **31**, 758-68 (2009).
127. Manning, G., Young, S.L., Miller, W.T. & Zhai, Y. The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc Natl Acad Sci U S A* **105**, 9674-9 (2008).
128. Lise, M.F. & El-Husseini, A. The neuroligin and neurexin families: from structure to function at the synapse. *Cell Mol Life Sci* **63**, 1833-49 (2006).
129. Srivastava, M. et al. Evolution of the LIM homeobox gene family in basal metazoans. *BMC Biology* **8**, 4 (2010).
130. Richards, G.S. et al. Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. *Curr Biol* **18**, 1156-61 (2008).
131. Galliot, B. et al. Origins of neurogenesis, a cnidarian view. *Dev Biol* **332**, 2-24 (2009).
132. Sakarya, O. et al. A post-synaptic scaffold at the origin of the animal kingdom. *PLoS ONE* **2**, e506 (2007).
133. Tessmar-Raible, K. The evolution of neurosecretory centers in bilaterian forebrains: insights from protostomes. *Semin Cell Dev Biol* **18**, 492-501 (2007).
134. Hook, V. et al. Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. *Annu Rev Pharmacol Toxicol* **48**, 393-423 (2008).
135. Reznik, S.E. & Fricker, L.D. Carboxypeptidases from A to z: implications in embryonic development and Wnt binding. *Cell Mol Life Sci* **58**, 1790-804 (2001).
136. Mohamed, M.M. & Sloane, B.F. Cysteine cathepsins: multifunctional enzymes in cancer. *Nat Rev Cancer* **6**, 764-75 (2006).
137. Martinez, A. & Treston, A.M. Where does amidation take place? *Mol Cell Endocrinol* **123**, 113-7 (1996).
138. Han, M. et al. *Drosophila* uses two distinct neuropeptide amidating enzymes, dPAL1 and dPAL2. *J Neurochem* **90**, 129-41 (2004).
139. Huang, K.F., Liu, Y.L., Cheng, W.J., Ko, T.P. & Wang, A.H. Crystal structures of human glutaminyl cyclase, an enzyme responsible for protein N-terminal pyroglutamate formation. *Proc Natl Acad Sci U S A* **102**, 13117-22 (2005).
140. Gauthier, S.A. & Hewes, R.S. Transcriptional regulation of neuropeptide and peptide hormone expression by the *Drosophila dimmed* and *cryptocephal* genes. *J Exp Biol* **209**, 1803-15 (2006).
141. Renden, R. et al. Drosophila CAPS is an essential gene that regulates dense-core vesicle release and synaptic vesicle fusion. *Neuron* **31**, 421-37 (2001).
142. Walchli, S., Colinge, J. & Hooft van Huijsduijnen, R. MetaBlasts: tracing protein tyrosine phosphatase gene family roots from Man to *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. *Gene* **253**, 137-43 (2000).
143. Schioth, H.B. & Fredriksson, R. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen Comp Endocrinol* **142**, 94-101 (2005).
144. Muller, W.E., Blumbach, B. & Muller, I.M. Evolution of the innate and adaptive immune systems: relationships between potential immune molecules in the lowest metazoan phylum (Porifera) and those in vertebrates. *Transplantation* **68**, 1215-27 (1999).

145. Muller, W.E. & Muller, I.M. Origin of the Metazoan Immune System: Identification of the Molecules and Their Functions in Sponges. *Integr Comp Biol* **43**, 281-292 (2003).
146. Inohara, Chamaillard, McDonald, C. & Nunez, G. NOD-LRR proteins: role in host-microbial interactions and inflammatory disease. *Annu Rev Biochem* **74**, 355-83 (2005).
147. Sarrias, M.R. et al. The Scavenger Receptor Cysteine-Rich (SRCR) domain: an ancient and highly conserved protein module of the innate immune system. *Crit Rev Immunol* **24**, 1-37 (2004).
148. Wheeler, G.L., Miranda-Saavedra, D. & Barton, G.J. Genome analysis of the unicellular green alga *Chlamydomonas reinhardtii* Indicates an ancient evolutionary origin for key pattern recognition and cell-signaling protein families. *Genetics* **179**, 193-7 (2008).
149. Pahler, S., Blumbach, B., Muller, I. & Muller, W.E. Putative multiadhesive protein from the marine sponge Geodia cydonium: cloning of the cDNA encoding a fibronectin-, an SRCR-, and a complement control protein module. *J Exp Zool* **282**, 332-43 (1998).
150. Wiens, M. et al. Toll-like receptors are part of the innate immune defense system of sponges (demospongiae: Porifera). *Mol Biol Evol* **24**, 792-804 (2007).
