"The Dynamic Genome of *Hydra*"

Jarrod A. Chapman[1,*], Ewen F. Kirkness[2,*], Oleg Simakov[3,4,*], Steven E. Hampson[5,#], Therese Mitros[4], Thomas Weinmaier[6], Thomas Rattei[6], Prakash G. Balasubramanian[3], Jon Borman[2], Dana Busam[2], Kathryn Disbennett[2], Cynthia Pfannkoch[2], Nadezhda Sumin[2], Granger G. Sutton[2], Lakshmi Devi Viswanathan[2], Brian Walenz[2], David M. Goodstein[1], Uffe Hellsten[1], Takeshi Kawashima[4], Simon E. Prochnik[1], Nicholas H. Putnam[1,4,+], Shengquiang Shu[1], Bruce Blumberg[7,8], Catherine E. Dana[8,9], Lydia Gee[7,8], Dennis F. Kibler[5], Lee Law[7,8], Dirk Lindgens[7,8], Daniel E. Martinez[10], Jisong Peng[7,8], Philip A. Wigge[11,@], Bianca Bertulat[3], Corina Guder[3], Yukio Nakamura[3], Suat Ozbek[3], Hiroshi Watanabe[3], Konstantin Khalturin[12], Georg Hemmrich[12], André Franke[12], René Augustin[12], Sebastian Fraune[12], Eisuke Hayakawa[13], Shiho Hayakawa[13], Mamiko Hirose[13], Jung Shan Hwang[13], Kazuho Ikeo[13], Chiemi Nishimiya-Fujisawa[13], Atshushi Ogura[13,=], Toshio Takahashi[14], Patrick R.H. Steinmetz[15], Xiaoming Zhang,[16] Roland Aufschnaiter[17], Marie-Kristin Eder[17], Anne-Kathrin Gorny[17,§], Willi Salvenmoser[17], Alysha M. Heimberg[18], Benjamin M. Wheeler[19], Kevin J. Peterson[18], Angelika Böttger[20], Patrick Tischler[6], Alexander Wolf[20], Takashi Gojobori[13], Karin A. Remington[2,$], Robert L. Strausberg[2], J. Craig Venter[2], Ulrich Technau[15], Bert Hobmayer[17], Thomas C.G. Bosch[12], Thomas W. Holstein[3], Toshitaka Fujisawa[13], Hans R. Bode[7,8], Charles N. David[20], Daniel S. Rokhsar[1,4], Robert E. Steele[8,9]

[1]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598

[2]The J. Craig Venter Institute, Rockville, Maryland 20850

[3]Institute of Zoology, Department of Molecular Evolution and Genomics, University of Heidelberg, D-69120 Heidelberg, Germany

[4]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720

[5]Department of Computer Science, University of California, Irvine, CA 92697-3435

[6]Department of Genome-Oriented Bioinformatics, Technische Universität München, D-85354 Freising, Germany

[7]Department of Developmental and Cell Biology, University of California, Irvine, CA 92697-2275

[8]Developmental Biology Center, University of California, Irvine, CA 92697-2275

[9]Department of Biological Chemistry, University of California, Irvine, CA 92697-1700

[10]Department of Biology, Pomona College, Claremont, CA 91711

[11]The Salk Institute, La Jolla, CA 92037

[12]Zoologisches Institüt, Christian-Albrechts-University, D-24098 Kiel, Germany

[13]National Institute of Genetics, Yata 1,111, Mishima 411-8540, Japan

[14]Suntory Institute for Bioorganic Research, Osaka 618-8503, Japan

[15]Department of Molecular Evolution and Development, University of Vienna, A-1090 Vienna, Austria

[16]Department of Anatomy and Cell Biology, The University of Kansas Medical Center, Kansas City, Kansas 66160

[17]Institute of Zoology and Center for Molecular Biosciences, University of Innsbruck, A-6020 Innsbruck, Austria

[18]Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755

[19]Department of Computer Science, North Carolina State University, Raleigh, North Carolina 27695

[20]Department of Biology II, Ludwig-Maximilians-University, D-82152 Planegg-Martinsried, Germany

[*]These authors contributed equally to this work.

[#]Deceased, July 5, 2007

[@]current address: Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, United Kingdom

[§]current address: Institute of Human Genetics, University of Heidelberg, D-69120 Heidelberg, Germany

[$]current address: Center for Bioinformatics and Computational Biology, National Institute of General Medical Sciences, Bethesda, MD 20892-6200

[+]current address: Department of Ecology and Evolutionary Biology, Rice University, Houston, TX 77251-1892

[=]current address: Ochadai Academic Production, Ochanomizu University, Ohtsuka, Bunkyo, 1128610, Tokyo, Japan

Table of Contents

S1. Choice of *Hydra* species for genome sequencing

The genus *Hydra* consists of multiple clades[1,2,3], with the main distinction being between the brown hydras and the green hydras. The latter contain symbiotic algae. The brown hydra species *Hydra magnipapillata* (strain 105), a member of the *vulgaris* clade, was chosen for genome sequencing for several reasons. First, a large-scale EST project had been carried out using this strain before the genome project was initiated. This strain has been used extensively for studies of developmental patterning, regeneration, cell differentiation, and stem cell biology. This strain was also used for the *Hydra* Peptide Project, a project that has identified a number of novel bioactive peptides from *Hydra*[4,5]. Mutants have been isolated from *Hydra magnipapillata* by inbreeding of animals collected from the wild[6]. Such mutants include one that fails to regenerate a head, one with temperature-sensitive interstitial cells, ones with altered sizes, and ones with defects in nematocyte production. Genome sizes have been determined for several species of *Hydra* using densitometry on Feulgen-stained nuclei[2,7]. All of the brown *Hydra* species have genomes of very similar sizes[7] with the 105 strain of *H. magnipapillata* having a haploid genome size of 1.3 x 10^9 base pairs[2]. All *Hydra* species, both brown and green, have a diploid chromosome number of 30[7]. While the green hydra genome is approximately four times smaller than the genomes of the brown hydras and was thus potentially less expensive to sequence, green hydras are not nearly as widely used for experimental studies as brown hydras. Finally, a method for making transgenic *Hydra*[8] has been developed using a brown hydra strain of the *vulgaris* clade (*Hydra vulgaris* strain AEP) that produces embryos readily in the laboratory (*Hydra magnipapillata* 105 does not readily produce embryos in the laboratory; Fujisawa et al., unpublished observations). Taking into consideration all of these features, the 105 strain of *Hydra magnipapillata* was selected as the best *Hydra* model for genome sequencing.

S2. *Hydra* strains and culture

The species *Hydra magnipapillata* was originally described in 1947 by Ito[9]. The 105 strain of *H. magnipapillata* originated from a single polyp collected by Dr. Tsutomu Sugiyama in September, 1973 from a swamp adjacent to the National Institute of Genetics in Mishima, Japan (T. Fujisawa, unpublished information). This strain of *H. magnipapillata* has been propagated in labs in Japan, Europe, and North America since then. For isolation of DNA for genome sequencing, the 105 strain was re-cloned from a single polyp in the laboratory of Dr. Hans Bode at UC Irvine in 2004 under standard culture conditions. About 1000 polyps were harvested from the resulting culture, frozen, and shipped to the J. Craig Venter Institute in 2004. All of the DNA used for genome sequencing was obtained from this sample. For construction of some of the cDNA libraries for EST sequencing, the sf-1 and nem-1 strains of *Hydra magnipapillata* were used. The sf-1 strain was produced by inbreeding in the lab (F5 generation) of wild strains originally collected from Hachirogata, in northern Japan. This strain has a temperature-sensitive mutation that results in the loss of the interstitial cell lineage when the animals are cultured at 23°C or higher. The nem-1 strain of *H. magnipapillata* undergoes sexual differentiation readily in the lab and thus was used to make cDNA libraries from sexual male and female polyps.

Hydra was cultured using standard methods that include feeding with *Artemia* nauplii[10]. In 2005,

voucher specimens of the 105 strain of *H. magnipapillata* from UC Irvine were fixed in 10% neutral buffered formalin, transferred to 70% alcohol, and deposited in the Invertebrate Zoology Collection of the Peabody Museum of Natural History, Yale University, under catalog numbers YPM35906, YPM 35907, YPM 35908, and YPM 35909. An aliquot of the DNA sample used for constructing the shotgun libraries for genome sequencing has been transferred from the J. Craig Venter Institute to the Peabody Museum, where it is cryopreserved. Requests for DNA samples or access to the voucher specimens should be directed to the Senior Collections Manager at the Peabody Museum.

S3. Genome sequencing and assembly

Sequencing

The genome of the 105 strain of *Hydra magnipapillata* was sequenced using the whole genome shotgun method. Total DNA was extracted from polyps and was used to construct libraries in the plasmid pHOS2, or the fosmid pCCFOS1. End-sequencing of clones from each library was conducted using a standard capillary platform (ABI 3730), and yielded 10,214,521 good traces (96% paired) with a mean clear read length of 838 bases. All sequences were generated at the J. Craig Venter Institute, and have been deposited in the NCBI Trace Archive. Sequence traces were derived from inserts of 3-4 kb (48%), 8-10 kb (21%), 10-12 kb (31%), and 35-40 kb (0.1%).

The six fosmid libraries that were constructed (35-40 kb inserts) yielded very low titers, sequenced with <50% success, and yielded 10,200 sequences with an A+T content (66%) that deviates from the genome average of 71%. Attempts to make a bacterial artificial chromosome (BAC) library from *H. magnipapillata* for end-sequencing were unsuccessful (K. Osoegawa, R. Steele, K. Hall, and P. de Jong, unpublished observations), which we attribute to the known instability in *E. coli* of large insert clones from A+T-rich genomes[11]. This inability to produce large-insert (i.e. fosmid or BAC) clones from the A+T-rich *Hydra* genome limits the length of sequence scaffolds.

To test for sequence completeness, 454 sequencing was carried out at JCVI on the same sample of *Hydra magnipapillata* genomic DNA that was used for Sanger sequencing. A total of 7.6 million reads (1.6 Gb) were generated. More than 92% of the 454 reads (average length 277 bases) could be aligned for >95% of their lengths with >95% nucleotide identity to the CA assembly (see below). The no-hit 454 reads (3% of total) were generally of low quality and short read length (average 184 bases). Thus, 454 sequencing revealed little sequence that could not also be found in the Sanger sequencing reads.

Assembly

Although *Hydra magnipapillata* has been maintained in the lab for decades as a clonal, asexually reproducing, diploid population, it is not "inbred" in any sense. The genomic DNA used for sequencing therefore contains two haplotypes per locus. The assembly of polymorphic diploid genomes is challenging, and no general method for such assembly exists. For genomes with low polymorphism levels (*e.g.*, human, ~0.1% SNP rate) both alleles can be integrated into a single, albeit mosaic, reference consensus, in which each chromosomal locus is represented by a single

assembled sequence. Conversely, at the high level of polymorphism (>3-5%) seen in some broadcast spawning invertebrates (e.g., *Ciona savignyi*[12] and amphioxus[13]) it is difficult to combine haplotypes into a single consensus, and may be more practical to assemble the haplotypes separately (at a two-fold increase in sequencing cost). In this extreme, a single chromosomal locus is represented by two assembled sequences.

Genomes with intermediate levels of polymorphism, as found in *Hydra* (see below), are the most challenging to assemble, as assembly programs may collapse some loci but assemble haplotypes apart at others, resulting in a partially artifactually duplicated sequence. For *Hydra*, we performed two independent assemblies (described below). For the purpose of the detailed annotation and analysis described in the manuscript, we focused our attention on the RP assembly which had less apparent haplotypic redundancy, as described below. Both assemblies have been deposited in GenBank and are available for use by the research community, recognizing the subtle difference in haplotype separation between them.

Celera Assembler assembly (CA)

The "CA" assembly was produced at the J. Craig Venter Institute using a modified version of the Celera Assembler[14,15,16] Release 3.10 (http://wgs-assembler.sourceforge.net, cvs tag HYDRA7, July 2006). The Overlap Based Trimming module was used to determine the clear region of each read. Overlaps were computed at up to 6% error using 22-mer seeds, ignoring 22-mers present more than 40 times in the trimmed fragments. The error rate of each overlap was corrected after inferring sequencing errors in a single read from the mutlialignment of all overlapping reads. Unitigs[15] were built using only overlaps with a maximum of 1.5% corrected error, the default value of the assembler. The -g parameter to the unitig module was set to 700,000,000, which has the effect of biasing the a-stat calculation towards labeling unitigs as unique. No special settings were used to build contigs and scaffolds from unitigs and mate pairs. 8,397,288 reads (82.2%) were placed in the assembly, 696,247 reads (6.8%) were in 'degenerate' contigs, 1,034,629 reads (10.1%) were unassembled, the remaining reads were discarded during trimming. This assembly has been deposited with NCBI under Accession Number ABRM00000000.

Ringer-Phrap assembly (RP)

An independent assembly ("RP") was produced at UC Berkeley using a new assembly protocol described here. *Hydra* genome reads were obtained from the NCBI Trace Archive and collected into "library groups" by reported insert size. Observed insert sizes self-consistently determined from the assembly are shown in Table S4. Reads were trimmed based on clipping coordinates provided by JCVI. Only reads of at least 400 bp with a mate-pair of at least 400 bp were used in this assembly. Table S4 describes the library groups after trimming and clipping. Read-to-read pairwise alignments were calculated using the MALIGN aligner module[17]. MALIGN used the co-occurrence of at least 17 distinct 16-mers between pairs of reads to trigger banded, semi-global Needleman-Wunsch alignments. 16-mers occurring more than 50 times in the data set were not allowed to trigger alignments.

Alignments of at least 100 bp 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 (Fig. S1). The 1-ring neighborhood of a read includes all reads with valid alignments to it. The n-ring neighborhood of a read is the union of 1-ring neighborhoods of all reads that belong to its n-1-ring neighborhood. Single-

linkage clustering was performed using read-read alignments of at least 100 bp and 98% identity (using the make_clusters program). Alignments were rejected if both reads involved had a 2-ring neighborhood of more than 45 reads. This step effectively removes reads derived from repetitive regions of the genome. 166,514 clusters of at least 4 reads were generated containing a total of 7,781,128 reads (*i.e.,* 83.5% of reads are clustered).

Each read-cluster was assembled with phrap (version 0.960731) (http://www.phrap.org/) using parameters: -minmatch 35 -minscore 55. Quality scores were not used, which allows allelic variants to be assembled together into the same contig (data not shown). The resulting contigs were ordered and oriented into scaffolds using the perl script phrapOut2Scaffolds in which contigs are iteratively merged via a greedy ordering based on the number and consistency of read-pair linkages between them. The RP assembly has been deposited with NCBI under Accession Number ACZU00000000.

Comparison of the CA and RP *Hydra* genome assemblies: general contiguity and scaffolding

Tables S5, S6, and S7 provide summary statistics of scaffold sizes, contig sizes, and gap sizes for the two assemblies (scaffolds corresponding to microbial contaminants have been removed from both assemblies). The CA assembly is longer in total scaffold and contig length, with essentially all of the differences coming from scaffolds of length less than 100 kb. We argue below that this difference is largely accounted for by the combination of redundant haplotype assemblies and short assembled repetitive regions found in the CA assembly.

Conventionally, a measure of assembly quality is the "N50" length, that is, the length of contig or scaffold such that half the assembled sequence is in pieces longer than this threshold. The CA assembly has contig and scaffold N50 values of 12.8 kb and 63.4 kb, respectively. The RP assembly has a somewhat shorter contig N50 length (9.7 kb) but longer scaffold N50 length (92.5 kb). These numbers are not directly comparable, however, since the CA assembly is longer in total length. Considering instead a common reference point near half the genome size, we find that the CA assembly captures 500 Mb in 3,526 scaffolds longer than 84.4 kb, while the RP assembly captures 500 Mb in 3,385 scaffolds longer than 77.0 kb. Both assemblies capture ~420 Mb in scaffolds longer than 100 kb.

Intra-scaffold gaps are comparable in the two assemblies comprising 70.5 Mbp (5.6%) of the CA assembly and 67.2 Mbp (7.6%) of the total RP scaffold sequence, with similar median and mean gap sizes (255 bp and 682 bp for CA, 275 bp and 630 bp for RP). As noted above, a comparable fraction of input reads are left unassembled in the two assemblies (17.8% in CA, 16.5% in RP). Thus, approximately 10% of the reads are expected to correspond to un-captured (inter-scaffold) gaps in both assemblies. The extrapolated sequenced genome size can then be estimated from the total contig length and unplaced read fraction to be 1.5 Gb and 1.0 Gb for the CA and RP assemblies, respectively. This extrapolation assumes uniform depth of sequence coverage for the unplaced read cohort, and does not account for the splitting of haplotypes into separate regions of the assembly. We show below that there is significant separate assembly of haplotypes in the CA assembly, which partially accounts for its longer length.

Comparison of the CA and RP Assemblies: completeness and redundancy relative to ESTs

We aligned *Hydra* EST assemblies with longer than 400 bp to each assembly using BLAT[18]. Sequences that hit neither assembly were manually reviewed, identifying and removing obvious non-*Hydra* contaminants. The two assemblies were found to be comparable in terms of completeness with respect to coding sequences; more than 99% of *H. magnipapillata*-derived ESTs are represented in both assemblies.

The assemblies differed in their level of apparent redundancy, as measured by the fraction of EST-derived ORFs that had multiple high confidence hits to the genome. While 15-20% of the EST-derived ORFs hit the CA assembly twice, only 3-5% hit the RP assembly twice (the ranges reflect different cutoffs on the length of the EST alignment to the assembly). These redundancies represent either recent paralogs (with accumulated sequence divergence) or allelic variants (with segregating variation). The ESTs were derived from multiple *H. magnipapillata* populations and, in some cases, closely related species.

Comparison of the CA and RP Assemblies: sequence redundancy

Due to the diploid nature of the *Hydra* genome, very similar pairs of sequences within an assembly could represent (1) assembly of the two haplotypes into distinct sequences, and/or (2) very recently diverged paralogous genes and/or families of repetitive elements. We evaluated the RP and CA assemblies for sequence redundancy in two ways.

First we counted the number of occurrences of each 21-mer in the two assemblies (Table S8). Assuming one discrepancy per ~100 bp (~1% polymorphism), both haplotypes should share ~80% of their 21-mers, since each isolated SNP creates 21 haplotype-distinct 21-mers that overlap the SNP. So 21-mers that occur two times in an assembly may reflect allelic regions. Not surprisingly, the longer CA assembly contains more distinct 21-mers. However, the RP assembly contains more *unique* 21-mers (*i.e.*, 21-mers that occur only a single time in the assembly). So the bulk of additional sequence found in the CA assembly represents either repeated sequence or redundantly assembled haplotypes.

As a second measure of redundancy on a longer length scale than provided by 21-mers, assemblies were first soft-masked for 21-mers that occurred five times or more in the respective sequences to remove high copy repeats, and then aligned to themselves using BLASTN -e 1e-199 -U -F 'm D' -W 21. For these parameters, only HSPs longer than ~350 bp are detected. As found in the EST analysis, the CA assembly has more two-copy regions (~30% in CA vs. ~10% in RP). We note that after correcting the 21-mer percentages by dividing by 80%, the 21-mer repetitiveness and self-alignment repetitiveness measures consistently estimate 25-30% redundancy in the CA assembly and 10% in the RP assembly.

Review of inter-assembly alignments

To understand better the relationship between the assemblies, we compared them using the same BLAST protocol described for self-comparison. A characteristic pattern of two-to-one scaffold alignment coverage is demonstrated in Fig. S2, which shows RP scaffolds at the top and the corresponding CA scaffolds aligned below, color-coded by percent identity. We note that the CA

scaffolds align over their entire length to the RP scaffold indicating no obvious global inconsistency between the RP and CA assemblies. Approximately 22% (see Fig. S2) of the RP assembly is doubly covered by CA scaffolds, which is consistent with the redundancy summarized in Table S8. Furthermore, regions of the CA assembly that mapped redundantly onto the RP assembly had lower depth of coverage, and little or no polymorphism, as expected if the CA assembly were assembling allelic variants separately in these regions. In principle, this pattern of alignment could also be consistent with very recent, large-scale duplication within the *Hydra* genome. The distribution of polymorphic sites in the RP assembly excludes this scenario, as we show next.

Distribution of SNPs is consistent with a simple coalescent model

Large-scale segmental duplication at some defined time in the past would yield a Poisson distribution of substitutions, with a rate defined by the time of divergence for the two copies. In contrast, for segregating variation within a population, the coalescence times for alleles in different regions of the genome are different, and the distribution is not Poisson. The details depend on populations dynamics, but in the simplest model (assuming constant populations size) the result is a geometric distribution for the number of polymorphic sites within a given window size (see for example, Nordborg[19]). For practical applications of this model to polymorphic genomes see Vinson *et al.*[12] (*C. savignyi*) and Putnam *et al.*[13] (amphioxus). In agreement with the simple population genetics model, we observe a geometric distribution that is quite different from the Poisson expectation for large-scale duplication (Fig. S3).

To evaluate the haplotypic polymorphism rate and distribution, all reads were re-aligned to the RP assembly with blastn (using parameters: -W 24 -U -F 'm D' -b 10 -v 10 -K 10 -e 1e-100). A read was placed on the assembly if its highest scoring alignment exceeded the score of any alternate alignment by at least 5%. We used re-alignment rather than the placement of the reads by the assembler to avoid artifacts of assembly in the face of polymorphism. The average depth of coverage peaks broadly at ~6.5X but the distribution is broader and more skewed relative to expectation for a uniform sampling of a random sequence.

SNPs were detected by aligning reads to regions covered by 8-10 reads, and identifying positions where two or more reads were discrepant with consensus but agreed with each other on the variant base call. The analysis was limited to regions of 8-10-fold coverage to minimize sampling effects. We then measured the number of SNPs in a 500 bp window, and computed a distribution across all such windows. As shown in Fig. S3, the Poisson description expected for synchronous duplication is clearly rejected in favor of the geometric distribution expected in a panmictic population. We therefore conclude that the large-scale-duplication/assembly-collapse scenario is unlikely and that the polymorphism in the *Hydra* genome is consistent with that expected for segregating variation. Presumably, this variation arose in an ancestral sexual population, and the asexually reproducing laboratory population of *H. magnipapillata* strain 105 captures two haplotypes from this population. We estimate the SNP rate to be 0.69 +/- 0.04% (or 1 SNP per 144 bp on average).

***Hydra magnipapillata* genome size**

By Feulgen staining of nuclei, the genome size of the 105 strain of *Hydra magnipapillata* has been estimated to be ~1.3 Gb[2]. Measurements of the average genome sizes for other brown

Hydra species vary from 1.15-1.45 Gb[7] (excluding the substantially smaller genome of the green *Hydra viridissima*). We note that measurements from different cell types within the same *Hydra* species yielded 5-10% variation (e.g., 1.18-1.33 Gb in *Hydra vulgaris*[7], a close relative of *H. magnipapillata*). Variations in Feulgen-estimated genome sizes between cell types are thought to depend on different levels of DNA compaction[20]. In any event, the *H. magnipapillata* cells tested stain at a similar level to chicken erythrocytes, suggesting a comparable genome size. While the chicken genome is widely accepted to be 1.25 Gb, the current chicken genome assembly spans 1.05 Gb[21]. Similarly, the *Drosophila melanogaster* genome is 180 Mb in total length, but the genome assembly spans 120 Mb, with the remaining 60 Mb in unsequenceable and/or unassemblable heterochromatin[22].

To estimate the *Hydra magnipapillata* genome size (1C) from the two assemblies, we must make a rough correction for residual redundancy on top of the correction for unassembled reads made above. Assuming 10% redundancy, the RP assembly projects to 90% (1.0 Gb) = 0.9 Gb. Assuming 30% redundancy, the CA assembly projects to 70% (1.5 Gb) = 1.05 Gb. These estimates do not account for cloning or sequencing biases, and only crudely capture the allelic redundancy by a single factor. Nevertheless, the two assemblies show reasonable agreement. We note that most of the difference between the two assemblies lies in the smaller scaffolds and contigs (170 Mb in scaffolds shorter than 5 kb in CA, versus 42 Mb in such scaffolds for RP). Nevertheless, the discrepancy between reported Feulgen-estimated DNA content in *Hydra* and the assemblies is comparable to the discrepancy between the reported genome size of chicken and its assembly. In any event, the completeness of the assemblies relative to expressed genes is high, suggesting that euchromatin is well-sampled.

Selection of the RP assembly for further analysis

Based on the analyses presented above, we selected the RP assembly as the reference assembly for further annotation and analysis. Although this assembly is somewhat shorter than the CA assembly, it has significantly less allelic redundancy as measured by 21-mer counts, EST alignments, and self-alignment. This was the dominant consideration, as both assemblies are similarly complete with respect to expressed protein-coding genes, based on comparison to ESTs, and both show comparable linkage in that ~420 Mb (roughly half the genome) is captured in scaffolds longer than 100 kb in both assemblies. Manual review of several genomic regions of interest confirmed that the RP assembly captured the genome in a non-redundant form preferred for genomic analysis. Both assemblies are deposited in GenBank, providing a useful resource to the community.

S4. Construction of cDNA libraries

A total of 17 different *Hydra magnipapillata* cDNA libraries were constructed. The features of the libraries are listed in Table S2. The libraries were from animals representing various biological states. These included animals producing gametes, animals regenerating heads or feet, and animals treated with alsterpaullone[23] to induce ectopic axes. In addition two suppression subtractive hybridization libraries were made with normal versus interstitial cell-depleted animals and normal animals without buds versus budding and regenerating animals.

S5. EST sequencing

EST sequencing was carried out at the National Institute of Genetics in Mishima, Japan and at the Genome Sequencing Center (GSC) at Washington University, St. Louis. During the initial stages of the project, clones sequenced at the GSC were sequenced from both the 5' and 3' ends. In later stages of the project, sequencing at the GSC was carried out on only the 5' ends of clones. Sequences generated at the NIG were from the 5' end only. Some sequencing of 3' ends at the GSC was carried out using a clamped oligo-dT primer to avoid problems associated with sequencing through the polyA tail. In accordance with the guidelines adopted by the genome sequencing community for the distribution and use of large-scale sequencing data, all of the EST sequences from both the GSC and the NIG have been deposited in dbEST at NCBI.

EST datasets typically cover significantly less than all of the genes in the genome. This is due to insufficient sequencing depth, incomplete representation of developmental stages, tissues and organs, and physiological states, etc. Because the adult *Hydra* polyp is relatively simple in construction and is continuously producing new tissue, undergoing cell differentiation, and patterning, we might expect it to express a broader range of genes than is typical of an adult animal. By adding regenerating, sexually differentiated, and alsterpaullone-treated animals to the set of samples from which cDNA libraires were made, we increased the chances of obtaining ESTs for most of the genes.

S6. Construction of gene models

Homology based gene modeling was done with GenomeScan[24] using the RP genome assembly. Putative loci were found by blastx of soft-masked genomic scaffolds versus the proteomes of *Nematostella vectensis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Additionally, *Hydra* EST assemblies were aligned to the genome with BLAT[18]. The BLAT output was processed so that the best hit to the genome, as well as any other hit within 97% coverage of the best hit were considered matches. Possible pseudogene matches were filtered out by disallowing secondary matches with only one 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 from each organism to GenomeScan and the resulting models were run through PASA[25], which verified/improved some of the models.

AUGUSTUS 2.0.3[26] was trained on 1061 EST assemblies suggested by PASA to be full-length and good training set candidates. These candidate genes were filtered so that they encoded at least 100 amino acids and more than one exon. The gene models were created by running AUGUSTUS on an unmasked version of the genome, incorporating *Hydra* and *Clytia* EST evidence. The majority of the produced models corresponded to transposable element proteins. The models were therefore filtered to remove genes that have more than 50% of their exonic length overlapping with an annotated transposable element (see Supplementary Information Section S9). The final gene model set was produced by running PASA on the filtered AUGUSTUS and the original GenomeScan predictions. About 9,000 models could be verified and/or improved by PASA. PASA-unverified models with no homology to any known protein in the GenBank nr database and no EST support (from *Hydra* or *Clytia*) were removed. If a locus had models from both predictors (GenomeScan and AUGUSTUS), the best model was selected based on its hit e-value to the GenBank nr database and EST support.

The final gene model set contains 31,452 genes. In addition to our annotation of the RP assembly, the CA assembly was annotated by NCBI using the Gnomon gene prediction pipeline (http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml). The 17,835 predicted protein sequences from this annotation have been deposited in GenBank.

The 31,452 predicted protein-coding loci from the RP assembly are an overestimate, as they may include predictions that are not *bona fide* protein-coding genes. Such spurious predictions could arise from unrecognized repetitive elements. Splitting of genes between scaffolds would also contribute to an over-estimate of gene number. Clustering of the 31,452 predicted *Hydra* genes with those of other metazoans identifies 22,083 *Hydra* gene models that have homology to other metazoan genes or are members of paralogous groups in the *Hydra* genome. Of these 22,083 models, 2,117 were found, after manual review, to be from transposons and other repetitive elements. Thus there are 19,966 predicted genes from the RP assembly that meet one of the following three criteria: (1) homology to other metazoan genes; (2) membership in a *Hydra*-specific paralogous group; (3) not from a transposon or repetitive element. This analysis does not capture *Hydra*-specific genes that are not members of paralogous groups.

While the remaining 9,369 predictions (31,452 minus 22,083) may contain additional *bona fide* *Hydra*-specific genes that are not part of gene families, most are likely missed repetitive sequences, artifacts of gene prediction algorithms, or pseudogenes. Only ~30% (2,892) have some (partial) support from ESTs, and of these many (902) are single or two-exon genes that appear to be enriched in pseudogenes. Taken together, we estimate that the *Hydra* genome encodes ~20,000 genes, although this is necessarily a rough estimate based on the above considerations.

S7. Genome browser

A GBrowse[27] type genome browser has been established using the RP assembly. The URL for the browser is http://hydrazome.metazome.net.

S8. Protein distance graphs

Gene families were built with a phylogeny-inferred clustering (http://www.metazome.net) using PSI-BLAST all-against-all scores. The following species were included in the clustering: *Homo sapiens*, *Strongylocentrotus purpuratus*, *Amphimedon queenslandica*, *Paramecium tetraurelia*, *Caenorhabditis elegans*, *Nematostella vectensis*, *Monosiga brevicollis*, *Drosophila melanogaster*, *Dictyostelium discoideum*, *Trichoplax adhaerens*, *Neurospora crassa*, *Hydra magnipapillata*, and *Arabidopsis thaliana*. Alignments of families that had at least three species and less than 500 sequences were built with MUSCLE[28]. The PROTDIST program from Phylip[29] was run with the default Jones-Taylor-Thornton model to calculate pairwise distances between all sequences in a given cluster. Smallest distances for *Homo sapiens* vs. *Nematostella vectensis*, and *Homo sapiens* vs. *Hydra magnipapillata* were collected for all clusters and plotted against each other. Linear regression was done for the data points within the interquartile range (Fig. S4).

S9. Analysis of repeated sequences

ReAS 2.02[30] was used to reconstruct the ancestral repeat sequences from the raw shotgun read data. In order to reduce memory requirements, two libraries were produced and analyzed separately: the first one based on 1% of all reads and 17-mers with a depth of at least 10 (this library is especially enriched in the CR1 class of retrotransposons), and the second one based on 10% of reads and 17-mer depth range of 10 to 100. In order to improve the quality of the assembly, only reads that have at least 100 high-depth 17-mers were considered.

ReAS was run on each of the libraries separately. After retaining only the assembled repeats of length larger than 500 nucleotides, and a minimal average depth value (as provided by the program) of 10, the two libraries contained 949 and 25,110 repeats each, respectively. These sequences were then pulled together. Since the initial ReAS assembly appears to be fragmented and there are many redundant sequences, the final version of the library was produced by running ReAS' join_fragments.pl and rmRedundance.pl scripts. The final library contained 3909 reconstructed repetitive elements with an average length of about 1500 nt.

The annotation steps included a TBlastX run against RepBase[31] (version 13.05), and a BlastX search using a custom non-redundant library from NCBI entries (keywords: retrotransposon, transposase, "reverse transcriptase", gypsy, copia). The elements were then automatically classified based on their homology to known repeats. The library, and especially the elements with a Blast e-value of larger than 1E-30, were manually curated to correct possible errors in the annotations. 2331 repeats could be annotated and verified, about 1500 repeats could not be annotated. The majority of the unannotated elements are putative non-autonomous repetitive elements, having no detectable ORFs and no homology to any other known repetitive elements. Some of the unannotated elements are pseudogenes.

Repeat Masking

Repeat masking was done with RepeatMasker 3.1.9[32] on all available *Hydra* contigs. The counts and the percentage of the genome masked for each of the repetitive elements were derived from the standard RepeatMasker output.

Active transposable element annotation in the EST dataset

Since the repeat library produced with ReAS should provide the ancestral, i.e. full-length and functional sequences, it can be used to annotate the EST assemblies. The requirement for an assignment of an EST contig to a certain class of repetitive element was at least 500 bp of alignment and 70% homology to the ReAS element.

S10. Trans-spliced leaders and operons

Prior to initiation of the *Hydra* EST and genome projects, we had identified two different trans-spliced leaders in *Hydra*[33]. To screen for additional spliced leaders we searched the 5' ends of *Hydra* ESTs for over-represented k-mers[34]. Using this approach we identified eight additional spliced leaders among the *Hydra magnipapillata* ESTs. Additional BLAST searching of the ESTs using these spliced leader sequences led to identification of two additional spliced leader sequences. All of the spliced leader sequences found by searching *H. magnipapillata* ESTs are

present in the genome and have a GT splice donor site. By alignment of spliced leader sequences identified in ESTs with genomic sequence reads, we extended all of the sequences to the same point as SL-B1, one of the two spliced leaders we originally identified.

To determine how many genes in *Hydra* produced transcripts that get trans-spliced we applied the following criteria:

1. The hit has to be at the end of the EST.
2. The hit can be the complement of the query, since some of the 3' ESTs reach to the 5' end of the cDNA. For example both the y read (the 5' read) and the x read (the 3' read) of an EST might both give a hit since the cDNA might have been a short one such that the 3' read goes all the way to the 5' end of the cDNA, and thus contains the complementary sequence of the spliced leader.
3. The hit has to end with one of the following five octamers:

GTAATAAG

TAAATAAG

TTAATAAG

TCAATAAG

TAATAAAG

We had previously identified two different spliced leaders, SL-A and SL-B, in the UCI strain of *Hydra vulgaris*. We have reported the characterization of a *H. magnipapillata* gene whose transcripts contain either SL-B or a newly identified spliced leader that we have termed SL-D[35]. Surprisingly we have not found SL-A, which is quite divergent in sequence from the other spliced leaders, in either the EST dataset for *H. magnipapillata* or the genome sequence. Thus SL-A either evolved in *H. vulgaris* after its divergence from *H. magnipapillata* or has been secondarily lost from *H. magnipapillata*. We prefer the former explanation since an analysis of SL genes from several brown *Hydra* species has not identified an SL-A gene (Stover and Steele, unpublished observations). If this explanation is correct, it suggests that evolution of new spliced leaders can occur relatively rapidly in *Hydra*.

Genes contained in operons in animals which carry out trans-spliced leader addition are separated by intergenic regions ranging from zero to two thousand base pairs. In *C. elegans*, where the most information is available[36], the average intergenic spacing between genes in operons is on the order of 100 bp. The majority of *C. elegans* operons show an intergenic spacing of less than 350 base pairs, but cases are known of spacings as large as 2 kb. The typical *C. elegans* operon contains two genes, with the largest known operon containing eight genes.

To identify potential operons in *Hydra*, we searched the genome for pairs of gene models in which the intergenic distance was unusually small (500 bp or less) and in which ESTs from the downstream gene model contained a spliced leader. 57 such gene pairs were identified. Manual curation confirmed that 32 of these gene pairs were potential operons. This represents a minimum number since some operons may be separated by more than 500 bp[36]. Nevertheless, it appears that the presence of trans-spliced leader addition in *Hydra* has led to the formation of operons as it has in other organisms that undergo trans-splicing[37,38,39]. The list of manually curated operons is in Table S11. An example of one such operon is in Fig. S9. This operon

contains two genes. The upstream gene (which encodes a cyclin H ortholog) contains no introns and the downstream gene (which encodes a von Hippel-Lindau binding protein 1 ortholog) contains four introns. The intergenic distance (from the polyadenylation site on the upstream gene to the trans-splicing acceptor site on the downstream gene) is 153 bp. In this operon, both the upstream gene and the downstream are trans-spliced, as indicated by the finding of spliced leader-containing ESTs for the upstream gene.

S11. The *Curvibacter sp.* genome (GenBank accession numbers: FN543101-FN543108)

The assembled *Hydra* genome contained a number of large scaffolds with unusually high G+C content (in contrast to the low G+C content of the *Hydra* genome) and no high copy repeat sequences typical of *Hydra* contigs. Following the hypothesis of putative bacterial origin of these scaffolds, a systematic screening of the assembly for high G+C content, low repeat proportion, and high coding density was carried out. Annotation of the resulting "not typical of *Hydra*" scaffolds indicated that they contained closely spaced single-exon ORFs with best hits to bacterial genes in the NCBI nr database. These results suggested that the genome of a bacterium had been sequenced together with the *Hydra* genome. To assess the completeness of the bacterial genome sequence we identified the corresponding scaffolds in the CA assembly (Table S12). This assembly is based on longer effective reads than the RP assembly, and as a consequence the assembled scaffolds are longer and have fewer and shorter gaps.

Taxonomic markers consisting of the 50 most universal bacterial COGs in the eggNOG database[40] have been searched in all putative bacterial scaffolds. Eight scaffolds (Table S12, Fig. S11) with similar best BLAST hits were identified, containing 49 out of the 50 taxonomic markers and representing a total of 4 Mb of chromosomal sequence. Thus we estimate the completeness of this bacterial chromosome to be 98%. Two additional scaffolds (HmaUn_WGA4669_1 and HmaUn_WGA72787_1) have similar G+C content but due to the lack of homologs to the 50 taxonomic markers and the differing taxonomy distribution of best BLAST hits (data not shown) their taxonomic origin remains unclear. The eight selected bacterial scaffolds were automatically annotated using PEDANT[41] and SIMAP[42]. A total of 3,782 genes were identified of which 3,544 possess homologs in other species. The single "missing" COG is present in the reads but was not co-assembled. The sequencing depth of these bacterial contigs (3.5X) was less than that of the *Hydra* contigs (6-8X), suggesting the presence of about 0.5 bacterial genomes per *Hydra* genome in the sequenced DNA. To confirm that all eight bacterial scaffolds belong to one bacterial genome, we did phylogenies of the COG sequences that are present on each scaffold. The phylogenies for all the COGs are nearly identical, confirming that the assembled scaffolds belong to one bacterial metagenome. A search for additional bacterial genomes in the *Hydra* assembly using the 50 universal bacterial COGs yielded no result.

To determine the phylogenetic position of the bacterial genome, we identified the nearest neighbors in the COG phylogenies. These indicated that the *Hydra*-associated bacterium is a member of the Comamonadaceae (Burkholderiales). Searching the 16S rRNA sequences of the Ribosomal Database Project[43] indicated the highest similarity to the novel freshwater isolate *Curvibacter delicatus* (98.4% pairwise identity, GenBank Accession AF078756)[44]. Since the identity is lower than the species threshold of 98.7%, we identify the *Hydra*-associated

bacterium as *Curvibacter sp*., an uncultured novel species within the genus *Curvibacter*[44]. Fig. S12 shows a phylogenetic tree of members of the Family Comamonadaceae. Several nodes of the tree are not well resolved, but the placement of our bacterial sequence within the *Curvibacter* branch and neighbored to the genera *Polaromonas*, *Rhodoferax*, and *Acidovorax* confirms our taxonomic assignment to *Curvibacter*. The eight *Curvibacter sp*. scaffolds have been deposited in GenBank under accession numbers FN543101-FN543108.

To infer possible physiological and ecological features of the *Curvibacter* bacterium associated with *Hydra*, the annotated genes were assigned to biochemical pathways in the KEGG database[45]. The occurrence of pathways in *Curvibacter* was generally similar to the other Comamonadaceae, e.g. *Polaromonas*, *Rhodoferax*, and *Acidovorax*. One striking difference was the large number of ABC transporters for sugar uptake present in *Curvibacter* and absent from *Polaromonas*, *Rhodoferax*, and *Acidovorax* (Table S14). These include uptake pathways for xylose, rhamnose, multisugar, sorbitol/mannitol, alpha-glycoside, and oligogalacturonide. Assuming *Curvibacter* is an ectosymbiont occupying a niche associated with the *Hydra* glycocalyx (Fig. S10) and/or the mucous layer surrounding the *Hydra* basal disk, it is reasonable to hypothesize that it uses component molecules at these locations as a source of sugars for nutrition.

Table S13 shows a comparison of orthologous genes between *Curvibacter sp*. and other species of Comamonadaceae. Roughly 60% of *Curvibacter sp*. genes (2200 genes) have orthologs in other species of Comamonadaceae. By comparison, the other species, which have slightly larger gene numbers, share about 2600 orthologs with each other. Of the roughly 1500 *Curvibacter sp*. genes without homologs in the other Comamonadaceae, some encode unknown proteins while others are distributed in various sugar metabolism pathways. None occur in KEGG pathways that were not already represented in the *Curvibacter sp*. genome.

To confirm that *Curvibacter sp*. was not simply a contaminant of the culture from which *Hydra* were harvested in 2004 for genome sequencing, we carried out PCR reactions on a single polyp from a *Hydra magnipapillata* 105 culture at UC Irvine in December, 2008 with primers designed to amplify segments of a protein-coding gene at positions 477054 to 478457 on scaffold HmaUn_WGA9493_1 of the *Curvibacter sp*. assembly. DNA was isolated from a single *Hydra magnipapillata* strain 105 polyp essentially according to the protocol developed for PCR from a single *Drosophila*[46]. A single unfed polyp was placed in a well of a 24 well microtiter plate, *Hydra* medium was removed from the well, and the polyp was washed by addition of Milli-Q water to the well. The washed polyp was transferred to a 0.5 ml tube, and the tube was spun briefly in a microcentrifuge to compress the polyp and make it easier to remove the last traces of water. 50 μl of lysis buffer was added to the tube and the sample was then incubated at 37°C for 30 minutes. Lysis buffer contained 10 mM Tris, pH 8.0, 1 mM disodium EDTA, 25 mM NaCl, and 200 μg/ml Proteinase K (Roche, PCR grade); Proteinase K was added fresh immediately before use from a 20 mg/ml stock solution. The tube was flicked periodically to aid dissolution of the polyp. The tube was then incubated at 95°C for 2 minutes to inactivate the Proteinase K. The sample was stored at -20°C until used for PCR. Two sets of primers for the target gene were designed using the GeneFisher2 server (http://bibiserv.techfak.uni-bielefeld.de/genefisher2/welcome.html)[47]. The sequences of the primers were as follows:

Primer set 1	Primer set 2
5'-GATGCCATCGATTGCTGA-3'	5'-GCATGCGGAGTCTCTCA-3'
5'-CGCAACTCTCGGGCAA-3'	5'-GCGTACCAGAGGCAGA-3'

Amplification was carried out using 1 μl of the *Hydra* polyp lysate and EncoTaq DNA polymerase (Lucigen) under conditions recommended by the supplier. An annealing temperature of 53°C was used for both primer sets. Reaction products were purified using a Qiagen QIAquick PCR Purification Kit, fractionated in a 1% agarose gel, and fragments of the expected size (515 bp for primer set 1 and 516 bp for primer set 2) were extracted from the gel using a Qiagen QIAquick Gel Extraction Kit. The fragments were cloned with an Invitrogen TA Cloning Kit and sequenced by Laguna Scientific. Sequencing confirmed that the products were the expected amplicons. Products of the expected size were not obtained from several other *Hydra* species/strains processed in parallel with *H. magnipapillata* strain 105, ruling out the possibility that *Curvibacter sp.* is a contaminant associated with the *Hydra* culture medium or one of the reagents used for DNA extraction or amplification. Thus the microbiome of *H. magnipapillata* strain 105 at UC Irvine still contains *Curvibacter sp.* four years after animals were harvested for the genome project.

A 16S rDNA sequence 99.8% identical to the *Curvibacter sp.* 16S rDNA sequence has been identified in the microbiome of a strain of *Hydra vulgaris* isolated from the Pohlsee in northern Germany[48], suggesting that *Curvibacter sp.* is associated with *Hydra* in Europe.

S12. Horizontal gene transfer: bacterial genes in the *Hydra* genome assembly

The close and stable association of bacteria with *Hydra*[48] and the absence of a sequestered germline in *Hydra*[49] create a potentially advantageous situation for horizontal gene transfer from bacteria to *Hydra*. A similar situation may have existed in early cnidarian precursors[50]. Hence it appeared worthwhile to carry out a systematic search for bacterial genes in the *Hydra* genome.

To identify horizontal gene transfers (HGTs) into the *Hydra* genome, we performed a blastx screen of the RP genome assembly against the NCBI nr database and retained putative non-metazoan hits, i.e. ORFs with a best hit to a non-metazoan sequence (excluding hits to annotated *Hydra* and *Nematostella* gene models) and with an alien index[44] greater than 30. Alien index (AI) is a bioinformatic measure of sequence relatedness. It compares the best non-metazoan homolog (e-value) to the best metazoan homolog (e-value). Only hits with an e-value better than E-10 were used to calculate alien indices and overlapping sequences were combined to single hits. These are potential candidates for HGT into the *Hydra* genome.

Examination of the putative HGT candidates in the genome assembly showed that most of the hits matched a gene model. Hits to models lying within repeat-masked segments of the genome were removed. The remaining gene models were examined for exon number, the presence of ESTs, the occurrence of a *Nematostella* homolog, and the occurrence of homologs in other metazoans. Finally, the alien index was calculated for the complete gene models and models

with AI > 30 were retained. In order to find genes that were overlooked by the blastx approach due to short exons and consequently low blast scores, we carried out a blastp screen of the NCBI nr database using the *Hydra* gene models. The alien index was calculated for each gene model and putative candidates for horizontally transferred genes were identified. The results of this second blastp screen were generally similar to the results of the blastx screen. Novel candidates found in the second screen were added to the list of putative horizontally transferred genes from the blastx screen.

For each of the HGT candidates we calculated phylogenetic gene trees, using both neighbor-joining and maximum-likelihood methods. A sequence alignment of the HGT candidate and up to 1000 best hits in the NCBI RefSeq database (e-value better than E-10) was created with the alignment program MUSCLE[28] using default parameters. Construction of phylogenetic trees by neighbor-joining was done using the tools PROTDIST and NEIGHBOR (both with default parameters)[29]; construction of phylogenetic trees by maximum-likelihood was done using RAxML[51] with parameters: -m PROTGAMMAJTT -x 12345 -N 5 -f a -T 2. Due to the taxonomic redundancy of prokaryotic protein sequences in the NCBI RefSeq database, a second set of maximum-likelihood phylogenetic trees was calculated using a reduced set of prokaryotic genomes. To create a reduced and non-redundant set of prokaryotes, only the species with the biggest genome per genus were selected in the NCBI RefSeq database. A sequence alignment of the HGT candidate, all eukaryotic hits in the NCBI RefSeq database and all hits in the reduced prokaryote set (e-value cutoff E-10 in both blast searches) was created using MUSCLE with default parameters. Construction of consensus phylogenetic trees from 100 bootstrap replicates by maximum-likelihood was done using RaxML with parameters: -x 12345 -N 100 -f a -T 4. The trees were displayed using the ITOL program[52] and can be viewed at:

http://webclu.bio.wzw.tum.de/hydrahgt/finalHgtGeneModels/overviewTable.html

The topologies of these trees were inspected individually and only HGT candidates that showed a clear discrepancy between the gene phylogeny and *Hydra*'s organism phylogeny (placing the Hydra branch within bacterial lineages) in at least one of the trees were retained.

Finally the HGT candidates were classified according to their confidence level based on the results from all three phylogenetic methods. Cases for which all three phylogenetic trees clearly support HGT are classified as high confidence candidates. For some HGT candidates, a phylogenetic tree using the reduced prokaryote set could not be calculated (indicated by 'NA' in Table S15), since the HGT candidate has hits in only a few bacteria, which were filtered out in the reduction step. If the remaining two phylogenetic trees show good support for HGT and there are no known metazoan homologs, these candidates were also classified as high confidence candidates. Cases in which only one or two phylogenetic trees showed good support for HGT or the phylogenetic tree using the reduced prokaryote set was not available and there are metazoan homologs, were classified as medium confidence candidates. These are candidates, which passed all of the initial filter steps and for which some of the phylogenetic trees indicate HGT, but for which we cannot rule out gene loss in non-cnidarian metazoans. Due to the lack of genome sequences for close relatives of *Hydra*, neither of the two hypotheses for the evolution of these genes can be confirmed at the moment.