151. Nedelcu, A.M. Comparative genomics of phylogenetically diverse unicellular eukaryotes provide new insights into the genetic basis for the evolution of the programmed cell death machinery. *J Mol Evol* **68**, 256-68 (2009).
152. Tzu, J. & Marinkovich, M.P. Bridging structure with function: structural, regulatory, and developmental role of laminins. *Int J Biochem Cell Biol* **40**, 199-214 (2008).
153. Cheng, Y.S., Champliaud, M.F., Burgeson, R.E., Marinkovich, M.P. & Yurchenco, P.D. Self-assembly of laminin isoforms. *J Biol Chem* **272**, 31525-32 (1997).
154. Team, R.D.C. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2009).
155. Kyrpides, N.C. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics* **15**, 773-4 (1999).
156. Inze, D. & De Veylder, L. Cell cycle regulation in plant development. *Annu Rev Genet* **40**, 77-105 (2006).
157. Claudio, P.P., Tonini, T. & Giordano, A. The retinoblastoma family: twins or distant cousins? *Genome Biol* **3**, reviews3012 (2002).
158. Walworth, N., Davey, S. & Beach, D. Fission yeast chk1 protein kinase links the rad checkpoint pathway to cdc2. *Nature* **363**, 368-71 (1993).
159. Shieh, S.Y., Ahn, J., Tamai, K., Taya, Y. & Prives, C. The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites. *Genes Dev* **14**, 289-300 (2000).
160. Fernandez-Guerra, A. et al. The genomic repertoire for cell cycle control and DNA metabolism in S. purpuratus. *Dev Biol* **300**, 238-51 (2006).
161. Bogliolo, M. et al. The *Fanconi anaemia* genome stability and tumour suppressor network. *Mutagenesis* **17**, 529-38 (2002).
162. Bork, P. et al. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J* **11**, 68-76 (1997).
163. Nichols, C.B., Fraser, J.A. & Heitman, J. PAK kinases Ste20 and Pak1 govern cell polarity at different stages of mating in *Cryptococcus neoformans*. *Mol Biol Cell* **15**, 4476-89 (2004).
164. Manning, G., Plowman, G.D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27**, 514-20 (2002).

165. Elbing, K., McCartney, R.R. & Schmidt, M.C. Purification and characterization of the three Snf1-activating kinases of *Saccharomyces cerevisiae*. *Biochem J* **393**, 797-805 (2006).
166. Scheid, M.P. & Woodgett, J.R. PKB/AKT: functional insights from genetic models. *Nat Rev Mol Cell Biol* **2**, 760-8 (2001).
167. Goldberg, J.M. et al. The dictyostelium kinome--analysis of the protein kinases from a simple model organism. *PLoS Genet* **2**, e38 (2006).
168. Templeton, D.J. Protein kinases: getting NEKed for S6K activation. *Curr Biol* **11**, R596-9 (2001).
169. Dinkova, T.D. et al. Dissecting the TOR-S6K signal transduction pathway in maize seedlings : relevance on cell growth regulation. *Physiologia Plantarum* (2007).
170. Dann, S.G. & Thomas, G. The amino acid sensitive TOR pathway from yeast to mammals. *FEBS Lett* **580**, 2821-9 (2006).
171. Deprost, D. et al. The *Arabidopsis* TOR kinase links plant growth, yield, stress resistance and mRNA translation. *EMBO Rep* **8**, 864-70 (2007).
172. Engelman, J.A., Luo, J. & Cantley, L.C. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat Rev Genet* **7**, 606-19 (2006).
173. Matsumoto, S., Bandyopadhyay, A., Kwiatkowski, D.J., Maitra, U. & Matsumoto, T. Role of the Tsc1-Tsc2 complex in signaling and transport across the cell membrane in the fission yeast Schizosaccharomyces pombe. *Genetics* **161**, 1053-63 (2002).
174. Wohrle, F.U., Daly, R.J. & Brummer, T. How to Grb2 a Gab. *Structure* **17**, 779-81 (2009).
175. Aspuria, P.J. & Tamanoi, F. The Rheb family of GTP-binding proteins. *Cell Signal* **16**, 1105-12 (2004).
176. Heymont, J. et al. TEP1, the yeast homolog of the human tumor suppressor gene PTEN/MMAC1/TEP1, is linked to the phosphatidylinositol pathway and plays a role in the developmental process of sporulation. *Proc Natl Acad Sci U S A* **97**, 12672-7 (2000).
177. Vojtek, A.B. & Der, C.J. Increasing complexity of the Ras signaling pathway. *J Biol Chem* **273**, 19925-8 (1998).
178. Dehal, P. et al. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157-67 (2002).
179. Putnam, N.H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-71 (2008).
180. Sodergren, E. et al. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941-52 (2006).
181. Pincus, D., Letunic, I., Bork, P. & Lim, W.A. Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc Natl Acad Sci U S A* **105**, 9680-4 (2008).
182. Tzolovsky, G., Millo, H., Pathirana, S., Wood, T. & Bownes, M. Identification and phylogenetic analysis of *Drosophila melanogaster* myosins. *Mol Biol Evol* **19**, 1041-52 (2002).
183. Abedin, M. & King, N. The premetazoan ancestry of cadherins. *Science* **319**, 946-8 (2008).
184. Ekert, P.G., Silke, J. & Vaux, D.L. Caspase inhibitors. *Cell Death Differ* **6**, 1081-6 (1999).

185. Filee, J., Pouget, N. & Chandler, M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol Biol* **8**, 320 (2008).
186. Miller, D.J. et al. The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. *Genome Biol* **8**, R59 (2007).
187. Bradham, C.A. et al. The sea urchin kinome: a first look. *Dev Biol* **300**, 180-93 (2006).
188. Meyer-Ficca, M.L., Meyer, R.G., Jacobson, E.L. & Jacobson, M.K. Poly(ADP-ribose) polymerases: managing genome stability. *Int J Biochem Cell Biol* **37**, 920-6 (2005).
189. Riemer, D., Wang, J., Zimek, A., Swalla, B.J. & Weber, K. Tunicates have unusual nuclear lamins with a large deletion in the carboxyterminal tail domain. *Gene* **255**, 317-25 (2000).
190. Prabhu, Y. & Eichinger, L. The Dictyostelium repertoire of seven transmembrane domain receptors. *Eur J Cell Biol* **85**, 937-46 (2006).
191. Pandey, R. et al. Analysis of histone acetyltransferase and histone deacetylase families of Arabidopsis thaliana suggests functional diversification of chromatin modification among multicellular eukaryotes. *Nucleic Acids Res* **30**, 5036-55 (2002).
192. Ledent, V. & Vervoort, M. The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. *Genome Res* **11**, 754-70 (2001).
193. Ochi, H., Pearson, B.J., Chuang, P.T., Hammerschmidt, M. & Westerfield, M. Hhip regulates zebrafish muscle development by both sequestering Hedgehog and modulating localization of Smoothened. *Dev Biol* **297**, 127-40 (2006).
194. Seack, J., Kruse, M. & Muller, W.E. Evolutionary analysis of G-proteins in early metazoans: cloning of alpha- and beta-subunits from the sponge *Geodia cydonium*. *Biochim Biophys Acta* **1401**, 93-103 (1998).
195. Zhang, N., Long, Y. & Devreotes, P.N. Ggamma in *Dictyostelium*: its role in localization of gbetagamma to the membrane is required for chemotaxis in shallow gradients. *Mol Biol Cell* **12**, 3204-13 (2001).
196. Mason, M.G. & Botella, J.R. Completing the heterotrimer: isolation and characterization of an Arabidopsis thaliana G protein gamma-subunit cDNA. *Proc Natl Acad Sci U S A* **97**, 14784-8 (2000).
197. van Dam, T.J., Rehmann, H., Bos, J.L. & Snel, B. Phylogeny of the CDC25 homology domain reveals rapid differentiation of Ras pathways between early animals and fungi. *Cell Signal* **21**, 1579-85 (2009).
198. Caffrey, D.R., O'Neill, L.A. & Shields, D.C. The evolution of the MAP kinase pathways: coduplication of interacting proteins leads to new signaling cascades. *J Mol Evol* **49**, 567-82 (1999).
199. Kloepper, T.H., Kienle, C.N. & Fasshauer, D. SNAREing the basis of multicellularity: consequences of protein family expansion during evolution. *Mol Biol Evol* **25**, 2055-68 (2008).
200. Behr, M., Riedel, D. & Schuh, R. The claudin-like megatrachea is essential in septate junctions for the epithelial barrier function in *Drosophila*. *Dev Cell* **5**, 611-20 (2003).
201. Wu, V.M., Schulte, J., Hirschi, A., Tepass, U. & Beitel, G.J. Sinuous is a *Drosophila* claudin required for septate junction organization and epithelial tube size control. *J Cell Biol* **164**, 313-23 (2004).
202. Boute, N. et al. Type IV collagen in sponges, the missing link in basement membrane ubiquity. *Biol Cell* **88**, 37-44 (1996).

203. Magie, C.R., Pang, K. & Martindale, M.Q. Genomic inventory and expression of Sox and Fox genes in the cnidarian *Nematostella vectensis*. *Dev Genes Evol* **215**, 618-30 (2005).
204. Matus, D.Q., Pang, K., Daly, M. & Martindale, M.Q. Expression of Pax gene family members in the anthozoan cnidarian, *Nematostella vectensis*. *Evol Dev* **9**, 25-38 (2007).