Table S15 shows the resulting list of 71 HGT candidates in the *Hydra* genome. The list also

includes a single hit that does not have a matching gene model but passed all the other filter steps. Several hits form clusters of paralogous sequences, in which one or more gene models have good EST support while others appear to be pseudogenes More than 70% of HGT candidates have ESTs and hence represent expressed genes. A large fraction (32%) of these HGTs have a trans-spliced leader sequence (Table S15), indicating unambiguously that they are derived from *Hydra* and not from associated organisms. Table S15 also shows the putative donor bacterial taxon of the HGTs. These were predicted from the nearest bacterial neighbor of the HGT candidate in the maximum-likelihood tree.

Fig. S14 shows the cumulative frequency of genes as a function of exon number in the HGT set and among all *Hydra* gene models. These results show clearly that genes in the HGT set have fewer introns on average than the bulk of genes in the genome. Roughly half of all HGTs are single exon genes compared to 25% of single exon genes in the whole genome. Assuming that introns are added continuously to genes over evolutionary time, this result suggests that genes in the HGT set entered the *Hydra* genome later than vertically inherited genes and consequently experienced less intron insertion. Furthermore, the result is not consistent with the idea that HGT candidates are in fact highly conserved ancestral genes, since in this case no difference in intron number between HGT candidates and vertically inherited genes is expected.

Most putative HGTs have no metazoan homologs and numerous bacterial homologs. These are clearly the best candidates for horizontal gene transfer from bacteria. For 20 candidates in Table S15 there are metazoan homologs scattered in the phylogenetic trees and hence these candidates could be ancestral genes, which were lost in multiple animal lineages. Most of these candidates are single exon genes in *Hydra* as expected of a HGT. There are, however, several putative HGTs, which have multiple introns and distant metazoan homologs in the phylogenetic trees. Although these could be false HGT predictions, closer examination of these cases indicated that they are likely HGTs. An example is Hma2.226010. This gene has an unusual combination of a globin domain and a FAD binding domain. Full-length homologs with e-value = E-70 are found in a large number of bacteria. The only good metazoan hit has an e-value of E-44 and lacks the globin domain. The metazoan hit has multiple introns, but the intron positions are different from the introns in Hma2.226010. Thus, despite multiple introns and a distant metazoan homolog, Hma2.226010 is a good HGT candidate.

HGT events which occurred after the separation of the anthozoan (*Nematostella*) and medusazoan (*Hydra*) lineages, about 550-600 million years ago, would lack homologs in the *Nematostella* genome. There are 26 such HGTs in *Hydra* (Table S15). Assuming that the rate of horizontal gene transfer is constant over time and that the rate of gene loss is low, this implies that a HGT was fixed on average every 20 My in the lineage leading to *Hydra*.

More than 70% of HGTs are expressed (Table S15) and thus could provide a selective advantage to *Hydra*. For the remaining HGTs we have no explicit information about expression and thus we cannot conclude that they have a function. Nevertheless, the fact that these genes have full-length open reading frames suggests that they are functional and presumably selected.

What types of horizontally transferred genes are fixed in the *Hydra* genome? Assignment of the transferred genes to KEGG pathways shows that these genes come from only a small subset of known biochemical pathways. Essentially none of the HGT candidates are in core pathways

involved in DNA replication, transcription or protein synthesis. Most of the genes are in KEGG pathways involved in carbohydrate metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, glycan biosynthesis, cofactor and vitamin metabolism, and xenobiotic metabolism. The biochemical functions of the HGTs in these pathways involve primarily transamination, methylation, and acetylation of sugars, polysaccharides, or glycoproteins. Hence they give rise to modifications to existing structures in cells.

Three HGT candidates encode a branch of the bacterial lipopolysaccharide (LPS) biosynthetic pathway

Most examples of horizontal gene transfer involve genes encoding proteins required for only a single step in a multi-step metabolic pathway. Only in the case of the LPS pathway do the HGTs constitute a biosynthetic pathway. LPS makes up the outer membrane of gram-negative bacteria, and is synthesized by a complex pathway involving intracellular and extracellular steps[45]. One branch of the LPS pathway leads to formation of an activated heptose sugar, which is added to 3-deoxy-D-manno-octulosonic acid, transferred across the bacterial membrane, and then further modified by addition of oligosaccharides[46]. Fig. S13 shows that three enzymes required for formation of the activated heptose are present in the *Hydra* genome. The fourth enzyme, GmhB, is absent, but its phosphatase activity can be replaced by other enzymes[47]. Thus *Hydra* apparently has the machinery to make an activated heptose, which could be added to polysaccharides, proteoglycans, or glycoproteins. It is not yet known what the potential heptose acceptor is in *Hydra* or which cell types express these HGT candidate genes.

Two HGT candidates are expressed in differentiating nematocytes

A homolog of the bacterial gene capA, which is involved in the biosynthesis of poly-gamma-glutamate, has recently been identified in ESTs from the hydrozoan *Clytia hemisphaerica* and shown to be expressed in differentiating nematocytes[53]. It was suggested that the capA homolog entered cnidarians by horizontal gene transfer from a bacterium and contributed to the evolution of nematocyst capsules. This observation led us to search for additional HGTs among nematocyte-specific genes. A comparison of the 71 HGT candidates (Table S15) with 58 genes that have been shown to be strongly expressed in differentiating nematocytes[54,55] only identified the capA homolog (Hma2.220724) and one other gene (Hma2.208759). The remaining 56 nematocyte-specific genes were not present in the HGT list. This indicates that most genes strongly expressed in nematocytes are not derived from bacteria. Thus the results do not provide support for the proposal that nematocytes arose by large-scale gene transfer from a bacterial donor. The results are more consistent with the idea that novel proteins involved in nematocyte formation evolved within the cnidarian lineage[56].

S13. MicroRNA sequencing and analysis

Hydra miRNAs were discovered using the protocols outlined by Wheeler et al.[57]. Briefly, total *Hydra* RNA was size-selected, a small RNA library was constructed, and 454 sequencing was carried out. After enforcing length and quality restrictions, a total of 9654 reads were mapped to the *Hydra* genome assembly. Reads were then clustered and folded with mfold[58] using a cutoff of -20 kcal/mol[59] to find potential pre-miRNA candidates. In addition, miRDeep[60] was

run on all 9654 reads to identify potential microRNA genes. Less stringent filtering parameters regarding the required presence of miR* sequences were used in miRDeep since most of the miRNAs in *Hydra* have just one 454 read assigned to them. There are only two miRNA loci that are supported by more than five reads. One of them is miRbase ID: miR-2022, which is potentially conserved between *Hydra* and *Nematostella*. The other appears to be novel. The combined set of predictions, excluding tRNA and rRNA, contains 51 candidates. After manual curation, only 17 candidates were considered as high-confidence miRNAs; five of these were confirmed as expressed by northern blot analysis (Table S16). All sequences have been submitted to miRBase[61].

In addition to 454 read data, whole-genome alignments were made between *Hydra* and *Nematostella*, and RNAz[62] was run to predict conserved sequence and structural elements. At that evolutionary distance, virtually all of the identified elements were tRNA and rRNAs. One candidate that has a hairpin structure resembling a miRNA and that was not repetitive could be identified on Contig33064:4803..4937, but it has a folding energy of the 70 bp precursor of -11 kcal/mol. The rest of the 51 predicted ncRNAs in *Hydra* are mostly located in repetitive regions since the same "mature" transcript sequence occurs multiple times in the genome. We believe that these genes are potential candidates for repeat-associated siRNAs (rasiRNAs) and piwi-associated RNAs (piRNAs). Elements such as Chapaev, CR1, Mariner, hAT, and P1 show 454 read support, with an average read length of 28 bp. Most highly expressed non-coding RNA genes in the *Hydra* genome, according to 454 read count, are tRNAs. In particular, AspGTC tRNA has 210 454 reads assigned to it with 5' homogeneity of 86%. Other tRNAs had less 454 support, but all of them appear to be cleaved specifically.

S14. Genes missing from the *Hydra* genome

We have found several interesting cases in which genes are missing from the *Hydra* genome. These are briefly described below. In all cases, the assembled genome, unassembled Sanger reads, and 454 reads were queried.

Circadian rhythm genes

Circadian clocks have been identified in a wide range of organisms. The core molecular components of the clock are encoded by the CLOCK, PER, BMAL1/CYC, and CRY genes[63]. Clock and BMAL are present in the anthozoan cnidarians *Nematostella* and *Acropora* [64] but are absent from *Hydra*. Thus central components of the circadian clock appear to have been secondarily lost from *Hydra*. Circadian rhythms have been well-studied in anthozoans but not in hydrozoans. The primary role of circadian rhythms in anthozoans is regulation of sexual reproduction in the wild (e.g. control of spawning in corals), and in the laboratory (spawning in *Nematostella*[65]). Sexual reproduction in *Hydra* appears to be tied to food intake rather than the light/dark cycle.

Fluorescent protein genes

Cnidarians are well known for their fluorescent proteins, such as GFP. Genes encoding fluorescent proteins have been identified in diverse members of the cnidarian classes Anthozoa

and Hydrozoa, and phylogenetic analyses support the hypothesis that the fluorescent proteins in modern cnidarians evolved from a fluorescent protein in the basal cnidarian[66]. The *Nematostella* genome contains four such genes. Queries of the *Hydra* genome with tblastn and amino acid sequences from anthozoan and hydrozoan fluorescent proteins did not identify any genes in *Hydra*. In addition pfam hits on the *Hydra* gene model dataset did not include fluorescent proteins. Thus it appears that genes for fluorescent proteins have been secondarily lost from the *Hydra* genome. In support of this conclusion, scanning of adult *Hydra* for fluorescence at a range of wavelengths did not produce a signal (Steele, unpublished observation).

Genes conferring pluripotency

Hydra contains a multipotent stem cell which gives rise to the various differentiated cell types of the interstitial cell lineage (germ cells, nerve cells, nematocytes, and secretory cells)[67,68,69]. The evolutionary history of stem cells is of obvious interest. In particular, one would like to know if the multipotent stem cell in the interstitial cell lineage of *Hydra* and stem cells in vertebrates have a shared evolutionary origin. Five genes have been shown to induce pluripotency when expressed in differentiated somatic cells of mammals, these being myc, nanog, klf4, Oct4, and Sox2[70]. We have queried the *Hydra* genome for homologues of these genes. Four myc homologues are present (GenBank accession numbers XP_002163473, XP_002164060, XP_002170328, and a fourth gene, RP gene model Hma2.222803, that was not identified by the NCBI annotation pipeline). In situ hybridization has shown that one of the myc genes (XP_002163473) is strongly expressed in all rapidly proliferating subpopulations of the *Hydra* interstitial stem cell system[71]. No nanog homlogue was found. Genes encoding members of the Krüppel-like family are present, but no homologue of kfl4 was found. Oct4 is a member of class V of the POU family. Members of the POU family are present in *Hydra*, but class V is absent. Class V family members are also absent from *Nematostella*[72], a finding that indicates that the stem cnidarian did not have class V genes. The *Hydra* genome contains two genes encoding members of the Sox B group (GenBank accession numbers XP_002161378 and XP_002154370), the group that includes vertebrate Sox2. However, the evolutionary relationship of cnidarian Sox B group genes to vertebrate Sox2 genes is not yet clear[73,74]. Taken together, these results suggest that the molecular circuitry responsible for establishing and maintaining multipotent stem cells in *Hydra* has evolutionary origins independent from mammals. Studies of diverse cnidarians support this scenario. Multipotent interstitial cells have been identified in other hydrozoans[75,76,77,78], but they appear to be absent from anthozoans[79], scypozoans[80], and cubozoans[81]. Thus the multipotent stem cells of the sort seen in hydrozoans likely arose as an independent event after the diversification of cnidarians. The identities of the genes conferring multipotency on interstitial stem cells in hydrozoans, and their evolutionary relationships to genes in other animals will be of obvious interest.

Head activator

Head activator is an 11 amino acid peptide that accelerates head regeneration in *Hydra*. It has been the focus of numerous studies of developmental processes in *Hydra*. Head Activator was originally isolated from the sea anemone *Anthopleura elegantissma*, and subsequently from *Hydra*[82]. It was also reported to be present in mammals[83]. A gene encoding a putative receptor for Head Activator has been cloned from *Hydra*[84]. Numerous efforts to clone a gene or a cDNA encoding Head Activator have been made, all unsuccessful. The Head Activator peptide

sequence is not encoded as such in the *Hydra* genome sequence. Thus the origin of the Head Activator peptide remains unresolved.

S15. Organizer genes

The *Hydra* head organizer is located at the apical tip of the hypostome and represents the major signalling center responsible for setting up and maintaining positional values along the oral-aboral body axis. It has the capacity to induce a secondary body axis when transplanted to an ectopic position along the body column[85]. The cnidarian oral pole is derived from a blastoporal signalling center acting at the late gastrula stage and establishing axial positional values in the developing primary polyp[86]. The molecular nature of the *Hydra* head organizer is incompletely understood. However, Wnt signalling seems to play a critical role in head organizer formation and in transmitting the capacity for head formation in *Hydra*[23,87,88,89,90]. Wnt/beta-catenin signalling has also been demonstrated to define the position of the blastoporal organizer and to act in axial patterning during embryogenesis in anthozoans and hydrozoans[91,92,93,94].

The evolutionary origin of axis inducing organizers is under discussion[95-96]. Recently, it has been proposed that an organizer was present at the base of chordate evolution[97] but it is currently unknown to what extent cnidarian and chordate organizers share elements of genetic regulation. To approach this question, we have searched the *Hydra* and *Nematostella* genomes for genes encoding signalling factors and transcription factors known to act in the best investigated signalling center, the Spemann-Mangold organizer in *Xenopus*[98]. The genomes of both *Hydra* and *Nematostella* contain a large number of orthologs of vertebrate organizer genes, particularly those encoding modulators of BMP and Wnt signalling and critical transcription factors. Also, many of these orthologs have been shown to be expressed in the *Hydra* head organizer and/or in the blastoporal organizer in *Nematostella* embryos. We therefore propose that a blastopore signalling center with a core set of genetic elements acting in axial patterning evolved before cnidarians and bilaterians diverged.

It should be added that our data also show notable differences between cnidarian and vertebrate organizers. Genes encoding the transcription factors Siamois, Anf-1, and Twn were not found in the cnidarian genomes. This may indicate that they evolved later associated with organizer functions specific for chordate/vertebrate development. Furthermore, some of the orthologs in Table S18 are present in cnidarian genomes, but are not expressed in the cnidarian organizers. In particular, several of the genes encoding BMP and Wnt modulating factors are activated outside of the blastopore at more aboral positions in *Nematostella* gastrulae. Here, additional data are required to more accurately define the orientation of BMP and Wnt signalling gradients and their role in axis determination in cnidarians in comparison to chordates/vertebrates.

S16. Neuromuscular junctions

Efforts to demonstrate that acetylcholine is a neurotransmitter in cnidarians in general and *Hydra* in particular have yielded conflicting results[99]. It remains unclear whether cnidarians produce and use acetylcholine, and in particular, whether they have cholinergic NMJs. It has

been suggested that neurotransmission in the ancestral cnidarian was performed by neuropeptides, of which *Hydra* has on the order of 500, but not by the classic small molecule neurotransmitters[4]. The studies searching for acetylcholine in cnidarians have used indirect methods, e.g. histochemical reactions or radioimmune assays. As far as we are aware, no chemical purification and identification of acetylcholine has been done on a cnidarian. The *Hydra* and *Nematostella* genome sequences offer the first opportunity to examine whether cnidarians have the necessary pieces to produce a canonical neuromuscular junction that uses acetylcholine as a neurotransmitter.

The neuromuscular junction is complex, but its defining protein components are relatively small in number. These are listed in Table S20. We have searched both the *Hydra* and the *Nematostella* genomes for genes encoding these proteins. The first step in the function of the NMJ is the uptake of choline by choline transporters in the synaptic regions of nerve cells. *Hydra* and *Nematostella* both have two genes encoding homologs of bilaterian choline transporters. One of the *Hydra* choline transporter genes (HyChT1, model Hma2.233321) is expressed in ectodermal epithelial cells, and the other (HyChT2, model Hma2.229026) is expressed in nerve cells (Takahashi and Fujisawa, unpublished observations). Once choline is taken up by neurons it is acetylated to produce the neurotransmitter acetylcholine. This reaction is carried out primarily by choline acetyltransferase (ChAT)[100]. Identification of genes for ChAT in cnidarian genomes is complicated by the fact that transferases that acylate carnitine (CrAT) are closely related in sequence to ChAT[101]. Tblastn queries of the *Hydra* genome with a bilaterian ChAT amino acid sequence yields hits to five gene models. In an attempt to determine whether these models are choline acetyltransferases or carnitine acyltransferases, we examined the *Hydra* sequences and sequences from a range of other animals for two residues known to be involved in substrate discrimination[102]. If these positions are V and N respectively, the enzyme has ChAT specificity. If the positions are T and R respectively, the enzyme has CrAT specificity. *Hydra* has no enzymes with V and N (all are T or S and R or P). We find the same situation in *Nematostella*. So a simple conclusion would be that the creation of separate enzymes for acetylcholine and carnitine happened after the divergence of bilaterians and cnidarians. An enzyme with T and R can use choline as a substrate, but its catalytic efficiency favors carnitine[102]. We have no information about the kinetic properties of the predicted *Hydra* enzymes, and we are focusing only on two residues. Other changes in the *Hydra* enzymes could conceivably allow them to use acetylcholine efficiently.

Once ACh is produced, it is transported into intracellular vesicles by a vesicular acetylcholine transporter (VAChT). We have been unable to identify a VAChT ortholog in either *Hydra* or *Nematostella*. Interestingly, the VAChT gene has been shown to be located in the first intron of the ChAT gene in mammals, amphioxus, *Drosophila*, and *C. elegans* and to share exon 1 with the ChAT gene[103]. Thus this arrangement must have been present in the basal bilaterian. Absence of a VAChT in *Hydra* and *Nematostella* does not, however, preclude release of ACh from cells of these animals. Release of ACh from some non-neuronal cells appears not to depend on vesicular sources and may in some cases involve organic cation transporters[104].

Upon release from the nerve cell, acetylcholine binds to the nicotinic acetylcholine receptor on the muscle cell. *Hydra* has seventeen genes encoding nicotinic acetylcholine receptor (nAChR) subunits. Expression of five of these genes could be detected by RT-PCR (Fig. S16). Expression of only one of the genes could be detected by in situ hybridization, with expression being

present in the ectodermal epithelial cells of the presumptive tentacle region in buds and in the tentacles of adult polyps (Fig. S17). The nAChR is clustered at the NMJ in vertebrates primarily by the combined actions of agrin (an extracellular matrix protein), MuSK (a receptor protein-tyrosine kinase), and rapsyn (a tetratricopeptide repeat-containing cytoplasmic protein)[105]. MuSK is part of a four member family that includes ROR, TRK, and DDR receptor-protein kinases[106]. *Drosophila* has a MuSK homolog[107], indicating that MuSK was present in the basal bilaterian. *Hydra* has genes that appear to encode homologs of ROR and DDR, but lacks TRK and MuSK homologs. In addition, *Hydra* lacks rapsyn and agrin homologs. *Nematostella* also lacks MuSK, rapsyn, and agrin homologs.

Recovery at the NMJ depends on acetylcholine degradation by acetylcholinesterase. *Nematostella* has two genes encoding putative acetylcholinesterases (GenBank accession numbers XP_001629673 and XP_001631050). A TBLASTN query of the *Hydra* genome with this *Nematostella* protein identifies a *Hydra* gene (Hma2.213104) that encodes a predicted esterase. Comparison of this protein to the *Nematostella* protein, and *Torpedo*, human, *Drosophila*, and *C. elegans* acetylcholinesterases indicate that the *Hydra* protein is not orthologous to these proteins. While the *Nematostella* protein contains all three residues of the catalytic triad[108], the predicted *Hydra* protein lacks two of the three catalytic triad residues. Proteins containing catalytically inactive AChE domains function as cell adhesion molecules and receptors[109]. Thus *Hydra* appears to lack a gene encoding a typical animal acetylcholinesterase. Recent work has demonstrated cholinesterase activity in medusae of the hydrozoan *Clytia hemisphaerica* using a histochemical assay[110]. In addition this study described three partial cDNA sequences predicted to encode carboxylesterases. Comparison of the protein sequences predicted from these cDNAs indicates that they, like the predicted *Hydra* protein, lack two of three catalytic triad residues and thus are not AChE orthologs. Given the presence of a gene encoding a typical AChE in *Nematostella*, it would appear that the absence of an AChE gene in *Hydra* is a derived feature.

S17. Evolution of cell-cell and cell-substrate junctions

Cnidarians have true epithelia, a characteristic shared with bilaterians, possibly with homoscleromorph sponges[111], but apparently not with most ancestral metazoans[112,113]. Phylogenetic distribution of metazoan cell contact-specific genes shows that the genomes of the cnidarians *Hydra* and *Nematostella* encode a nearly complete set of proteins known to participate in the formation of cell-cell and cell-substrate junctions of bilaterians (Fig. 3b). Included are genes encoding the structural and adhesive proteins connecting neighboring cells as well as the cytoplasmic proteins associated with the junctions. For *Hydra* and *Nematostella*, a filled box in Fig. 3b indicates a gene identified by the presence of protein domains, amino acid motifs, or conserved cysteine patterns diagnostic for the corresponding protein family. Protein domain patterns were confirmed by SMART/Pfam analysis. A representative set of SMART/Pfam-based protein structures of predicted *Hydra* cell-cell and cell-substrate junction proteins is shown in Fig. S18. For the choanoflagellate *Monosiga*, the sponge *Amphimedon*, the placozoan *Trichoplax*, and *Nematostella*, data were taken from the published EST and genome sequencing projects[114,115,116,117,118,119,120,121,122].The presence and function of cell-cell and cell-substrate junction proteins in *Drosophila* and humans have been extensively described earlier[123,124,125,126]. Accession

numbers of the predicted junction proteins from Fig. 3b that are not available in the literature are listed in Table S21.

Sakarya et al.[121] reported a classic cadherin with a bilaterian-type cytoplasmic domain in the sponge *Amphimedon* (question mark in Fig. 3b). Amino acid sequence alignment of this domain shows weak similarity to the highly conserved cytoplasmic domains of eumetazoan classic cadherins. Furthermore, no data are available to confirm that the *Amphimedon* cytoplasmic domain is able to bind to beta- and/or p120/delta-catenin proteins. Hence, it is presently uncertain whether the *Amphimedon* protein represents a classic cadherin. In contrast, experiments show clearly that *Hydra* cadherin interacts with beta- and p120/delta-catenin proteins (Kostron, Holstein, and Hobmayer, unpublished data).

There are 17 innexin genes in the *Hydra* genome assembly (Table S21). This striking expansion of innexin genes also occurs in *Clytia hemispherica*, a marine hydrozoan polyp. We have found 8 innexin genes in the publicly available *Clytia* EST collection at NCBI. Hence multiple innexin genes appear to be a feature of at least the class Hydrozoa within Medusozoa. In *Hydra* this expansion is associated with numerous gap junctions between epithelial cells, across the mesoglea between ectodermal and endodermal epithelial cells, between epithelial cells and nerve cells, and between nerve cells[127,128]. By comparison, there is only one innexin gene in the genome of the anthozoan *Nematostella* and there is currently no electron microscopic evidence for gap junctions in *Nematostella* or any other anthozoan.

Characteristic sequence features of innexin/pannexin genes are conserved in all cnidarian innexin genes (Fig. S19). The genes are generally about 400 amino acids long and have four transmembrane domains (TM). The N terminus and C terminus are predicted to be intracellular. This creates extracellular loops between TM1 and TM2 and between TM3 and TM4. These loops contain two conserved cysteine residues in invertebrate innexins and vertebrate pannexins and also in the *Nematostella* homolog. The *Hydra* and *Clytia* innexins have four cysteine residues in a slightly expanded loop between TM1 and TM2.

TM2 has two conserved proline residues separated by 12 hydrophobic residues. TM3 has a conserved lysine or arginine in the middle. These features are characteristic of almost all innexins and pannexins. They are also present in the *Hydra*, *Clytia*, and *Nematostella* innexins, suggesting that they represent ancestral features. There are no innexin genes in the choanoflagellate *Monosiga*, the placozoan *Trichoplax*, or the sponge *Amphimedon*. Thus, the innexin gene is one of the clearest examples of a novelty in the eumetazoan ancestor of Cnidaria and Bilateria.

S18. Telomeres

Traut et al.[129] have shown that a TTAGGG probe, corresponding to the most common telomeric repeat sequence in animals, hybridizes to the telomeres of *Hydra* metaphase chromosomes. Among the contigs from the *Hydra* genome assembly we have found ones that end in tandemly repeated arrays of TTAGGG (e.g GenBank accession number ABRM01014206). Interspersed among the TTAGGG repeats are the variant sequence TCAGGG. These data indicate that *Hydra* telomeres are typical of those found in other animals.

Supplemental Tables S1-S21 and Figures S1-S19

TABLE S1: Comparison of *Hydra magnipapillata* and *Nematostella vectensis*

Character	*Hydra magnipapillata*	*Nematostella vectensis*
Class	Hydrozoa	Anthozoa
Habitat	freshwater ponds, lakes, and streams	brackish estuaries
Asexual Reproduction	budding	fission
Sexual Reproduction	monoecious, dioecious, or hermaphrodite	dioecious
Larval Stage?	no	yes
Medusa Stage?	no (lost secondarily)	no (absent from all anthozoans)
Genome Size	1.3 x 10^9 base pairs	0.45 x 10^9 base pairs
Diploid Chromosome Number	30	30
Symmetry	radial	radial/bilateral
Gastrulation	multiple ingression	invagination
Polyp State	solitary	solitary
Cell Number in the Adult Polyp	50,000-100,000	not determined
Mitochondrial Genome	two 8 kb linears	single 16 kb circle

TABLE S2: *Hydra magnipapillata* cDNA libraries used for EST sequencing

Library Name	dbEST Library ID	Strain	Source of RNA	Library Features	Number of Sequences	Sequencing Center
Hydra cDNA Library	12062	105	adult polyps with all bud stages	pA, dT, D	1816	GSC
Hydra EST-II	12496	105	adult polyps with all bud stages	pA, dT, D	2487	GSC
Hydra EST-III	12992	105	adult polyps with all bud stages	pA, dT, D	10186	GSC
Hydra EST-IV	13867	105	adult polyps with all bud stages	pA, dT, D	14612	GSC
Hydra EST Kiel 2	15649	105	SSH, normal polyps without buds vs. budding and regenerating polyps	dT, ND	3634	GSC
Hydra EST Darmstadt I	15651	sf-1	adult polyps	pA, dT, D	30715	GSC
Hydra EST Kiel 5	15943	sf-1	SSH, normal vs. I cell depleted	dT, ND	5454	GSC
Hydra EST UCI 5	15944	105	adult polyps with stage 1-5 buds	pA, dT, D	14199	GSC
Hydra EST UCI 5 ALP	16043	105	ALP-treated adult polyps	pA, dT, D	9771	GSC
Hydra EST UCI 6	16418	105	adult polyps with all bud stages	pA, dT, D	7620	GSC
Hydra UCI 6-barcoded	16704	105	adult polyps with all bud stages	pA, dT, D	8235	GSC
Hydra EST UCI 7	16748	105	feet, peduncles, and regenerating foot ends	pA, dT, D	14905	GSC
Hydra EST UCI 8	16944	105	ALP-treated, smaller size fraction from cDNA used to make UCI 5 ALP library	pA, dT, D	3879	GSC

Hydra EST UCI 9	17834	nem-1	excised testis-containing regions	pA, dT, D	8085	GSC
Hydra EST UCI 10	17835	nem-1	excised egg-forming regions	pA, dT, D	8012	GSC
Hydra magnipapillata cDNA library	14408	105	adult polyps with all bud stages	pA, dT, D	19611	NIG

ABBREVIATIONS

pA = polyA+ RNA used as starting material

TR = total RNA used as starting material

dT = oligo-dT primer used for first strand cDNA synthesis

N = normalized

D = directionally cloned

ND = non-directionally cloned

ALP = alsterpaullone (inhibitor of glycogen synthase kinase-3β)

SSH = Suppression Subtractive Hybridization

GSC = Genome Sequencing Center, Washington University, St. Louis, Missouri

NIG = National Institute of Genetics, Mishima, Japan

TABLE S3: *Hydra* repeat classifications

Classification	Count	Bases masked	Percent genome
DNA transposons			**20.94**
"cut and paste"			
Mariner (Tc1, Pogo groups)	101915	33978052	4.17
hAT	160081	48803241	5.99
Kolobok	11063	4649949	0.57
PiggyBac	882	394512	0.05
Harbinger	2036	1029207	0.13
Transib	139068	30095968	3.69
P	25327	9922419	1.22
MuDR	90	38419	0.005
En/Spm	1707	542572	0.07
IS4EU	3736	1319150	0.16
Chapaev	61673	15307096	1.88
Rehavkus	216	78530	0.01
Unclassified	41996	12382422	1.52
"self-synthesizing" Polintons	9225	5533623	0.68
"rolling circle" Helitrons	17798	6494962	0.80
Retrotransposons			**21.15**
LTR retrotransposons			
Gypsy	74678	17195828	2.11
BEL	2828	1529047	0.19
Copia	1636	445088	0.05
Unclassified	576	309349	0.04
Non-LTR retrotransposons			
CR1 (CR1 and L2 groups)	320941	123939640	15.21
RTE	4391	1943415	0.24
L1 (L1, Tx1)	3882	2686185	0.33
R2	77	28894	0.004
Jockey	1170	594972	0.07

SINE	1146	196662	0.02
Unclassified	15284	3983094	0.49
Penelope	79159	17840761	2.19
ERV	1714	889518	0.11
Unclassified retrotransposons	2344	708519	0.09
Unclassified repeats/fragmented (usually no ORFs)	439092	126741600	**15.56**
Total repeats/TEs			**57.64**
Low-complexity	220518	14734514	1.81
Simple	140886	11711230	1.44
Other	15527	5633909	0.69

TABLE S4: Characteristics of library groups used in the RP *Hydra* genome assembly

Pseudo-Library Name	Number of Reads with length >400 bp and mate >400 bp	Nominal insert size (bp)	Observed insert size and standard deviation (bp)
HYDA	4,559,864	3,500	3,280 +/- 330
HYDB	758,416	9,000	8,910 +- 1,440
HYDC	1,099,964	9,500	8,930 +/- 1,220
HYDF	2,903,966	10,500	7,000 +/- 830
Total	9,322,210	n/a	n/a

TABLE S5: Scaffold size distributions for CA and RP assemblies

Scaffold size	Number of scaffolds (CA)	Total combined length (CA)	Number of scaffolds (RP)	Total combined length (RP)
any	132,902	1,256,186,585	39,356	881,879,712
≥ 2 kb	37,311	1,127,144,048	21,094	856,723,085
≥ 5 kb	24,885	1,086,810,796	15,466	839,426,677
≥ 10 kb	19,955	1,049,826,077	12,997	820,982,870
≥ 20 kb	14,450	971,422,772	9,907	776,340,964
≥ 50 kb	8,811	808,742,983	5,425	629,131,692
≥ 100 kb	2,670	421,532,502	2,425	418,217,852
≥ 200 kb	438	119,341,401	597	167,018,937
≥ 500 kb	8	4,659,988	11	6,890,898

TABLE S6: Contig size distributions for CA and RP assemblies

Contig size	Number of contigs (CA)	Total combined length (CA)	Number of contigs (RP)	Total combined length (RP)
any	236,282	1,185,657,612	145,808	814,657,765
≥ 2 kb	102,761	1,001,792,677	97,982	746,037,531
≥ 5 kb	57,752	860,547,878	51,353	596,842,955
≥ 10 kb	32,482	681,977,602	23,320	398,243,952
≥ 20 kb	12,551	399,661,136	5,663	155,253,616
≥ 50 kb	1,058	67,917,841	107	6,160,673

TABLE S7: Intrascaffold gap size distributions for CA and RP assemblies

Gap size	Number of gaps (CA)	Number of gaps (RP)
≥ 10 bp	103,380	106,452
≥ 100 bp	61,209	65,499
≥ 1 kb	22,107	21,509
≥ 5 kb	1,330	744

TABLE S8: Copy number computed by 21-mer counting and self-BLAST

copy number	CA		RP	
	21-mer (% of all 21-mers)	Bases self-aligned (% of all bases in aligned scaffolds)	21-mers (% of all 21-mers)	Bases self-aligned (% of all bases in aligned scaffolds)
0	n/a	72,295,245 (6%)	n/a	69,622,620 (8%)
1	407 M (71%)	660,019,961 (55%)	430 M (87%)	655,567,974 (78%)
2	115 M (20%)	360,042,848 (30%)	37 M (8%)	84,270,732 (10%)
3+	54 M (9%)	113,121,174 (9%)	25 M (5%)	34,574,609 (4%)
Total distinct 21-mers	575 M	n/a	492 M	n/a

TABLE S9: Trans-spliced leaders in *Hydra*

Name	Sequence
SL-A*	CAAACTTCTATTTTCTTAATAAAG
SL-B1	ACGGAAAAAAACACATACTGAAACTTTTTAGTCCCTGTGTAATAAG
SL-B2	ACGGAAAGAGACACATACTGAAACTTTTTAGTCTTTGTGTAATAAG
SL-B3	ACGGAAAAAAACACATACTGAAACTTTTTATTCCCTGTGTAATAAG
SL-B4	ACGGAAAAAAACACATACTGAAACTTTTTAGTCTCTGTGTAATAAG
SL-C	ACGGAAAAAACGCATTATTAACTCTTGTTTTATTGCGTAAATAAG
SL-D	ACGGAAAAAACACAATAAACAAACAGTTCTATTTGTGTTAATAAG
SL-E	ACGGAAAAAACACAAACAAAACTCGACGTAGAATTTGTGTTAATAAAG
SL-F1	ACGGAAAAAAACACATCTAAACTTCATTTAAGTATTTGTGTCAATAAG
SL-F2	ACGGAAAAAAACACATCTAAACTTTTTTTAAGTATTTGTGTCAATAAG
SL-F3	ACGGAAAAAAACACATCTAAACTTGTTTTAAGTATTTGTGTCAATAAG
SL-G	ACGGAAAAAACACAAACAAAACTCAACGTTAAATTTGTGTTAATAAG

*SL-A is found in *Hydra vulgaris* but not in *Hydra magnipapillata*.

TABLE S10: Trans-spliced leader sequences in non-*Hydra* hydrozoans

Species	Sequence	References
Podocoryne carnea	GACAGATTAGAATACTCAAACTACTTCTAAGTCACTGAGTATAAG ACAGATAAAAAATACTCACACTACTTTTAAGTCCCTGAGTATAAG GAAAATACTCACACTACTTTCAAGTCCCTGAGTATAAG GAAAATACTCACACCTTTTTGAAGTCCCTGAGTATAAG ATACTCACACTACTTTTAAGTCCCTGAGTATAAG TACTCAAACTACTTCTAGGTCCCTGAGTTTAAG ACTCATACTACTTCTGAGTCCCTGAGTATAAG CACACTACTTTTAAGTCACTGAGTATAAG CACTACTTTTAGGTCCCTGAGTTTAAG CTACTTCTGAGTCCTTGAGTATAAG CTACTTTCAAGCCCCTGAATATAAG CTACTTTCAAGCCTTTGAGTATAAG CTACTTTCAAGCCTATGAGTATAAG TTCAAGTCCCTGAATATAAG	GenBank entires AY508722 and AYA493987; Bode et al., ESTs deposited in GenBank; Jürg Spring, personal communcation
Hydractinia echinata	GACGGATAAAAAAACCCACACTATTTCTAAGTCCCTGAGTTTAAG CGGATAAAAACTCAAACTACTTTTAGGTCCCTGAGTTTAAG AAAAAAACTCACACTATTTCTAGGTCCCTGAGTTTAAG AAAAACTCACACTTTTTCTAGGTCCCTGAGTTTAAG AAAAGCTCACACTATTTCTAGGTCCCTGAGTTTAAG AAAAACTCACACTATTTCTAAGTCCCTGAGTTTAAG AAACTCACACTATTTCTAGGTCACTGAGTTTAAG ACTATTTTTAGGTCCCTGAGTTTAAG	GenBank entries AY836588, AY944220, and AF312733; Bode et al., ESTs deposited in GenBank
Clytia hemisphaerica	GACAGATAAAAAAATTCACTCCATTAAGAATTAGTGAATAAG	GenBank entries DQ872898, EU374716, EU374717, and DQ138605; Derelle R. et al., ESTs deposited in GenBank
Eleutheria dichotoma	ACAGATAATACACAAACTAATCTTGAGTCACTGTGTATAAG	Kuhn et al. (1996)[130]
Tubularia sp.	AAATTATACTCACACATTTCTAGTCCCTGAGTAAAAG	Davis, R.E. (1997)[131]
Cladonema radiatum	ACAGATTAAATACACATACTAAACCGAGTCACTGTGTATAAG	Suga et al. (2008)[132]

Potential spliced leader sequences in non-*Hydra* hydrozoans were identified initially by manual inspection of sequences at 5' ends of hydrozoan cDNA sequences in GenBank and comparison of these sequences to the spliced leader sequences from *Hydra*. Potential *Podocoryne*, *Hydractinia*, and *Clytia* spliced leaders identified in this way were subsequently confirmed by

comparison to ESTs datasets as they became available (*Hydractinia* and *Podocoryne* - Bode *et al*., unpublished sequences deposited in GenBank; *Clytia* - Derelle *et al*., unpublished sequences deposited in GenBank) and by cloning of a gene encoding a spliced leader RNA in the case of *Podocoryne* (Jürg Spring, personal communication). The *Cladonema* spliced leader was identified by inspection of four cDNA sequences; one of these sequences is in GenBank but not published (accession number AB379658) and the other three have been published[132]. Potential *Eleutheria* and *Tubularia* spliced leaders were identified by inspection of a single published cDNA sequence from each of these organisms and thus should be considered provisional.

TABLE S11: Predicted operons in *Hydra*

Upstream Gene Model	Upstream Gene Product	Introns in upstream gene?	Downstream Gene Model	Downstream Gene Product	Introns in downstream gene?
Hma2.225661	70 kDa chaperone	no	Hma2.225660	5'-bisphosphate nucleotidase	yes
Hma2.207986	mitochondrial 3-oxoacyl-ACP synthase	no	Hma2.207987	smyd family protein	yes
Hma2.213225	DUF1757 protein	yes	Hma2.213226	thioredoxin/UBQ protein	yes
Hma2.229935	ski-interacting protein	yes	Hma2.229934	RNA Pol II, 14.5 kD subunit	yes
Hma2.228141	dihydrofolate reductase	no	Hma2.228140	ELAC	yes
Hma2.223885	TSSC1	no	Hma2.223886	GCP-2	yes
Hma2.212148	SUMO activating enzyme	yes	Hma2.212149	oligomeric golgi complex 8 component	yes
Hma2.226597	Snf 7 family	yes	Hma2.226596	abhydrolase	yes
Hma2.217873	ubiquinol-cytochrome c reductase hinge protein	yes	Hma2.217872	cdc2-related kinase	yes
Hma2.226582	hydroxypyruvate isomerase	yes	Hma2.226581	L antigen	no
Hma2.203742	stomatin-like	yes	Hma2.203741	PEST-containing nuclear protein	yes
Hma2.219194	RAD23 homolog b	yes	Hma2.219193	LPS biosynthesis protein	yes

Hma2.206682	NOL1/NOP2/Sun domain family	yes	Hma2.206683	nucleolar protein 5	yes
Hma2.230445	VMA21 domain protein	no	Hma2.230446	WD40 domain protein	no
Hma2.212449	macrophage erythroblast attacher	yes	Hma2.212448	novel	yes
Hma2.212213	THO complex 3	yes	Hma2.212212	mitochondrial ribosomal protein L9	yes
Hma2.208661	transaminase	yes	Hma2.208662	Paf1, RNA polymerase II associated factor	yes
Hma2.218695	zinc finger, HIT type 1	no	Hma2.218696	26 proteasome complex subunit DSS1	yes
Hma2.227180	NAD-dependent oxidoreductase	yes	Hma2.227179	WD40/Sof1	yes
Hma2.223338	WD40 family/ST kinase receptor associated protein	yes	Hma2.223336	annexin	yes
Hma2.211159	novel	yes	Hma2.211158	bHLHzip	yes
Hma2.231564	cyclin H	no	Hma2.231563	von Hippel-Lindau binding protein 1 (prefoldin subunit 3)	yes

Hma2.220318	RNA helicase	no	Hma2.220319	adenine phosphoribosyl transferase	yes
Hma2.220932	peptidase C19	yes	Hma2.220931	NADH-ubiquinone oxidoreductase 1 beta subcomplex 2; NDUFB2	yes
Hma2.227810	eIF-2B	no	Hma2.227808	ubi-d4/requiem	yes
Hma2.214911	cytochrome c-1	yes	Hma2214910	DUF1135 family	no
Hma2.201470	novel	no	Hma2.201471	AlkB	yes
Hma1.100486	coiled-coil protein	yes	Hma2.209875	Use1 membrane fusion protein	yes
Hma2.224949	novel	yes	Hma2.224948	nucleoporiln SEH1-like	yes
Hma2.226657	novel	no	Hma2.226658	MAM33 superfamily	yes
Hma2.206494	5-fTHF cyclo-ligase	no	Hma2.206495	RNA Pol I, polypeptide D	yes
Hma2.227534	Apc13p	yes	Hma2.227533	novel	yes

TABLE S12: *Curvibacter sp.* scaffolds in CA *Hydra* genome assembly

Scaffold	Length (bp)	% G+C	Number of coding sequences	Average gene size (bp)	% Coding sequence
HmaUn_WGA26082_1	77,787	60	75	903	87.1
HmaUn_WGA3814_1	110,796	61	108	928	90.5
HmaUn_WGA70434_1	132,027	59	129	929	90.8
HmaUn_WGA12741_1	314,796	57	303	937	90.2
HmaUn_WGA71069_1	564,641	59	548	915	88.8
HmaUn_WGA70176_1	583,683	60	554	962	91.3
HmaUn_WGA9493_1	642,420	60	606	960	90.6
HmaUn_WGA69518_1	1,568,541	58	1459	963	89.6
Sum or Average	3,994,691	59	3782	950	89.9

TABLE S13: Orthologous proteins of Comamonadaceae

	Hydra-associated *Curvibacter sp.*	*Rhodoferax ferrireducens* T118	*Acidovorax avenae* subsp. *citrulli* AAC00-1	*Delftia acidovorans* SPH-1	*Polaromonas naphthalenivorans* CJ2
Hydra-associated *Curvibacter sp.*	3782 (100%)	2292 (61%)	2179 (60%)	2253 (60%)	2175 (58%)
Rhodoferax ferrireducens T118	2359 (53%)	4418 (100%)	2368 (54%)	2539 (57%)	2581 (58%)
Acidovorax avenae subsp. *citrulli* AAC00-1	2219 (46%)	2360 (50%)	4709 (100%)	3177 (67%)	2607 (55%)
Delftia acidovorans SPH-1	2386 (40%)	2619 (43%)	3207 (53%)	6040 (100%)	2839 (47%)
Polaromonas naphthalenivorans CJ2	2247 (46%)	2603 (53%)	2511 (51%)	2719 (55%)	4929 (100%)

The cells show the numbers and percentages of orthologous proteins in the left genome, compared to the upper genome.

TABLE S14: Occurrence of sugar ABC transporters in *Curvibacter sp.* and closely related species

Sugar (transporter gene)	*Curvibacter sp.*	*Rhodoferax*	*Acidovorax*	*Delftia*	*Polaromonas*
Ribose (RsbB, RsbC, RsbA)	+	+	+	+	-
D-Xylose (XylF, XylH, XylG)	+	-	-	-	-
Fructose (FrcB, FrcC, FrcA)	+	+	-	-	+
Rhamnose (RhaS, RhaP, RhaQ, RhaT)	+	-	-	-	-
Multiple sugar (ChvE, GguB, GguA)	+	-	+	-	-
Multiple sugar (MsmE, MsmF, MsmG, MsmK)	+	-	-	-	-
Sorbitol / Mannitol (SmoE, SmoF, SmoG, SmoK*)	+	-	-	+	-
alpha-glucoside (AglI, AglF, AglG, AglK*)	+	-	-	-	-
Oligogalacturonide (TogB, TogM, TogN, TogA)	+	-	-	-	-

*AglK is missing in *Hydra*-associated bacterium and SmoK is missing in *Delftia acidovorans*

TABLE S15: Horizontal gene transfer candidates in the *Hydra* genome

Gene model	AI	Best hit description	Putative donor	Cluster ID	Exons	ESTs	SL	Nv	M	NJ supp.	ML supp.	Red. set ML supp.	CL
Hma2.214958	355	UDP-glucose/GDP-mannose dehydrogenase family protein	Algoriphagus sp. PR1	1	1	+			+	+	+	+	high
Hma2.232428	350	UDP-glucose/GDP-mannose dehydrogenase family protein	Pedobacter sp. BAL39	1	1				+	+	+	+	high
Hma2.214959	310	NAD-dependent epimerase/dehydratase family protein	Flavobacteriales	2	1	+		+	+	+	+	+	high
Hma2.233900	291	agmatine deiminase	Bacillus	3	1	+				+	+	+	high
Hma2.210435	290	agmatine deiminase	Bacillus	3	2					+	+	+	high
Hma2.210436	252	agmatine deiminase	Bacillus	3	1	+	+			+	+	+	high
Hma2.218485	106	agmatine deiminase	Bacillus	3	1	+				+	+	+	high
Hma2.223443	243	glycosyl hydrolase, family 5	Clostridium butyricum 5521	4	3	+		+		+	+	+	high
Hma2.211500	242	putative ADP-D-beta-heptose epimerase, hldD	Candidatus Protochlamydia amoebophila UWE25	5	1	+	+	+	+	2	2	+	high
Hma2.208759	233	conserved hypothetical protein	Microscilla marina ATCC 23134	6	10	+				+	+	+	high
Hma2.211670	228	extracellular solute-binding protein	Polynucleobacter necess. subsp. asymbioticus QLW-P1DMWA-1	7	1	+	+			+	+	+	high

Hma2.200124	221	extracellular solute-binding protein	Polynucleobacter necess. subsp. asymbioticus QLW-P1DMWA-1	7	1					+	+	+	high
Hma2.212103	65	extracellular solute-binding protein	Polynucleobacter necess. subsp. asymbioticus QLW-P1DMWA-1	7	1					+	+	+	high
Hma2.225886	213	hypothetical protein FBBAL38_08395	Flavobacterium	8	1	+		+	+	2	2	+	high
Hma2.231042	212	O-methyltransferase, family 2	Algoriphagus sp. PR1	9	1	+		+	+	2	2	+	high
Hma2.230045	164	O-methyltransferase, family 2	Algoriphagus sp. PR1	9	2			+	+	1,2	1,2	+	high
Hma2.214619	130	O-methyltransferase, family 2	Algoriphagus sp. PR1	9	1	+		+	+	1,2	1,2	+	high
Hma2.214618	86	O-methyltransferase, family 2	Algoriphagus sp. PR1	9	1					+	+	+	high
Hma2.205060	178	hypothetical protein FB2170_02100	Bacteroides	10	5	+				+	+	NA	high
Hma2.223079	132	hypothetical protein FB2170_02100	Bacteroides	10	4	+				+	+	NA	high
Hma2.221393	166	hypothetical protein BACUNI_02968	Bacteria	11	5	+		+		+	+	+	high
Hma2.207815	161	D-alanine aminotransferase	Mariprofundus ferrooxydans PV-1	12	1	+				+	+	+	high
Hma2.224009	159	fused heptose 7-phosphate kinase/heptose 1-phosphate adenyltransferase	Akkermansia muciniphila ATCC BAA-835	13	1	+	+	+	+	2	2	+	high
Hma2.228394	148	hypothetical protein BACOVA_04475	Bacteria	14	8					+	+	+	high
Hma2.205478	140	O-Glycosyl hydrolase family 30	Marinomonas sp. MED121	15	2	+				+	+	+	high
Hma2.201736	137	O-Glycosyl hydrolase family 30	Verrucomicrobia	15	2					+	+	+	high
Hma2.211682	128	hypothetical protein EUBVEN_01036	Marinomonas sp. MED121	15	2	+				+	+	+	high