205. Hadrys, T., DeSalle, R., Sagasser, S., Fischer, N. & Schierwater, B. The *Trichoplax PaxB* gene: a putative Proto-PaxA/B/C gene predating the origin of nerve and sensory cells. *Mol Biol Evol* **22**, 1569-78 (2005).
206. Ryan, J.F. et al. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biol* **7**, R64 (2006).
207. Matus, D.Q. et al. Molecular evidence for deep evolutionary roots of bilaterality in animal development. *Proc Natl Acad Sci U S A* **103**, 11195-200 (2006).
208. Pang, K., Matus, D.Q. & Martindale, M.Q. The ancestral role of COE genes may have been in chemoreception: evidence from the development of the sea anemone, *Nematostella vectensis* (Phylum Cnidaria; Class Anthozoa). *Dev Genes Evol* **214**, 134-8 (2004).
209. Yamada, A., Pang, K., Martindale, M.Q. & Tochinai, S. Surprisingly complex *T-box* gene complement in diploblastic metazoans. *Evol Dev* **9**, 220-30 (2007).
210. Martinelli, C. & Spring, J. Distinct expression patterns of the two *T-box* homologues *Brachyury* and *Tbx2*/3 in the placozoan *Trichoplax adhaerens*. *Dev Genes Evol* **213**, 492-9 (2003).
211. Chourrout, D. et al. Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* **442**, 684-7 (2006).
212. Monteiro, A.S., Schierwater, B., Dellaporta, S.L. & Holland, P.W. A low diversity of ANTP class homeobox genes in Placozoa. *Evol Dev* **8**, 174-82 (2006).
213. Hibino, T. et al. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol* **300**, 349-65 (2006).
214. Bosch, T.C. et al. Uncovering the evolutionary history of innate immunity: the simple metazoan *Hydra* uses epithelial cells for host defence. *Dev Comp Immunol* **33**, 559-69 (2009).
215. Zapata, J.M., Martinez-Garcia, V. & Lefebvre, S. Phylogeny of the TRAF/MATH domain. *Adv Exp Med Biol* **597**, 1-24 (2007).
216. Sullivan, J.C., Reitzel, A.M. & Finnerty, J.R. A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan *Nematostella vectensis*. *Genome Inform* **17**, 219-29 (2006).
217. Gauthier, M. & Degnan, B.M. The transcription factor NF-kappaB in the demosponge *Amphimedon queenslandica*: insights on the evolutionary origin of the Rel homology domain. *Dev Genes Evol* **218**, 23-32 (2008).
218. Kvennefors, E.C., Leggat, W., Hoegh-Guldberg, O., Degnan, B.M. & Barnes, A.C. An ancient and variable mannose-binding lectin from the coral *Acropora millepora* binds both pathogens and symbionts. *Dev Comp Immunol* **32**, 1582-92 (2008).
219. Low, D.H. et al. A novel human tectonin protein with multivalent beta-propeller folds interacts with ficolin and binds bacterial LPS. *PLoS ONE* **4**, e6260 (2009).
220. Huh, C.G. et al. Cloning and characterization of *Physarum polycephalum* tectonins. Homologues of *Limulus* lectin L-6. *J Biol Chem* **273**, 6565-74 (1998).

221. Wood-Charlson, E.M. & Weis, V.M. The diversity of C-type lectins in the genome of a basal metazoan, *Nematostella vectensis*. *Dev Comp Immunol* **33**, 881-9 (2009).
222. Perovic-Ottstadt, S. et al. A (1-->3)-beta-D-glucan recognition protein from the sponge *Suberites domuncula*. Mediated activation of fibrinogen-like protein and epidermal growth factor gene expression. *Eur J Biochem* **271**, 1924-37 (2004).
223. Kjaer, K.H. et al. Evolution of the 2'-5'-Oligoadenylate Synthetase Family in Eukaryotes and Bacteria. *J Mol Evol* (2009).
224. Reintamm, T. et al. Sponge OAS has a distinct genomic structure within the 2-5A synthetase family. *Mol Genet Genomics* **280**, 453-66 (2008).
225. de Jong, D. et al. Multiple dicer genes in the early-diverging metazoa. *Mol Biol Evol* **26**, 1333-40 (2009).
226. Nehyba, J., Hrdlickova, R. & Bose, H.R., Jr. Dynamic evolution of immune system regulators: The history of the interferon regulatory factor (IRF) family. *Mol Biol Evol* (2009).
227. Fernandez-Busquets, X. & Burger, M.M. Circular proteoglycans from sponges: first members of the spongican family. *Cell Mol Life Sci* **60**, 88-112 (2003).