Hma2.211683	123	hypothetical protein EUBVEN_01036	Bacteria	15	2					+	+	+	high
Hma2.205211	70	O-Glycosyl hydrolase family 30	Thermoanaerobacter tengcongensis MB4	15	4	+				+	+	+	high
Hma2.229328	70	O-Glycosyl hydrolase family 30	Bacteria	15	3	+	+			+	+	-	medium
Hma2.226901	138	arsenite S-adenosylmethyltransferase	Bacteria	16	1	+		+	+	2	2	+	high
Hma2.226900	136	hypothetical protein SPV1_03478	Firmicutes	16	1	+		+	+	2	2	+	high
Hma2.215597	135	hypothetical protein M23134_03444	Bacteroidetes	17	1	+	+			+	+	+	high
Hma2.215647	102	hypothetical protein M23134_03444	Microscilla marina ATCC 23134	17	1	+				+	+	NA	high
Hma2.228525	132	hypothetical protein WH7805_01667	Prochlorococcus marinus subsp. pastoris str. CCMP1986	18	1	+	+	+		+	+	+	high
Hma2.223081	126	hypothetical protein FB2170_02100	Bacteroides	19	3	+	+			+	+	NA	high
Hma2.224869	125	hypothetical protein FB2170_02100	Bacteroides	19	1	+	+			+	+	NA	high
Hma2.212555	118	phosphoheptose isomerase	Mycobacterium tuberculosis	20	1					+	+	+	high
Hma2.214152	82	phosphoheptose isomerase	Methanococcales	20	1					+	+	+	high
Hma2.218542	113	hypothetical protein YpsIP31758_2789	Yersinia	21	2	+	+			+	+	+	high
Hma2.233810	110	crystal protein	Bacillus thuringiensis	22	1					+	+	NA	high
Hma2.204168	101	crystal protein	Bacillus thuringiensis	22	1	+				+	+	NA	high
Hma2.215262	37	crystal protein	Bacillus thuringiensis	22	1	+				+	+	NA	high
Hma2.221265	105	hypothetical protein slr1628	Thermosynechococcus elongatus BP-1	23	1	+				+	2	+	high
Hma2.214160	91	hypothetical protein MC7420_1971	Lentisphaera araneosa HTCC2155	24	4	+	+			+	+	+	high

Hma2.226010	90	nitric oxide dioxygenase	Lentisphaera araneosa HTCC2155	25	6	+	+		+	+	+	+	high
Hma2.221201	87	hypothetical protein Sma0592	Bacteria	26	2					+	+	+	high
Hma2.231356	84	saxitoxin N1-hydroxylase	Bacillus	27	6			+		+	+	NA	high
Hma2.222682	80	phytanoyl-CoA dioxygenase family protein	Alphaproteobacteria	28	5	+		+	+	1	1	+	high
Hma2.222754	62	Pirin-related protein	Burkholderia	29	4	+		+	+	1	1	+	high
Cont27899:2031-2249	46	hypothetical protein HH0003	Actinomycetales	30	1	+	+	+		+	+	+	high
Hma2.220724	43	hypothetical protein HH0003	Bacteroides	30	4			+		+	+	-	medium
Hma2.215873	36	sensory box histidine kinase/response regulator	Lentisphaera araneosa HTCC2155	31	2					+	+	+	high
Hma2.208802	373	COG1292: Choline-glycine betaine transporter (ISS)	Psychrobacter	32	9	+		+		1	1	-	medium
Hma2.219075	268	BCCT transporter (ISS)	Psychrobacter	32	5	+	+	+		+	+	-	medium
Hma2.222799	201	COG1292: Choline-glycine betaine transporter (ISS)	Psychrobacter	32	4			+		+	+	-	medium
Hma2.222800	89	COG1292: Choline-glycine betaine transporter (ISS)	Bacteria	32	1	+		+		+	+	-	medium
Hma2.203040	331	hypothetical protein VAS14_12879	Psychrobacter	33	2	+				+	+	-	medium
Hma2.230779	329	polysaccharide deacetylase family protein	Acidithiobacillus ferrooxidans ATCC 23270	34	2	+	+	+		+	+	-	medium
Hma2.206073	170	GCN5-related N-acetyltransferase	Bacteria	35	1	+				+	+	-	medium
Hma2.222145	125	maltose O-acetyltransferase	Bacteria	36	1	+				+	2	-	medium

Hma2.213745	122	hypothetical protein PM8797T_26465	Flavobacteriales	37	2	+	+			+	+	-	medium
Hma2.203165	109	hypothetical protein BACCAC_01538	Bacteroides	38	2	+		+		+	+	-	medium
Hma2.213996	107	YD repeat-containing protein	Bacteria	39	4				+	1	1	NA	medium
Hma2.214592	79	YD repeat-containing protein	Bacteria	39	6				+	1	1	NA	medium
Hma2.221669	75	anhydro-N-acetylmuramyl-tripeptide amidase	Rickettsia	40	1	+			+	+	+	-	medium
Hma2.233203	60	putative lipopolysaccharide biosynthesis protein	Parabacteroides distasonis ATCC 8503	41	1	+		+		+	+	-	medium
Hma2.222272	42	metallo-beta-lactamase superfamily protein	Bacteria	42	1	+			+	2	2	-	medium
Hma2.206652	40	RelA/SpoT domain-containing protein	Clostridium botulinum E3 str. Alaska E43	43	1	+				+	+	-	medium
Hma2.209415	37	Ion transport 2 domain protein	Bacteria	44	9			+	+	1	1	-	medium
Hma2.229612	36	Bacterial extracellular solute-binding protein, family 3	Bacteria	44	12	+		+	+	1	1	-	medium

Abbreviations:

Gene Model: Hma numbers refer to gene models in the *Hydra* genome browser (http://hydrazome.metazome.net)

AI: Alien Index (Ref. 44)

Best hit description: best Blastp hit in the NCBI nr database (excluding *Hydra* and *Nematostella* gene models)

Putative donor: nearest bacterial neighbor(s) in ML phylogenetic tree

Cluster ID: putative HGT candidates were clustered to identify paralogs

Exons: exon number in Hma gene model

EST: + indicates presence of one or more ESTs with >98% identity to the genome sequence

SL: + indicates one or more ESTs with a spliced leader sequence

Nv: + indicates that a homolog is present in the *Nematostella vectensis* genome

M: + indicates metazoan hits (except *Nematostella*) in the MJ or ML phylogenetic trees

NJ: + indicates support for HGT in neighbor joining phylogenetic tree

ML: + indicates support for HGT in maximum likelihood phylogenetic tree

In the NJ and ML trees that contain metazoan hits, the HGT prediction was further supported based on a different exon/intron structure compared to the metazoan hit (1) or based on a different exon number compared to the metazoan hit (2)

Red. set ML: + indicates support for HGT in maximum likelihood phylogenetic tree based on a reduced prokaryote set (only the species with the biggest genome per genus); - indicates no clear support for HGT; NA indicates that no tree is available

CL: represents an estimate for the confidence level of the HGT candidate; high indicates that all three phylogenetic trees support HGT or that the two available trees support HGT and there are no metazoan homologs; medium indicates that only one or two phylogenetic trees show good support for HGT or that there are only two trees available and there are metazoan homologs

TABLE S16: microRNAs from *Hydra*

miRBase ID	Location	Mature	miR* support	Read support	RNAfold dG, ~70 bp	Northern
hma-miR-3002	Contig35599:5316..5415	TAGAAGATAAGATTGTGGCCAG	no	1	-21.9	
hma-miR-3003a	Contig35197:56773..56872	TAAATGTGCAAACTGGTAATGA	yes	3	-31.9	
hma-miR-3003b	paralogue of hma-miR-3003a	TTAAATGTGCGAACTGGTAATG	yes	?	-39.1	Y
hma-miR-3004	Contig37346:30501..30600	AACTATTTGTTTTGGTCTTGG	no	2	-22.9	
hma-miR-3005	Contig38041:108152..108251	TTGGTGGATGATTGATTCCTAA	no	13	-40.8	Y
hma-miR-3006	Contig38272:140947..141046	TGATTTATAATAGGAGAGCATATGATT	no	3	-22.8	
hma-miR-3007	Contig38799:92550..92649	TGTTCGGCGTTTTTACCTGCA	yes	2	-46.3	
hma-miR-3008	Contig38964:135673..135772	TCAGCTGTAGTTGACGACAAAT	no	1	-39.4	
hma-miR-2022	Contig38718:107549..107648	TTTGCTAGTTGCTTTTGTCCCC	no	7	-31.9	Y
hma-miR-3009	Contig38968:76227..76326	TTTCATCAAAGTCACTAATAA	no	3	-24.4	Y
hma-miR-3010	Contig36420:11485..11584	TGAGAGCGCAAAAATAAATTTG	no	2	-33.0	Y
hma-miR-3011	Contig39235:74153..74252	TTGATTTTATGGAATGATTTAC	no	1	-29.7	
hma-miR-3012a	Contig37241:32552..32651	TCAGTGGATTTGTAGAAAGCTTT	no	2	-30.2	
hma-miR-3012b	Contig39307:331986..332085	TCAGTGGATTTGTAGAAAGTTTTA	no	2	-26.2	

hma-miR-3013	Contig32576:29320..29419	TGGTCTGACTTCGCTATAGAGAATG	no	2	-21.0	
hma-miR-3014	Contig32768:28952..29051	CCAACAGACACAAAACCGACG	no	1	-24.5	
hma-miR-2030	Contig39224:149555..149654	CGTCTTACACTGCTGTGCTAC	no	1	-33.6	

TABLE S17: Correlation between homeobox gene loss and life cycle stage loss in *Hydra*

Species	Class	Larval Stage Present?	Polyp Type	Medusa Stage Present?	Eve Gene Present?	Emx Gene Present?	Reference
Hydra magnipapillata	Hydrozoa	no	solitary	no	no	no	Chourrout et al., 2006[133]
Sarsia sp.	Hydrozoa	yes	colonial	yes	yes	N.D.	Bridge and Steele, unpublished, GenBank Accession Number AF326771
Clytia hemisphaerica	Hydrozoa	yes	colonial	yes	yes	N.D.	Chiori et al., 2009[134]
Hydractinia echinata	Hydrozoa	yes	colonial	no	N.D.	yes	Mokady et al., 1998[135]
Podocoryne carnea	Hydrozoa	yes	colonial	yes	N.D.	yes	Bridge and Martinez, unpublished
Acropora millepora	Anthozoa	yes	colonial	no	yes	yes	Miller and Miles, 1993[136]; de Jong et al., 2006[137]
Nematostella vectensis	Anthozoa	yes	solitary	no	yes	yes	Finnerty and Martindale, 1997[138]; Chourrout et al., 2006[133]

N.D., Not determined

TABLE S18: Cnidarian orthologs of signaling and transcription factors known to act in the Spemann-Mangold organizer in *Xenopus*

	Hydra	*Hydra* head organizer expression	*Nematostella*	*Nematostella* blastopore organizer expression
Secreted Signaling Factors				
Wnt	**HyWnt1, -2, -3, -5, -7, -8, 9/10a, -9/10b, -9/10c, -11, and -16**	yes	**NvWnt1, -2, -3, -5, -7a, -7b, -8a, 8b, 10, -11, and -16**[92,139]	yes
Chordin	**HyChordin**[140]	yes*	**NvChordin**[141,142,143]	yes
Noggin	**HyNoggin** *Hma2.221925*	n.d.	**NvNoggin1, -2**[141,142]	no
Follistatin	**HyFollistatin-like**[144]	n.d.	**NvFollistatin, NvFollistatin-like**[141,142]	no
Admp	**HyBmp5-8b**[145]	no	**NvBmp2/4 and NvBmp5-8**[141,142,143,146]	yes
Dkk1	**HyDkk1/2/4**[147]	yes*	**NvDkk1/2/4**[139]	no
Cerberus	**HyCerberus-like** *XP_00215421/Hma2.208432* *Holstein et al., unpublished*	yes	no	---
Frzb1/Frzb2/Crescent	**HysFRP1/2/5** *Hma2.22298* *Hobmayer et al., unpublished*	yes*	**NvsFRP1/2/5** *XP_001638620* *Hobmayer et al., unpublished*	no
sFRP	**HysFRP3/4** *XP_002162073/Hma2.229127* *Hobmayer et al., unpublished*	yes*	**NvFrizzled-related proteins 1, 2, 3** *XP_001638660* *XP_001624600* *XP_001623764*	n.d.
Transcription Factors				
goosecoid	**Cngsc**[148]	yes	**NvGsc**[141]	no
forkhead	**Budhead**[149]	yes	**NvFoxA, -B**[150,151,152]	yes
otx-2	**CnOtx**[153]	yes*	**NvOtxA, -B, -C**[154]	yes*
bra	**Hybra1, -2**[155,156]	yes	**NvBra1**[157]	yes
lim-1	**Hm lim-1**	n.d.	**NvLhx1**[120,158]	yes

	XP_002155267			
irx-3	**Hm irx-related** *XP_002166749*	n.d.	**Nvlrx**[72]	n.d.
Ldb	**HyLdb** *XP_002161782/Hma2.217495*	n.d.	**NvLdb**[158]	n.d.
not-2	**HyNot-2** *Hma2.206051*	n.d.	**NvNot-like B**, **-C**, **-D**, **-E**, **-F**[72]	n.d.
siamois	no	---	no	---
anf-1	no	---	no	---
twn	no	---	no	---

*transient expression in the developing organizer; n.d.: not determined

TABLE S19: Inventories of genes encoding structural and regulatory muscle proteins in *Hydra* and *Nematostella*

Integrin-binding costamere proteins

	vertebrates	*Nematostella*	*Hydra*
Talin	present	present	present
melusin/chord/integrin-β1 binding protein	present	present	present
Vinculin	present	present	present

Dystroglycan-associated costamere proteins

	vertebrates	*Nematostella*	*Hydra*
Dystrophin	present	present	present
α/β-syntrophin	present	present	present
α/β-Dystroglycan	present	present	/
α/ε-Sarcoglycan	present	present	/
β-Sarcoglycan	present	present	/
δ/γ/ζ-Sarcoglycan	present	present	present
α/β-Dystrobrevin	present	present	present
γ-syntrophin	present	present	/

Motor and regulatory proteins

	vertebrates	*Nematostella*	*Hydra*
Myosin heavy chain type II	present	present (2)	present (2)
Myosin essential light chain	present	present (8)	present (2)
Myosin regulatory light chain	present	present (5)	present (1)
Tropomyosin-like	present	present (14)	Present (4)
Troponin T	present	/	/
Troponin I	present	/	/
Troponin C	present	/	/
Calmodulin	present	present	present
Caldesmon	present	/	/
Calponin	present	transgelin/calponin	transgelin/calponin

TABLE S20: Neuromuscular Junction Proteins in *Hydra*

Function	Diagnostic Protein	Present/Absent in *Hydra*?
Uptake of choline by neurons	choline transporter (ChT)	Two putative ChTs are present, one expressed in nerve cells
Acetylation of choline to produce acetylcholine	choline acetyltransferase (ChAT)	ChAT/CrAT family acetyltransferases present in both *Hydra* and *Nematostella*, but with catalytic domain residues that indicate primitive carnitine acetyltransferase activity
Uptake of acetylacholine into vesicles	acetylcholine transporter (VAChT)	Not found in *Hydra* or *Nematostella*. In bilaterians, this gene is located in the first intron of the ChAT gene, sharing exon 1.
ligand-gated ion channel	nicotinic acetylcholine receptor (nAChR)	Six subunits encoded in the *Hydra* genome; one expressed in ectodermal epithelial cells of presumptive tentacle region of buds and in the tentacles of adult polyps
Clustering of receptors at neuromuscular junction	agrin (extracellular matrix protein) rapsyn (tetricopeptide repeat-containing cytoplasmic protein), MuSK (muscle-specific receptor protein-tyrosine kinase)	Agrin and rapsyn are absent from the *Hydra* genome. While no MuSK gene is found, genes encoding homologs of the related ROR and DDR receptor protein-tyrosine kinases are present.
Recovery by degradation of acetylcholine	acetylcholinesterase (AChE)	Divergent AChE domain present as part of larger protein

TABLE S21: Accession numbers and/or contig positions for the junction proteins shown in Figure 4.

predicted protein	accession/EST number
Hydra magnipapillata	
hm classic cadherin	contig 37169
hm p120/delta-catenin	contig 38288
hm alpha-catenin	XP_002156697
hm vinculin	XP_002156269
hm alpha-actinin	contig 37562 (35228-60384)
hm afadin	XP_002161471
hm claudin	XP_002154363/contig 38160 (36222-48040)
hm ZO-1	XP_002164770
hm stardust/PALS1	XP_002165318
hm PATJ	XP_002167180/contig 37321 (67997-108034)
hm PAR3	XP_002165602
hm PAR6	contig 36553 (448-4206)
hm neurexin	XP_002157821
hm neuroglian	XP_002165052
hm contactin	XP_002154651
hm gliotactin	XP_002162717 / contig 35977 (56713-54802)
hm discs large	XP_002168436 / contig 38190 (68002-37333)
hm scribble	XP_002165472
hm coracle	XP_002160141 / contig 39288 (115680-112603)
hm innexin1	XP_002154796
hm innexin1A	XP_002160488
hm innexin2	XP_002160488
hm innexin3	XP_002155819
hm innexin4	XP_002166931

hm innexin5	XP_002164746
hm innexin6	XP_002170247
hm innexin7	XP_002170241
hm innexin8	XP_002170409
hm innexin9	contig 37896 (84059-85535)
hm innexin9A	contig 37896 (83310-81793)
hm innexin10	XP_002166542
hm innexin11	XP_002165350
hm innexin12	XP_002156556
hm innexin13	XP_002155033
hm innexin14	XP_002166566
hm innexin15	XP_002160898
hm integrin-alpha1	XP_002161020
hm integrin-alpha2	XP_002159658/XP_002161969
hm integrin-alpha3	XP_002163512/XP_002159522
hm integrin-alpha4	contig 31383 (5189-22897)
hm integrin-beta1	XP_002159375/XP_002164638
hm integrin-beta2	contig 35322 (49962-16597)
hm integrin-beta3	contig 38319 (65309-86575)
hm talin	XP_002154525
hm paxillin	XP_002168161
hm tensin	XP_002156645
hm ILK	XP_002158769
hm FAK	AAW21807/contig 39326 (9312-19136)

Clytia hemispherica

ch innexin1	CU442601/CU435687/CU439715/CU436674
ch innexin2	CU430212/CU442174/CU442627
ch innexin3	AM750354

ch innexin4	CU432478
ch innexin5	CU440840/CU426883/CU432586
ch innexin6	CU432056
ch innexin7	AM755630/CU425525
ch innexin8	CU428875

Nematostella vectensis

nv alpha-actinin	XP_001633290
nv afadin	XP_001624542
nv claudin	XP_001642058
nv ZO-1	XP_001633912
nv stardust/PALS	XP_001625677
nv PATJ	XP_001625806
nv PAR3	XP_001637950
nv PAR6	XP_001627416
nv neurexin	XP_001637897
nv neuroglian	XP_001637406
nv contactin	Nemve1/scaffold_17:486517-498167
nv gliotactin	XP_001629673
nv discs large	XP_001638123
nv scribble	XP_001625532
nv coracle	XP_001622324
nv innexin	XM_001623849
nv tensin	Nemve1/scaffold_26:113107-118134
nv ILK	Nemve1/scaffold_2:491819-498567

Trichoplax adhaerens

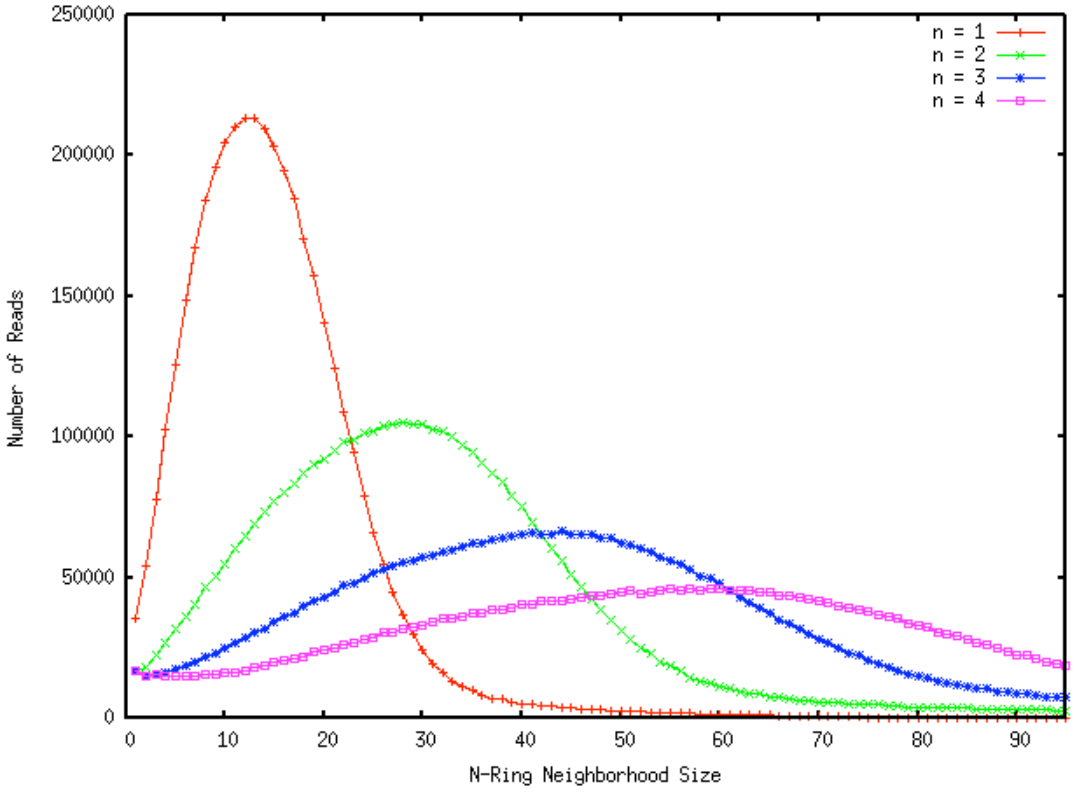
ta alpha-actinin	XP_002113237
ta afadin	XP_002112781
ta ZO-1	XP_002116319
ta stardust/PALS	XP_002109073
ta PATJ	XP_002113893
ta PAR3	XP_002110754
ta PAR6	XP_002108295
ta neurexin	XP_002116615
ta neuroglian	XP_002117985
ta contactin	XP_002117979
ta gliotactin	XP_002117140
ta discs large	XP_002108800
ta scribble	XP_002114951
ta coracle	XP_002114293
ta tensin	XP_002110466
ta ILK	XP_002116044

Monosiga brevicollis

mb vinculin	EDQ87727
mb alpha-actinin	XP_001746581
mb discs large	XP_001743865
mb talin	XP_001743667
mb paxillin	XP_001742769

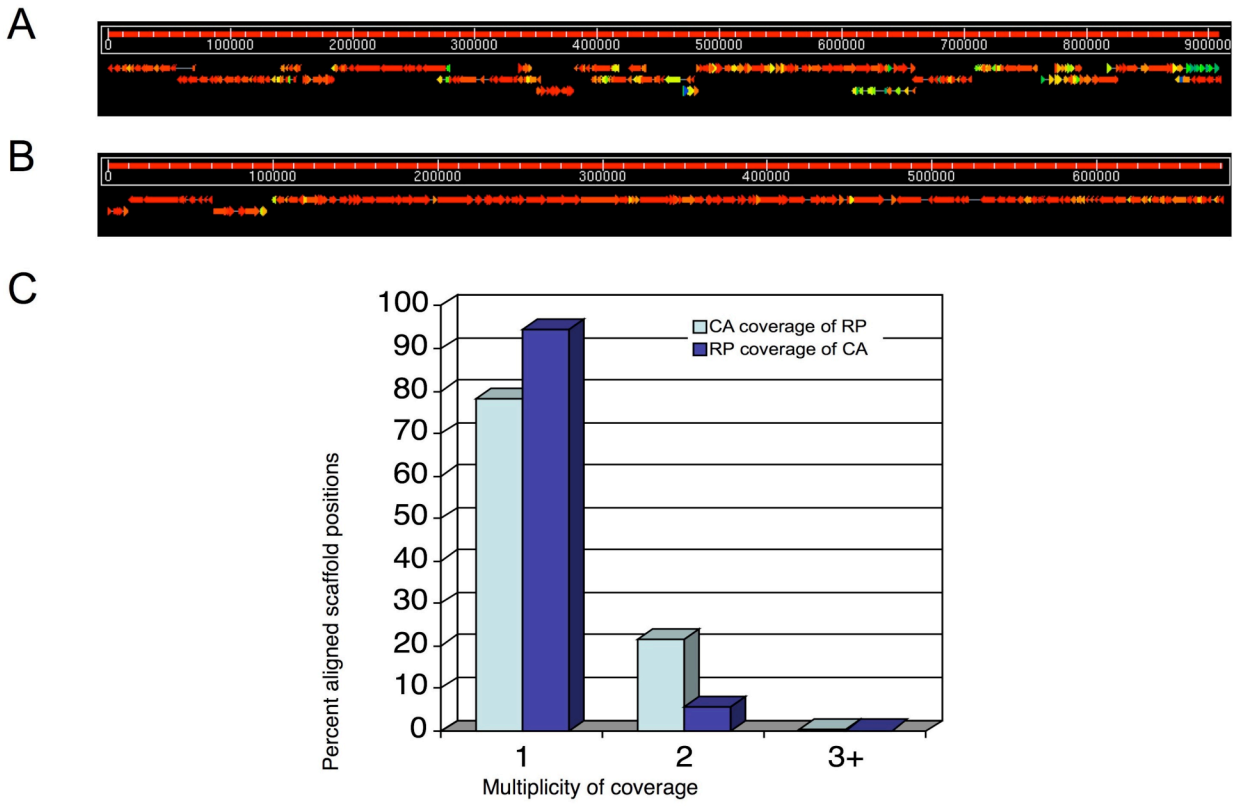
Contig numbers are from the *Hydra* genome browser and the *Nematostella* genome browser at JGI.

FIGURE S1: Alignment Neighborhood Size Distributions



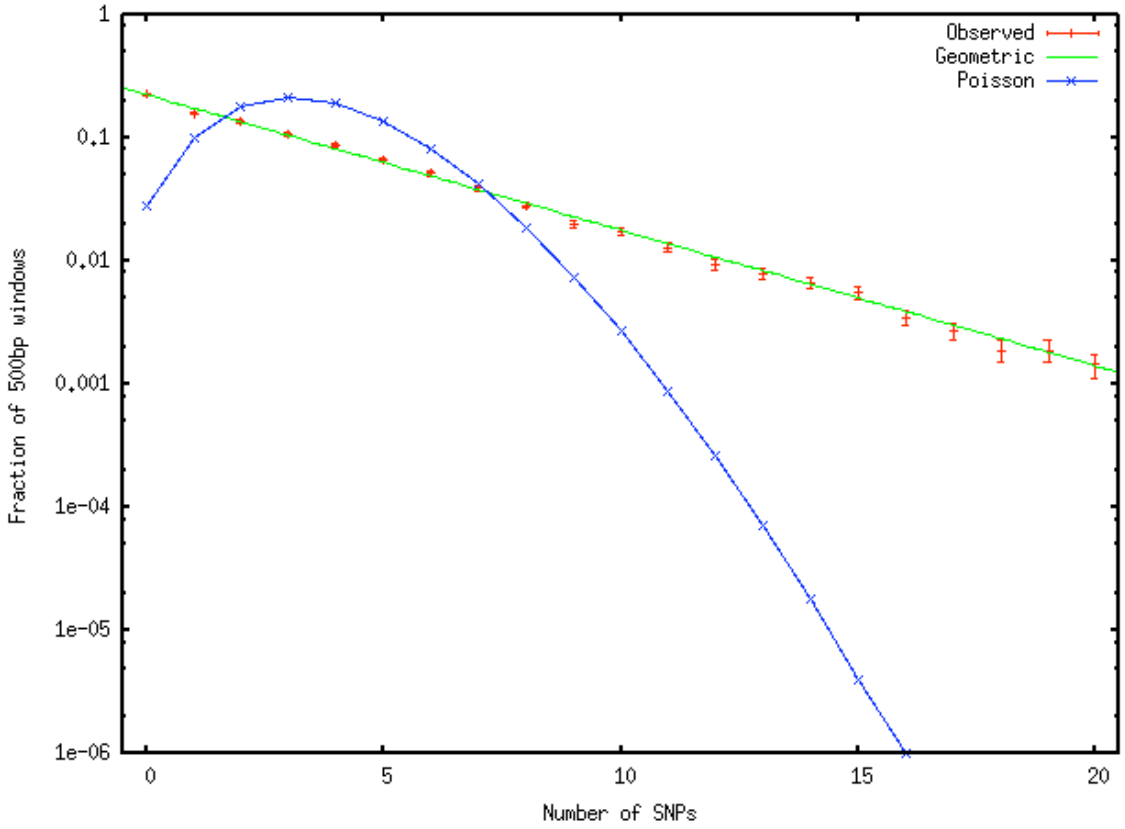
The distribution of the number of N-degree neighbors (N=1-4) is shown. Alignments between reads with more than 45 second-degree neighbors are excluded from clusters.

FIGURE S2: Inter-Assembly Alignments



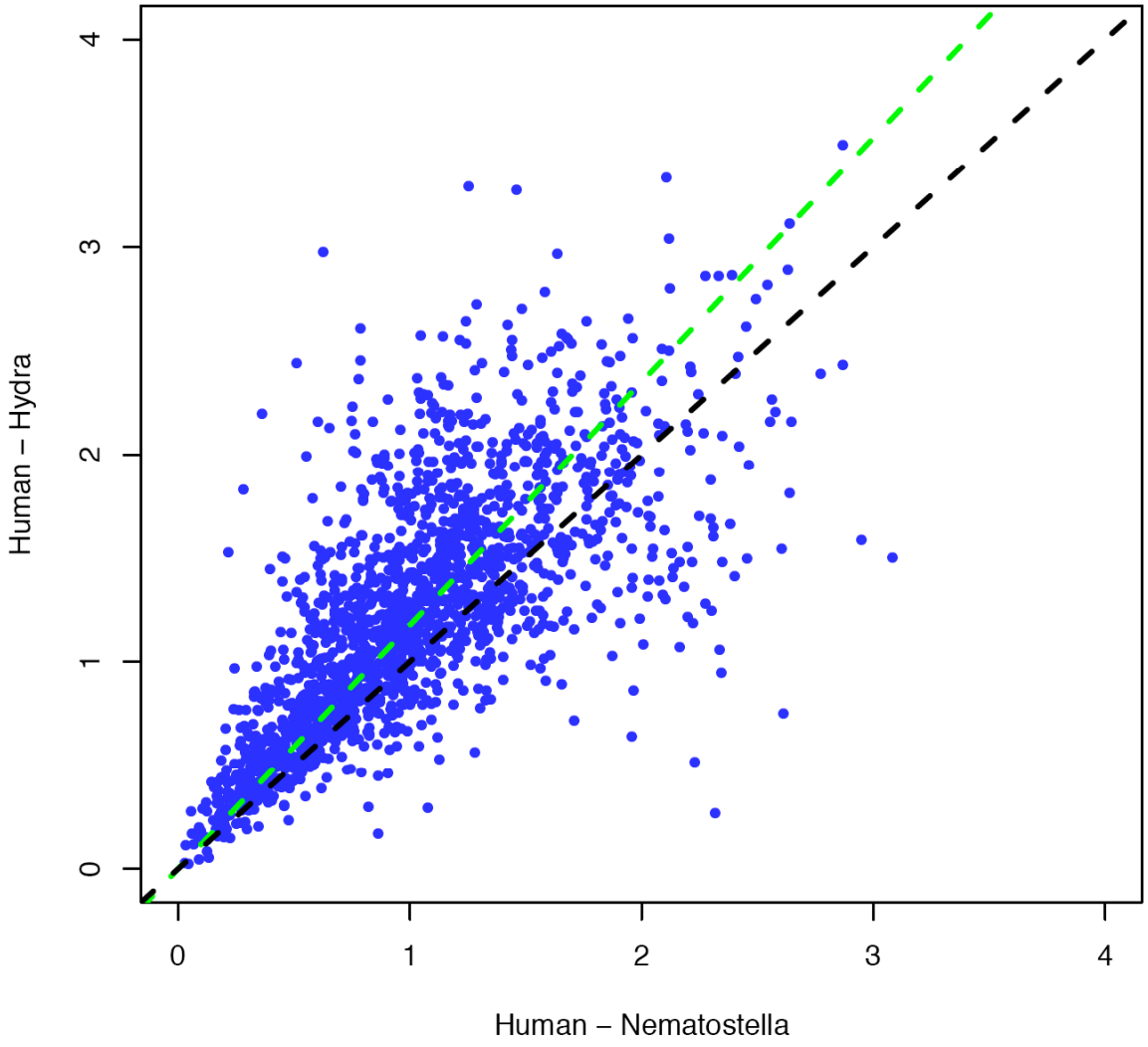
(A) Alignments of CA scaffolds to the largest RP scaffold. (B) Alignments of RP scaffolds to the largest CA scaffold. High-scoring segment pairs of at least 5 kb are shown colored by percent identity. (red, >99.9; orange, 98%; green, 96%; blue, 94%). (C) Distribution of inter-assembly scaffold alignment coverage for all scaffolds larger than 100 kb. Over 400 Mb of aligned scaffold sequence 22% of RP is doubly covered by CA, and 5.6% of CA is doubly covered by RP.

FIGURE S3: Distribution of SNPs in the High-Depth Fraction of the RP Assembly



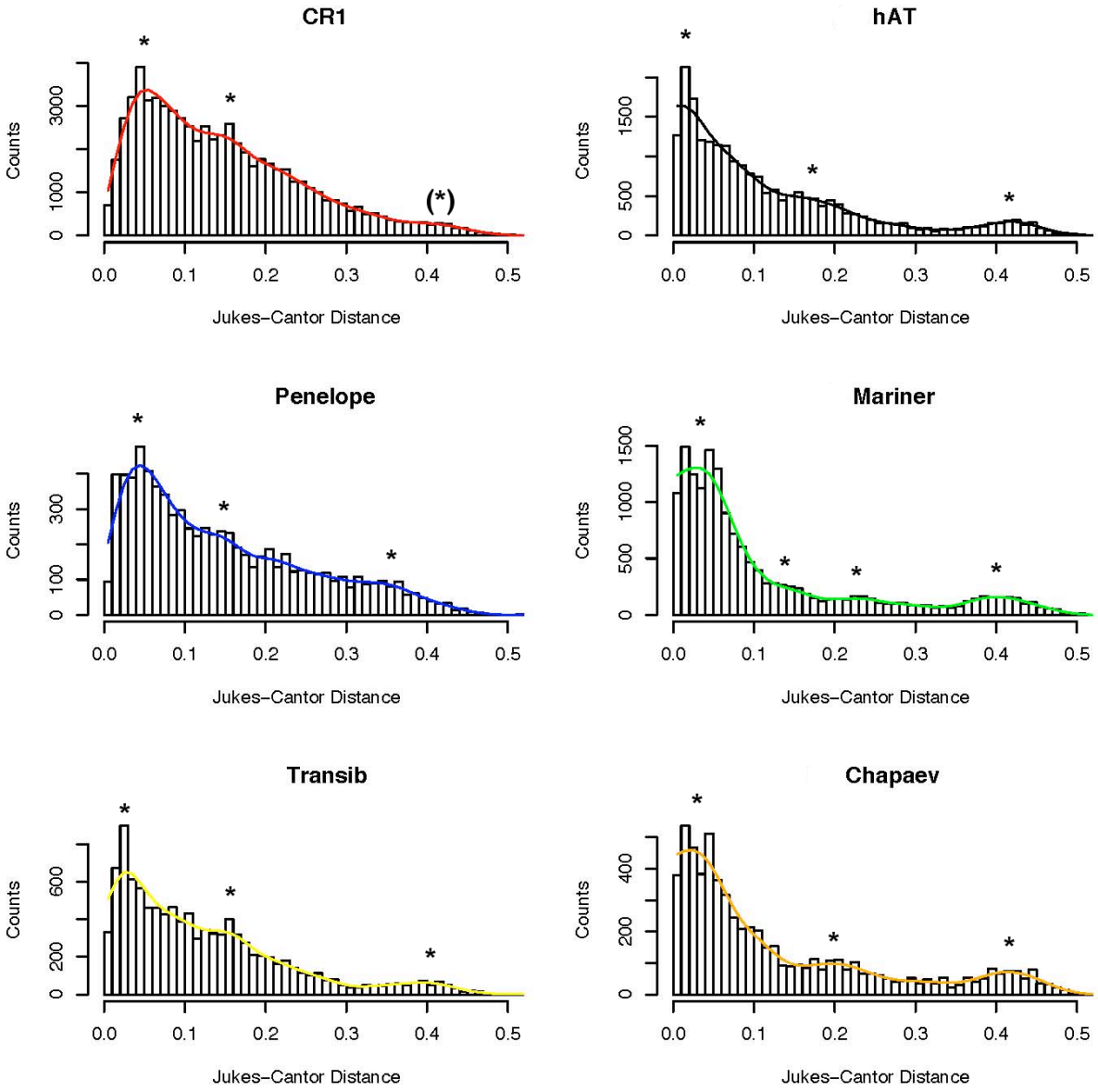
The observed distribution of SNPs per 500 bp window of scaffold sequence (red points) with mean depth (8-10x) is well-fit by the geometric distribution of $P*(1-P)^n$ with P=0.23 (green line). This corresponds to a mean SNP rate of 3.45 SNPs per 500 bp. A Poisson distribution with the same mean value is shown in blue.

FIGURE S4: Protein Distance Graph for *Nematostella* vs. Human and *Hydra* vs. Human.



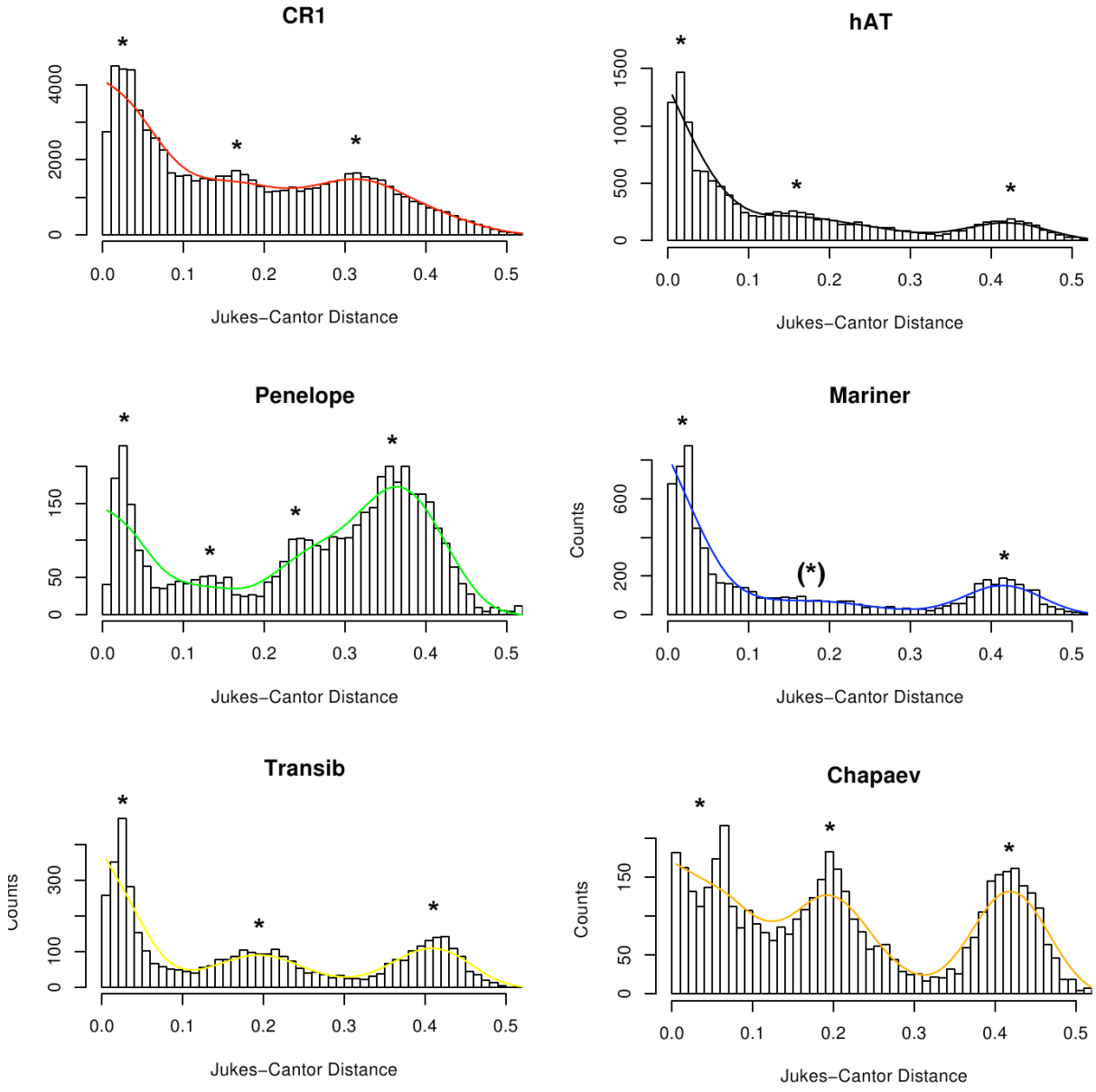
Proteins from *Hydra* are on average more distant from their human orthologs than *Nematostella* proteins are from their human orthologs. Green line: linear model fit to the data; black line: line of slope 1. Theoretical gene families were built using a mutual-best hit clustering algorithm based on BLAST scores. Pairwise distances between orthologs in each gene family were computed with the PROTDIST program from Phylip. Each dot on the graph represents a gene family where distances between orthologous proteins (*Nematostella* to human and *Hydra* to human) were plotted against each other. The axes show PROTDIST units derived from the Dayhoff PAM matrix.

FIGURE S5: Periods of Repeat Expansion in the *Hydra* Genome



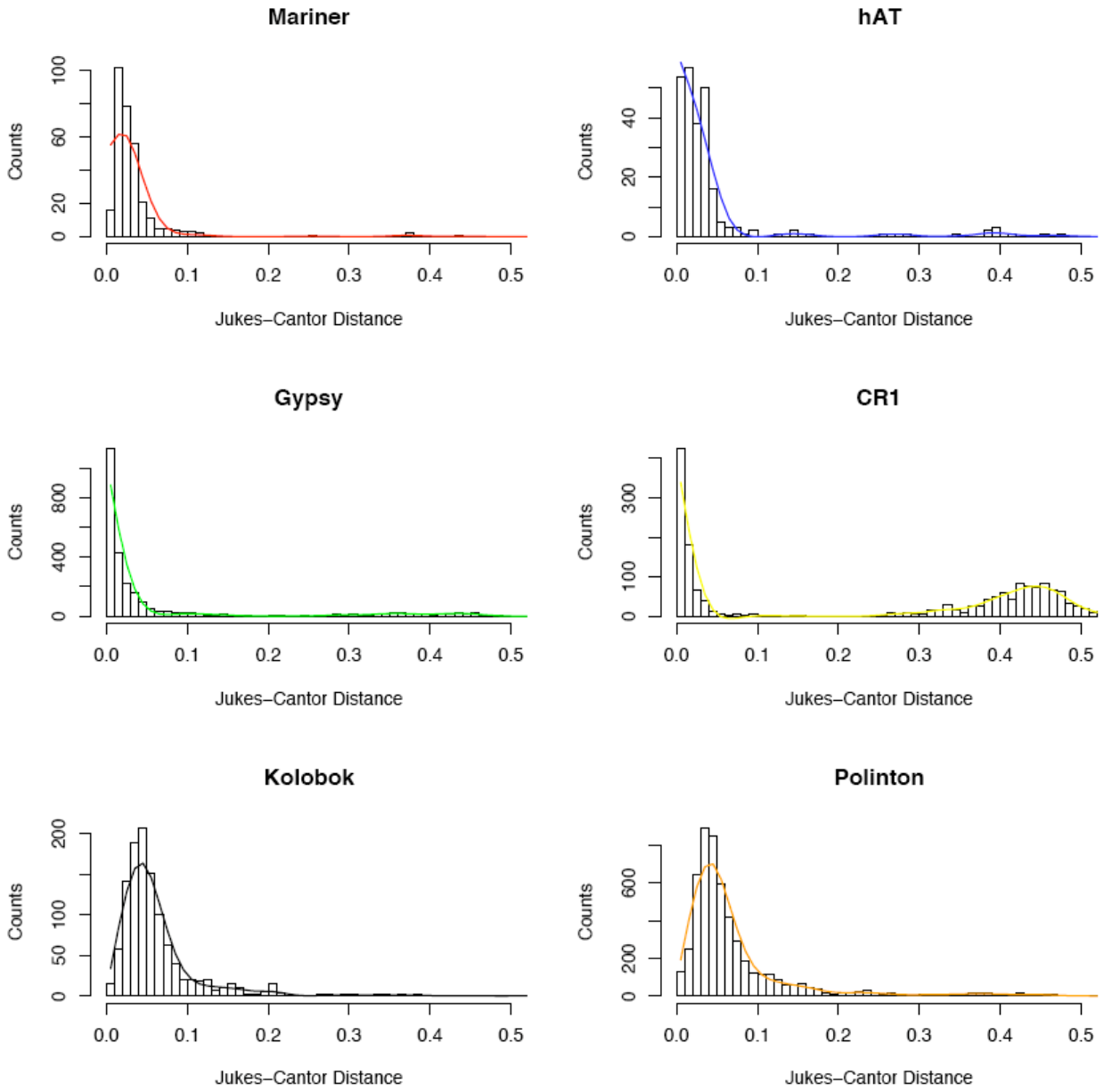
Presence of ancient repeat copies and periodic expansions of the six most abundant repeat families in *Hydra* using the ReAS repeat library for genome masking. Most elements show an increase of activity during periods around Jukes-Cantor distances 0.05, 0.15-0.25, 0.35-0.45.

FIGURE S6: Periods of Repeat Expansion in the *Hydra* Genome Identified Using the RepBase Library



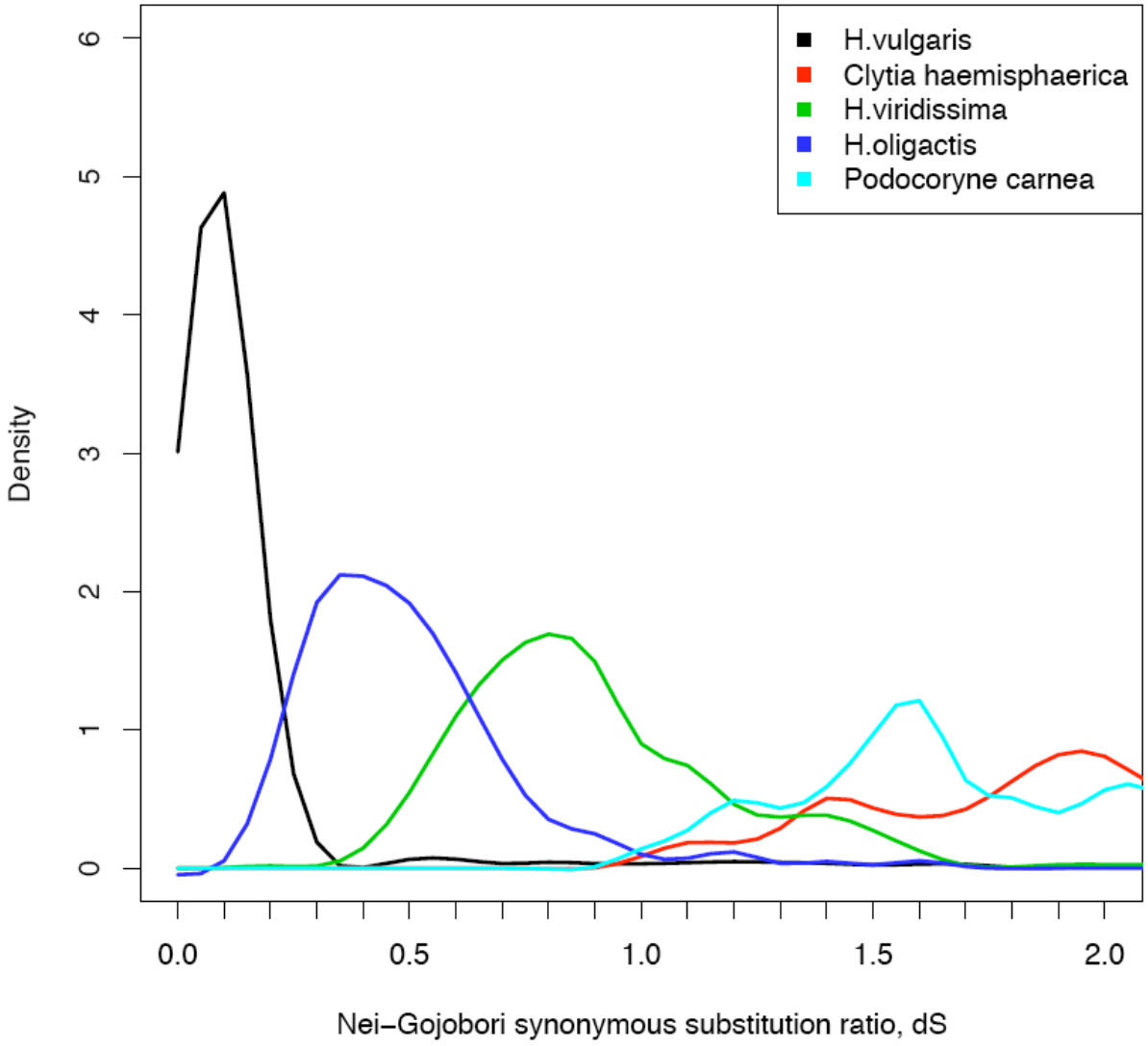
Presence of ancient repeat copies and periodic expansions of the six most abundant repeat families in *Hydra* using RepBase 14.01[159] as the repeat library for genome masking. Distinct peaks of activity can be observed for most of the elements for periods around Jukes-Cantor distances 0.05, 0.15-0.25, 0.35-0.45. Artificial enhancement of the most ancient peak (JC 0.4) is probably due to a not yet complete RepBase library for *Hydra*.

FIGURE S7: The *Nematostella* Genome Lacks Ancient Transposon Copies



The six most abundant *Nematostella* transposon classes are shown. With the exception of CR1, the most abundant repeat families in *Nematostella* are only 10% diverged from their consensus sequences. Genome masking was done with RepBase 14.01.

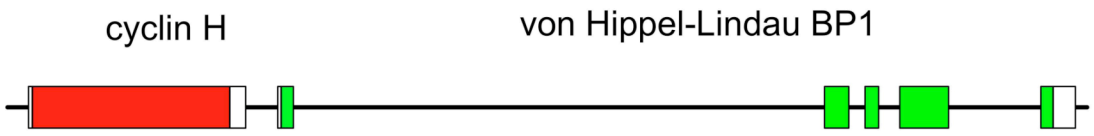
FIGURE S8: Sequence Divergence among Hydrozoans Based on Synonymous Substitutions in ESTs



Synonymous substitution rates in the ESTs of several hydrozoans relative to *H. magnipapillata*. ESTs were clustered with a simple mutual best hit algorithm using BLAST scores. PAML[160] program was used to compute Nei-Gojobori substitution ratios (substitutions per site). Saturation is reached after dS = 1.

FIGURE S9: Example of a *Hydra* Operon

A

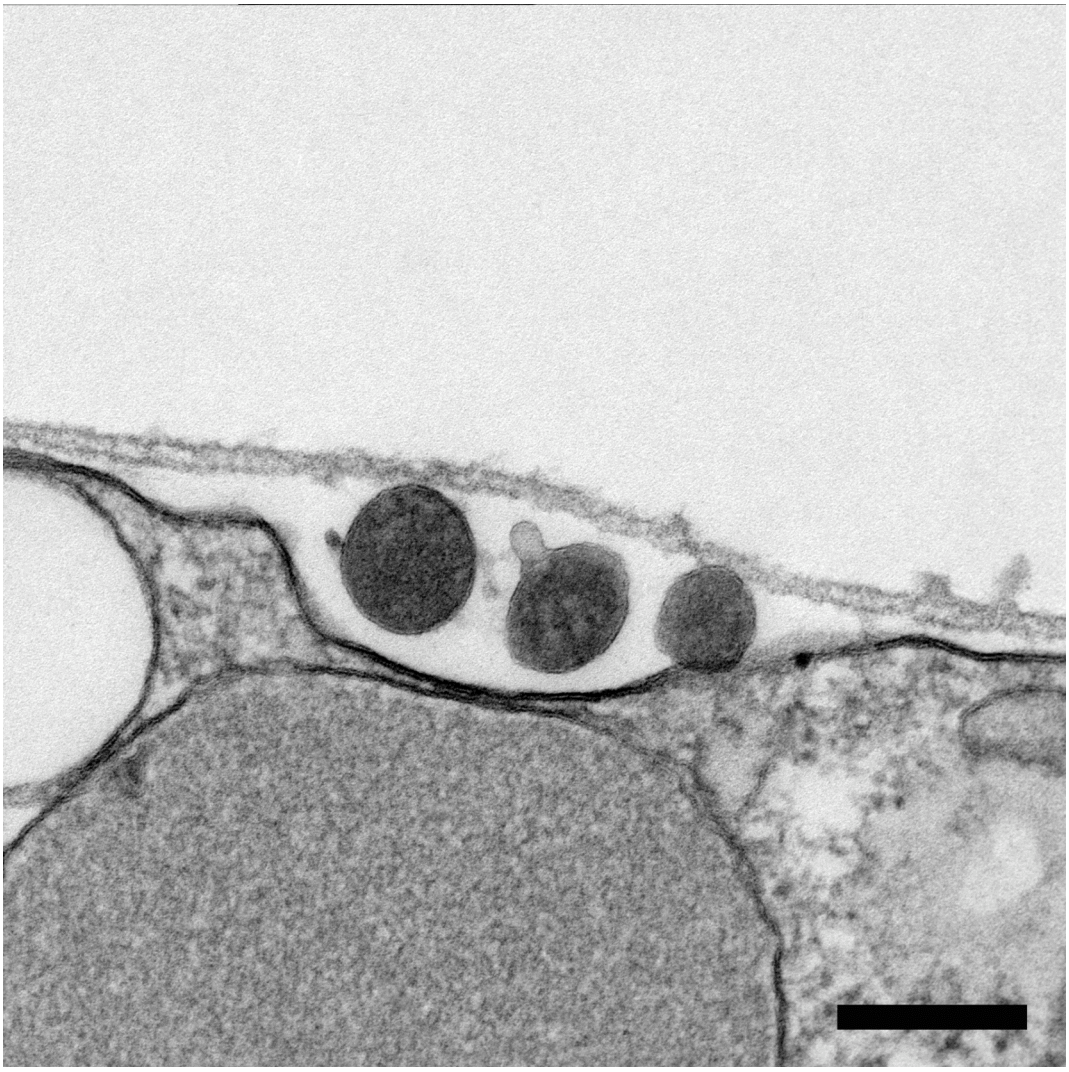


B

```
Hydra_cyclin_H      MFHTSSQRKHWIFESIDKINSNRAASNNNF---IEMHRKLKPNAKIEYLNPNEEKLLLTY 57
Human_cyclin_H      MYHNSSQKRHWTFSSEEQLARLRADANRKFRCKAVANGKVLPND-PVFLEPHEEMTLCKY 59

Hydra_cyclin_H      YTQFVFNICRRFKPVPLSLIGTSLSYFKRFYLYTSVMEFHPKDIAYLCVYLACKIDEYN 117
Human_cyclin_H      YEKRLLEFCSVFKPAMPRSVVGTACMYFKRFYLNNSVMEYHPRIIMLTCAFLACKVDEFN 119

Hydra_cyclin_H      VSIDQFMEQAVVKRS--LNMQKFLIDNELVLLQKLNYHLTVHSPYRPLEGFLIDVKTKKS 175
Human_cyclin_H      VSSPQFVGNLRESPLGQEKALEQILEYELLLIQQLNFHLIVHNPYRPFEGFLIDLKTRYP 179

Hydra_cyclin_H      IP-DIEKHRTNIEQFLTSSLLTDVILLFTPSQLALAAIENGVGREQLIGYLQLTLDTDSI 234
Human_cyclin_H      ILENPEILRKTADDFLNRIALTDAYLLYTPSQIALTAILSSASRAGITMESYLSESLMLK 239

Hydra_cyclin_H      EKVLCKLKRIQEIVCSIKVADQTEIFAIEDKLKLCKDYENDPFSDLYHQREISRQEAKEM 294
Human_cyclin_H      ENRTCLSQLLDIMKSMRNLVKKYEPPRSEEVAVLKQKLERCHSAELALNVITKKRKGYED 299

Hydra_cyclin_H      QKRKKFQELYDIKHAEEKMLAESLD 319
Human_cyclin_H      DDYVSKKSKHEEEEWTDDDLVESL  323


Hydra_von_Hippel-Lindau_BP                MTETPLVKKKHRGIPEALFLNDVDAYMVK--ESSSETAL 37
Human_von_Hippel-Lindau_BP      MAAVKDSCGKGEMATGNGRRLHLGIPEAVFVEDVDSFMKQPGNETADTVL 50

Hydra_von_Hippel-Lindau_BP      QKLDEQFQKYRFMESNLLNKKIRLSTQIPDIKATLSSINFLKNKKNEKEP 87
Human_von_Hippel-Lindau_BP      KKLDEQYQKYKFMELNLAQKKRRLKGQIPEIKQTLEILKYMQKKKESTNS 100

Hydra_von_Hippel-Lindau_BP      LKTQFMLSDQLFVHAKVPTTDKVCLWLGANVMLEYNIDEADELLKKNLSA 137
Human_von_Hippel-Lindau_BP      METRFLLADNLYCKASVPPTDKMCLWLGANVMLEYDIDEAQALLEKNLST 150

Hydra_von_Hippel-Lindau_BP      AESQLLELDNDLDYLRDQITTTEVSMARIYNWDVKRRQKLKISS    181
Human_von_Hippel-Lindau_BP      ATKNLDSLEEDLDFLRDQFTTTEVNMARVYNWDVKRRNKDDSTKNKA 197
```

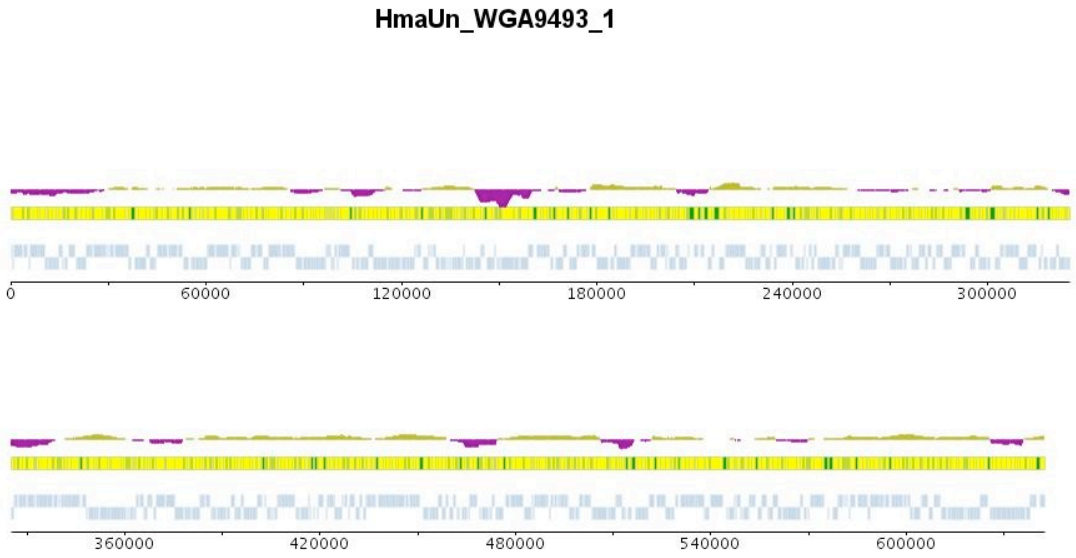
Panel A shows the arrangement of genes in a two-gene operon. Exons are in red (cyclin H gene, gene model Hma2.231564) and green (von Hippel-Lindau Binding Protein 1 gene, gene model Hma2.231562). Untranslated regions are in white. Panel B shows alignments of the predicted protein sequences for the two genes with their human orthologs. Identical amino acids are highlighted in yellow.

FIGURE S10: Electron Micrograph of Bacteria Under the *Hydra* Glycocalyx.

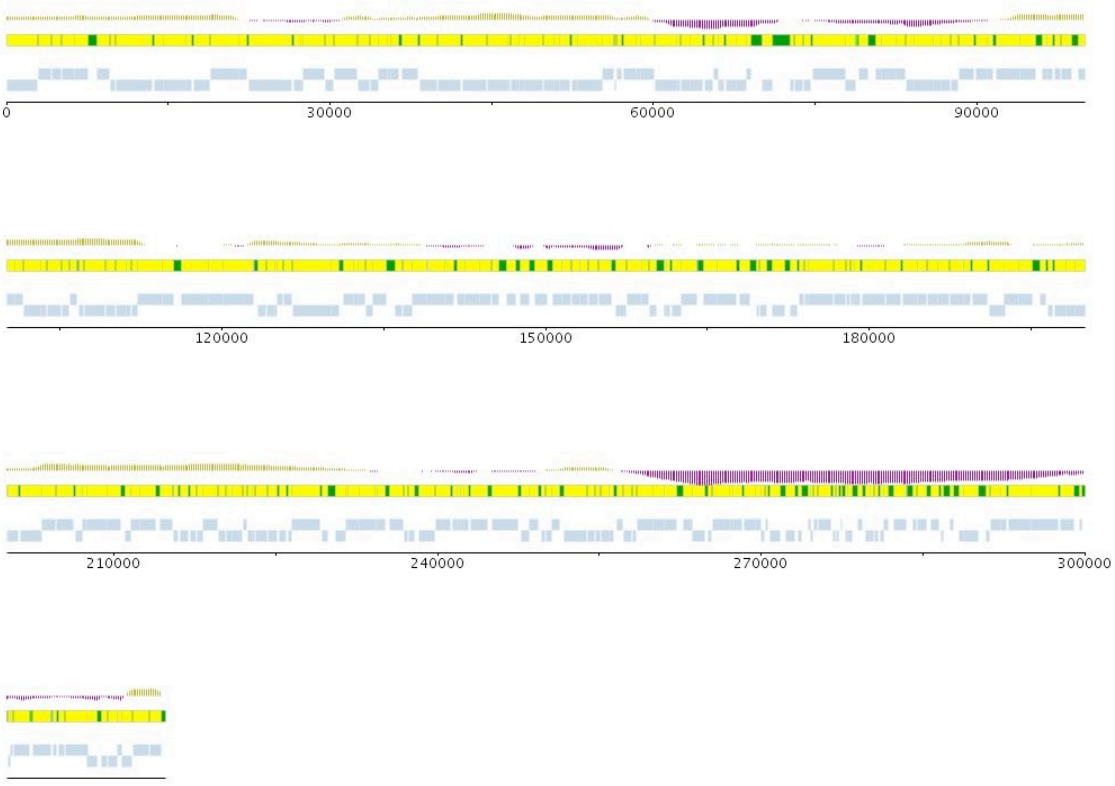


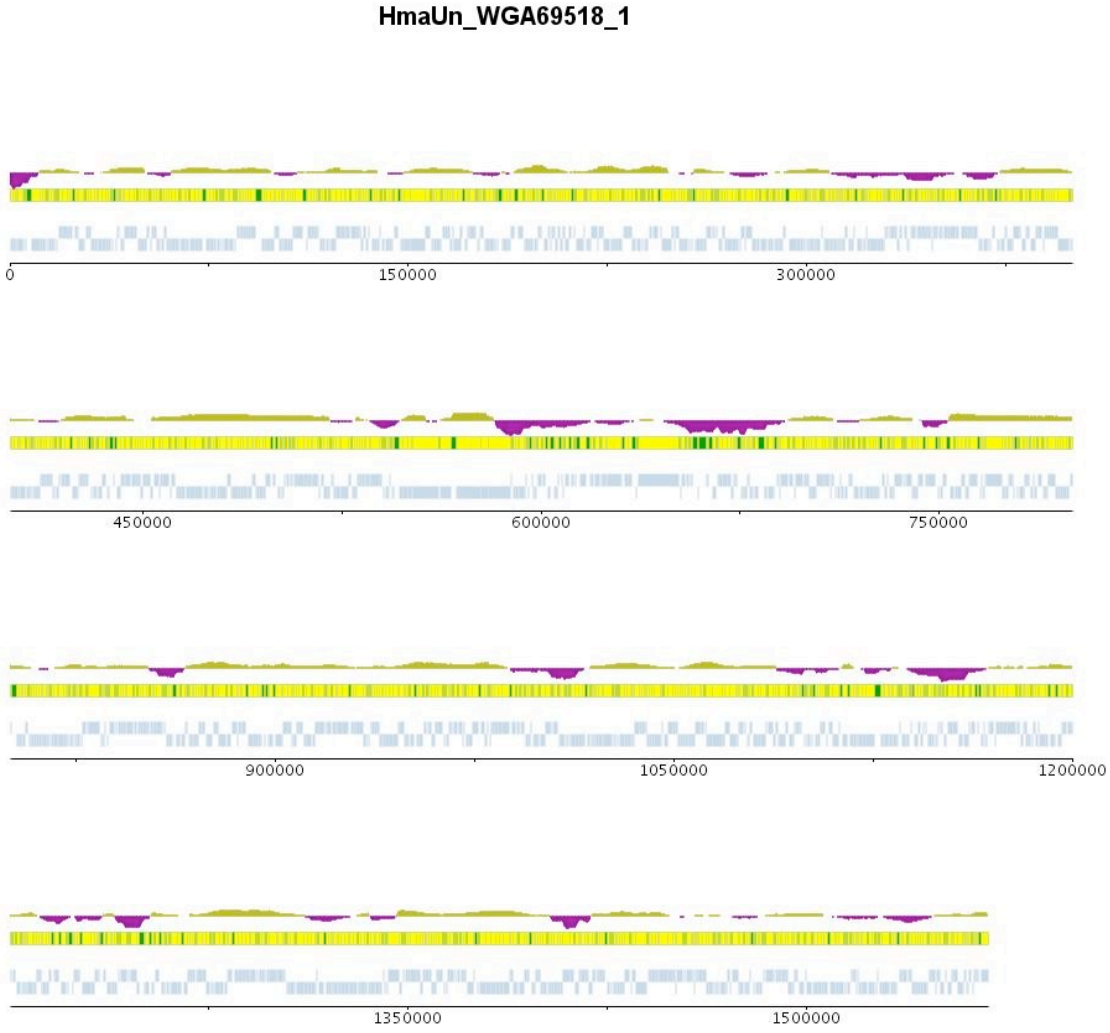
Electron micrograph of *Hydra* epithelium showing three gram-negative bacteria under the glycocalyx but outside the cell membrane of the epithelial cell. Scale bar is 200 nm.

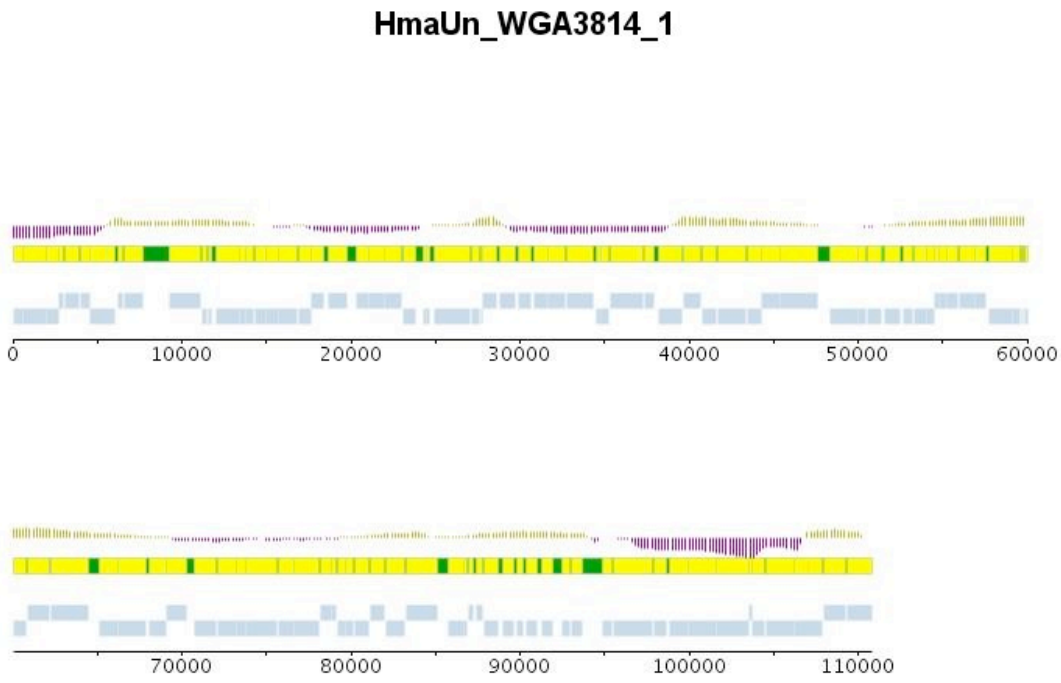
FIGURE S11: Gene Maps of *Curvibacter sp.* Scaffolds in the CA *Hydra* Genome Assembly

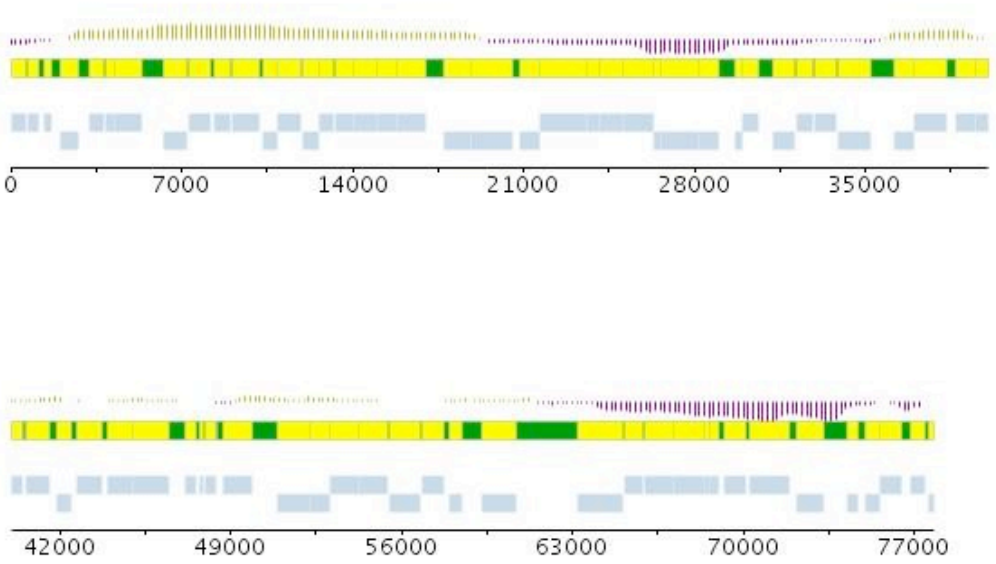


HmaUn_WGA12741_1

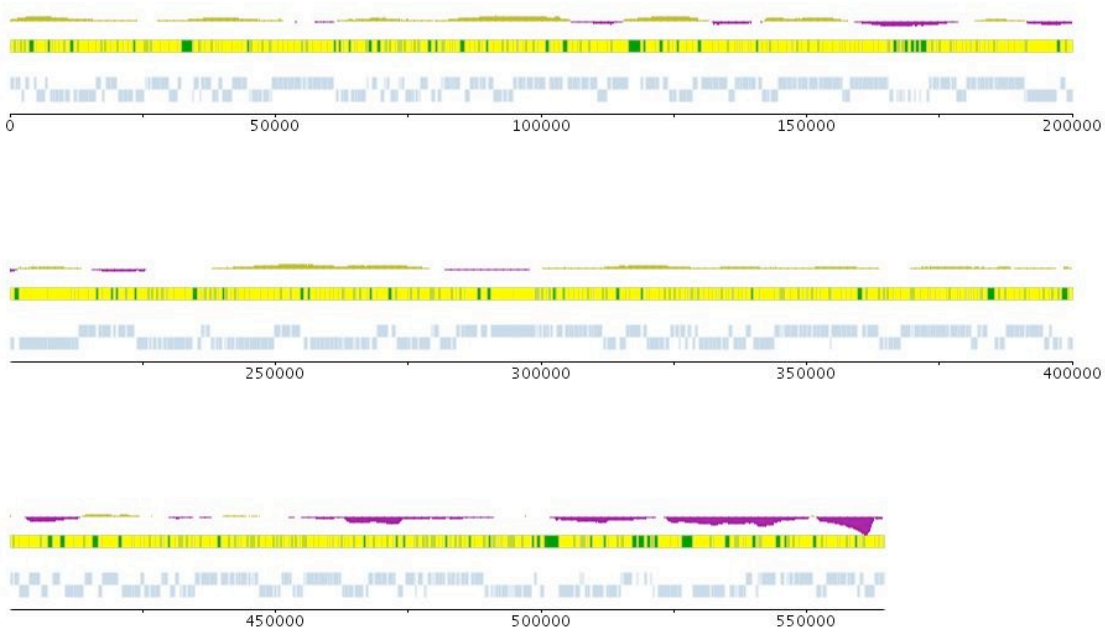



HmaUn_WGA69518_1
0
150000
300000
450000
600000
750000
900000
1050000
1200000
1350000
1500000

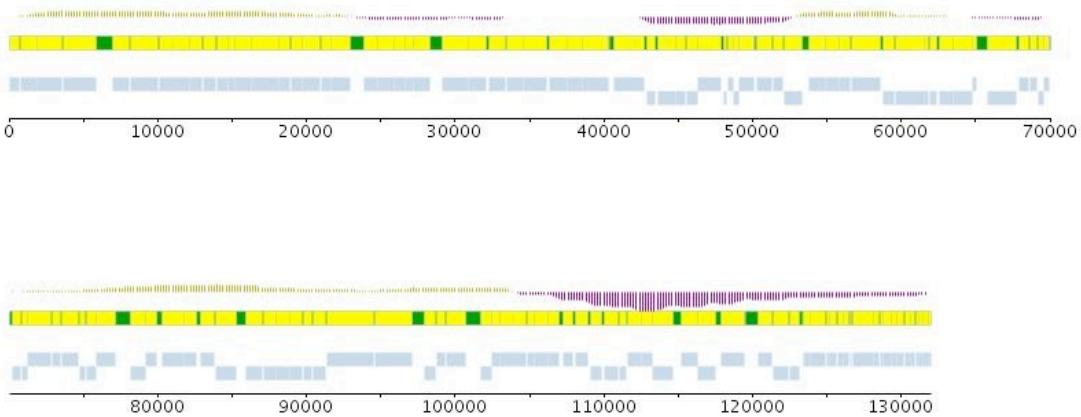

HmaUn_WGA3814_1
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
110000

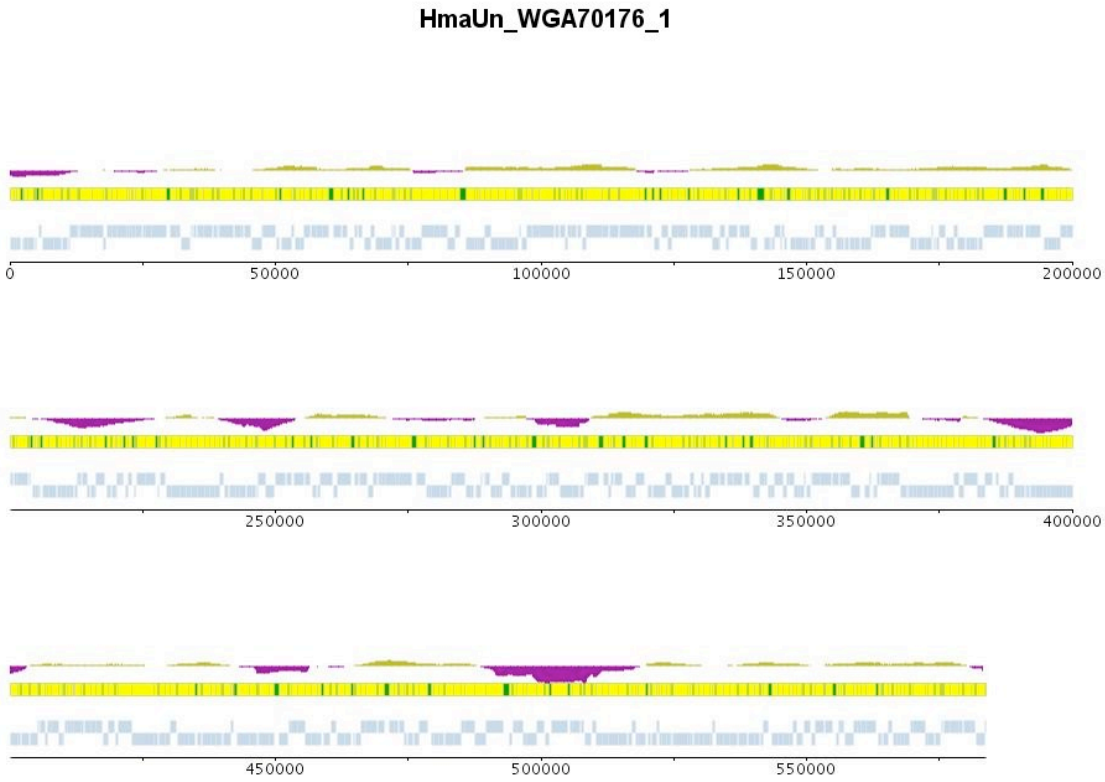

HmaUn_WGA26082_1
0
7000
14000
21000
28000
35000
42000
49000
56000
63000
70000
77000

HmaUn_WGA71069_1



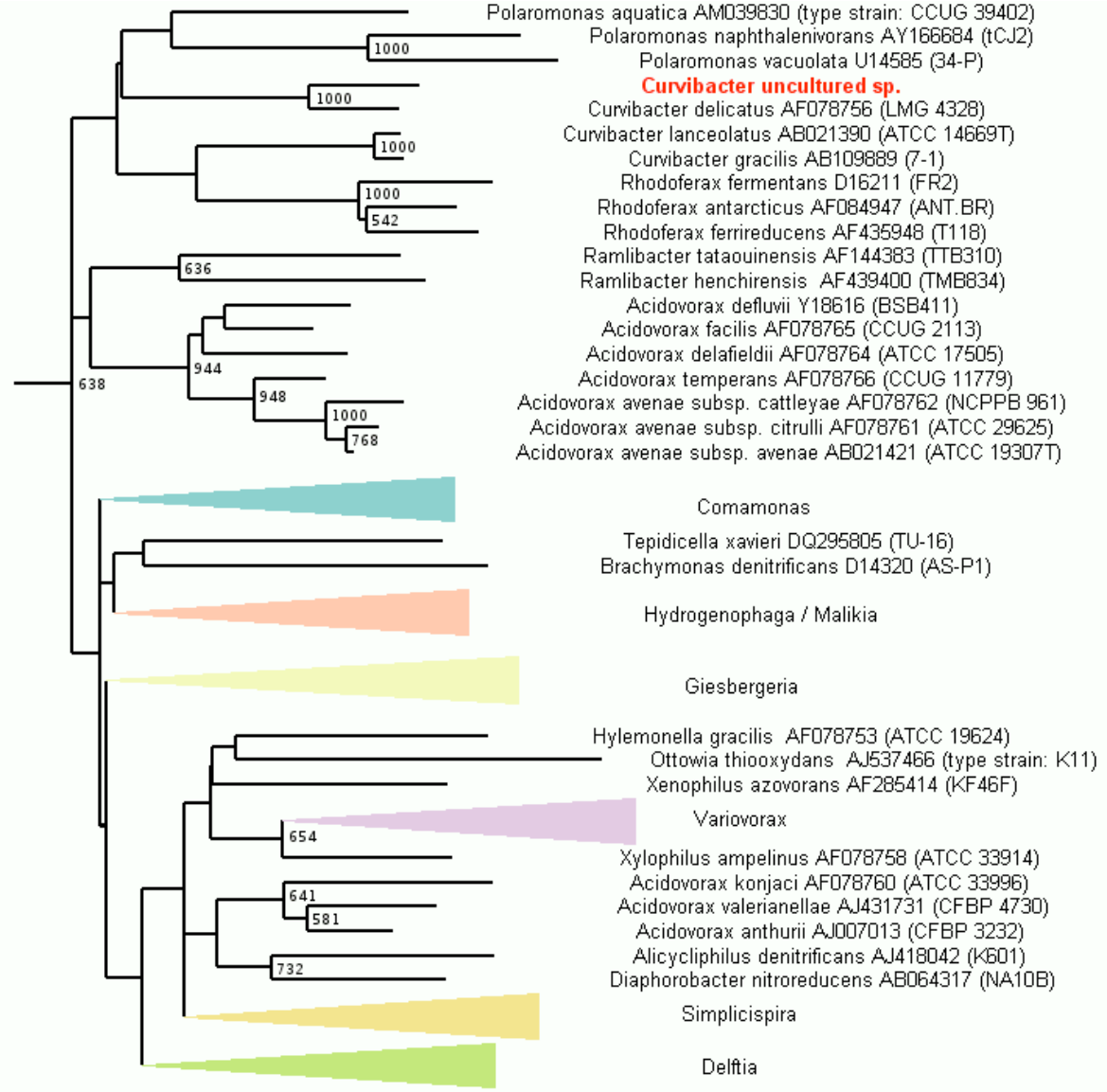
HmaUn_WGA70434_1





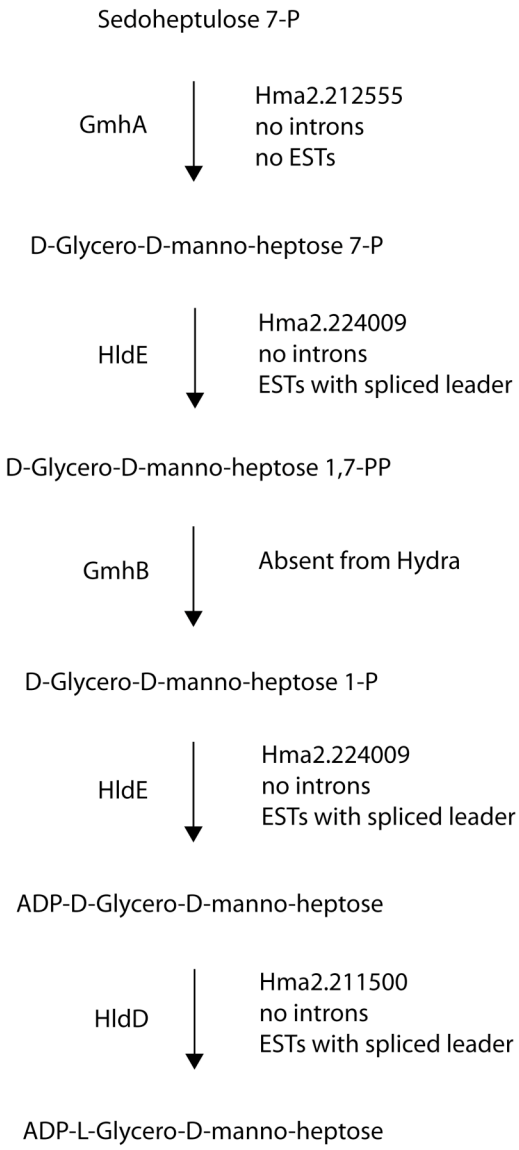
Eight scaffolds in the CA assembly at NCBI encode 4 megabases of sequence from the *Curvibacter sp*. chromosome. Top bar, deviation from the average G+C content; middle bar, coding sequence (yellow), non-coding sequence (green); bottom bar, annotated genes on the two strands of DNA. Positions on the assembly in base pairs are shown at the bottom.

FIGURE S12: Neighbor-Joining Phylogeny of 16S rRNA Sequences from Comamonadaceae



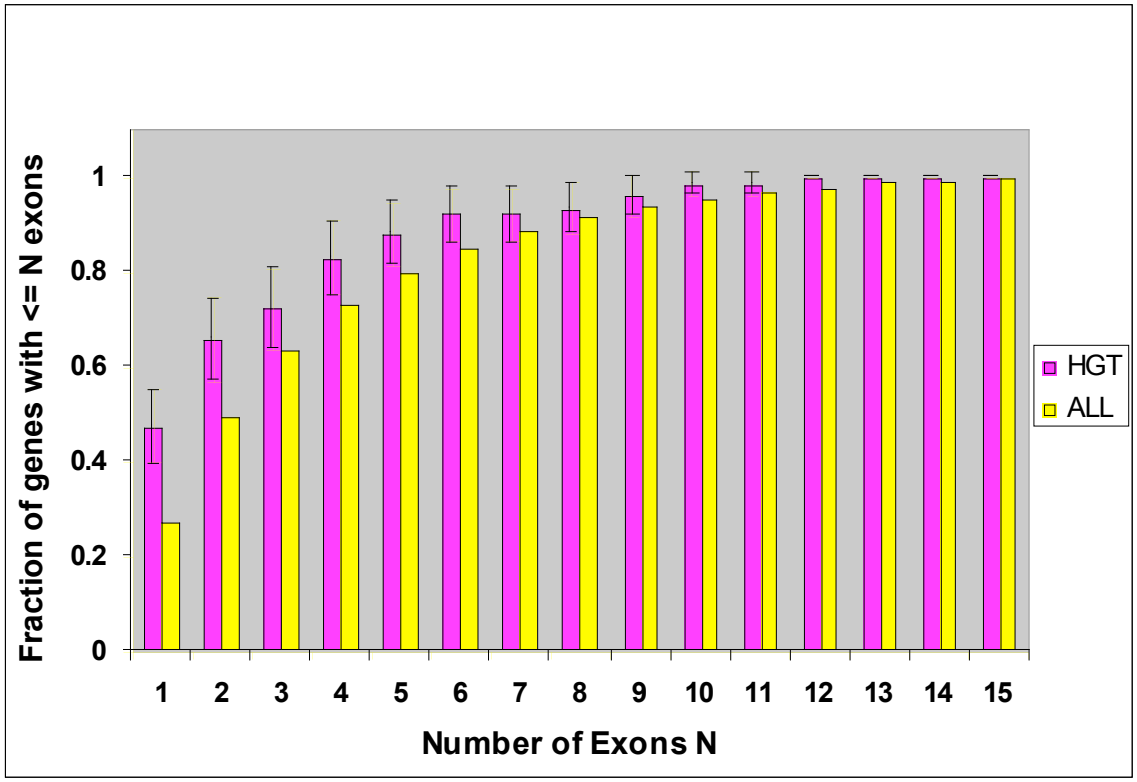
The sequences in the phylogeny are derived from the RDP database. The values are bootstrap values for 1000 replicates. The *Curvibacter sp.* 16S sequence is found on scaffold HmaUn_WGA71069_1 from position 171770 to 170304.

FIGURE S13: LPS Pathway Genes in *Hydra*



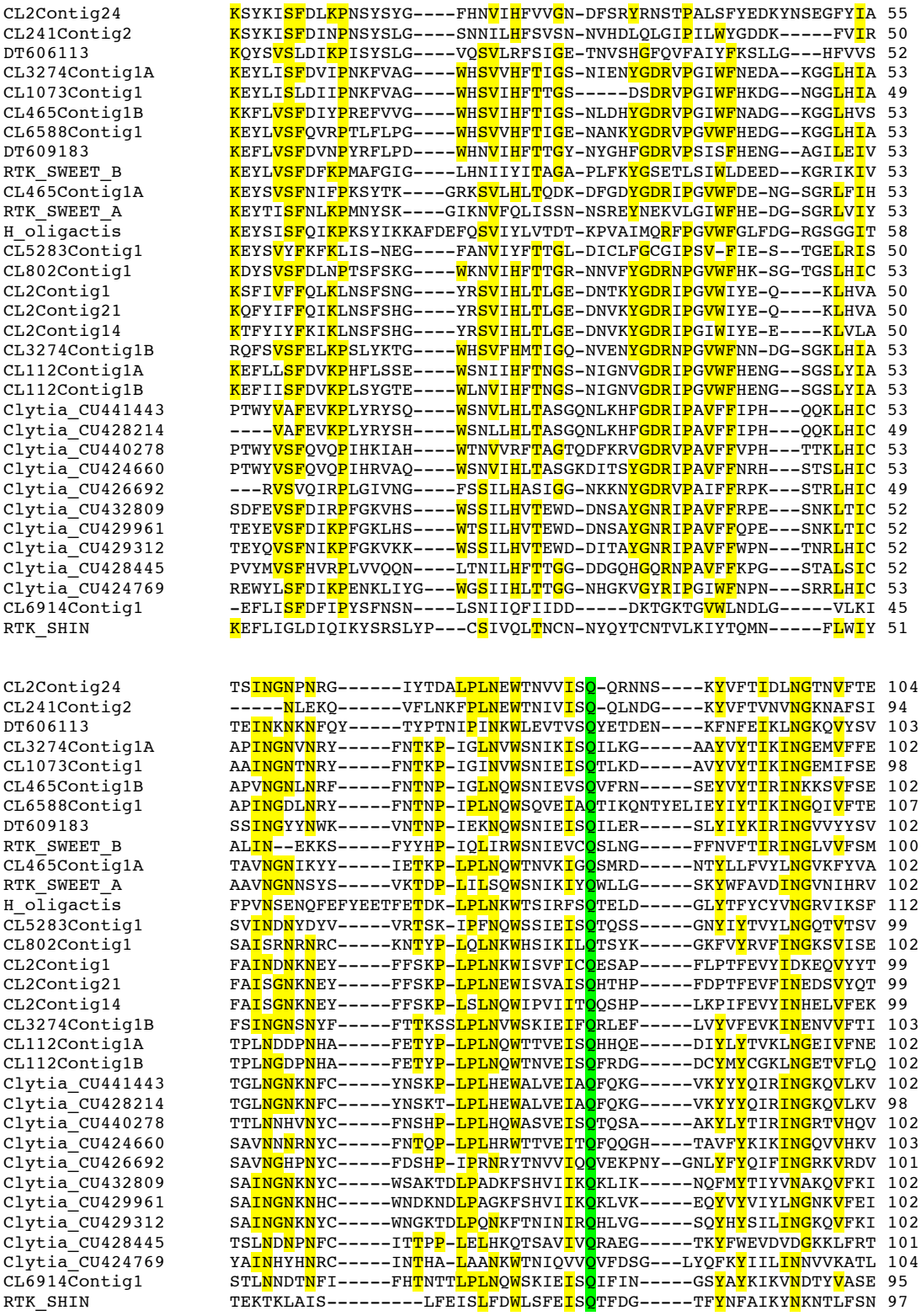
This figure shows the pathway for ADP-L-Glycero-D-mannose-heptose synthesis in bacteria. The names of the enzymes that catalyze the various steps in the pathway are shown on the left and their corresponding *Hydra* gene model numbers and gene characteristics are shown on the right.

FIGURE S14: Cumulative Distribution of Exon Number in Horizontally-Transferred Genes and in the Total Gene Model Set



Cumulative distribution of exon counts of the 76 candidate horizontal transfer genes (HGT) versus all 31,570 annotated genes. The HGT set shows error bars (95% C.I.) due to the small size of the HGT sample.

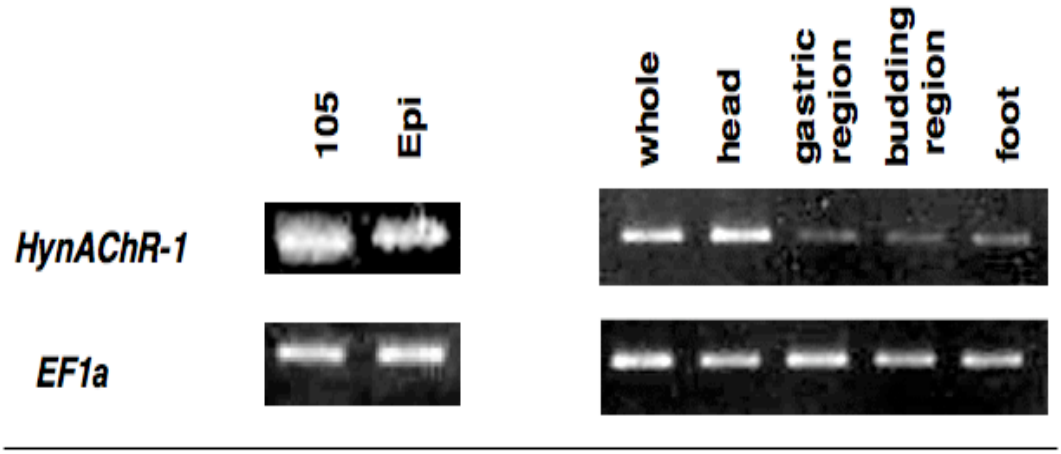
FIGURE S15: SWT Domain Sequences from *Hydra* and *Clytia*



```
CL2Contig24          RNNKPQNFNNVKVFASDPWYPSHDGSIKNLIIEN 138
CL241Contig2         ENKNTDVFENVVVYASNPWNLPFDGSIKKLTLEN 128
DT606113             VNNHPVDHSNVKVYASDPYFPALTGLIRNFRFSN 137
CL3274Contig1A       INNQAQYFDNVKVYASDPWYEVQDGSIRNLFLSN 136
CL1073Contig1        INNQAQYFDNVKVYASDPWYEVQDGSIKNLYIIN 132
CL465Contig1B        INNQVQSFDNVKVYAADPWHDVQNGLIKDFFVVN 136
CL6588Contig1        KNQQPQVFKNVKVYASDPWYEAQNGQIKNLYILN 141
DT609183             TNNQSQSFDNVKVYVSNPWYDAQDGSIKNLFVIN 136
RTK_SWEET_B          LNNQTMDFINVKIYASNIWDKVQNGTIKNFFIVN 134
CL465Contig1A        ENTDARDFKNVKVYASDPWYDAQNGLISNILIVN 136
RTK_SWEET_A          ENFLVSHFNEIKIYTSNLWDDAHNGSISDLLIVN 136
H_oligactis          ENQYPNSFKDVKVYLADPWNDVQDASIKNFKIVN 146
CL5283Contig1        FNKKPQMFPNVNVYATHPWRNGAVGFIKDLKIIN 133
CL802Contig1         INKDARVYKEVKVYVSDPWYAPQSGFIKNLKIVN 136
CL2Contig1           ENRNQKVFTNINVYAGDPWYEVQDGSIKEFAVFN 133
CL2Contig21          KNLNQKEFTNINVYAGDPWYEVQDGSIKHFLIFD 133
CL2Contig14          ENLNQKEFTNINVYAGDPWYEVQDGSIKDFFVYN 133
CL3274Contig1B       NNNDARDFKNVKVYVSDPWYDAQPGLVKNVKIIN 137
CL112Contig1A        QNEHAKSFKNVKVFASDPWHNAQNGSIKNLYIIN 136
CL112Contig1B        ENAVPRSFKNVKVFSSDPWHLPQQGLIRNFYTIN 136
Clytia_CU441443      LNKNARYFMGVDVYRGDPFYRQAPAKIRNFVYHN 136
Clytia_CU428214      LNKNARYFMGVDVYRGDPFYRQAPAKIRNFV--- 129
Clytia_CU440278      ENKGARYYMGVNVFRGDNFYGPAPARIRQLTYHN 136
Clytia_CU424660      QNKNVRFFPNVQVYRGNPFYTAAPARIRAFTYKN 137
Clytia_CU426692      LNTKPQVFRNVQYFASNPFYQPAIANIRNFRLT- 134
Clytia_CU432809      VNTRPITLYRAKVFLSSPWHPAAKAFVQRVRVTT 136
Clytia_CU429961      VNTRPITLYRAKVFLSSPWYAAAHAFVKRVRVTT 136
Clytia_CU429312      VNTKPITLYRAKVYLSNPWYLAAEAFVKNVVVTT 136
Clytia_CU428445      ENKNPQYFRQIQVYCSNRWHTPASGTMRGLYYDN 135
Clytia_CU424769      INSRPRSFQNVQVWTADPWHHAANAFLRNIVFQN 138
CL6914Contig1        INTQVRSFKNVVLYASNPWNVSQNGSIKNVFIAN 129
RTK_SHIN             VYNEVQVFNDIKIYASDPWSDSLDAAIRNFKIIN 131
```

Alignment of SWT domain sequences encoded by *Hydra* and *Clytia* ESTs. Alignment was done using CLUSTAL W. Yellow shading indicates amino acids conserved in greater than 50% of the sequences. Green shading indicates the glutamine residue conserved in all sequences. *Hydra* sequences are from assembled ESTs (Compagen[161]), previously characterized receptor protein-tyrosine kinase genes (RTK_SHIN and RTK_SWEET[162,163]), and a *H. oligactis* sequence in GenBank (accession number ACF20992). The *Clytia* sequences are from *Clytia* ESTs in GenBank. Manual examination of the assembled ESTs and their corresponding genomic sequences indicate that the SWT domain is present in one or more copies in predicted secreted proteins that lack other domains.

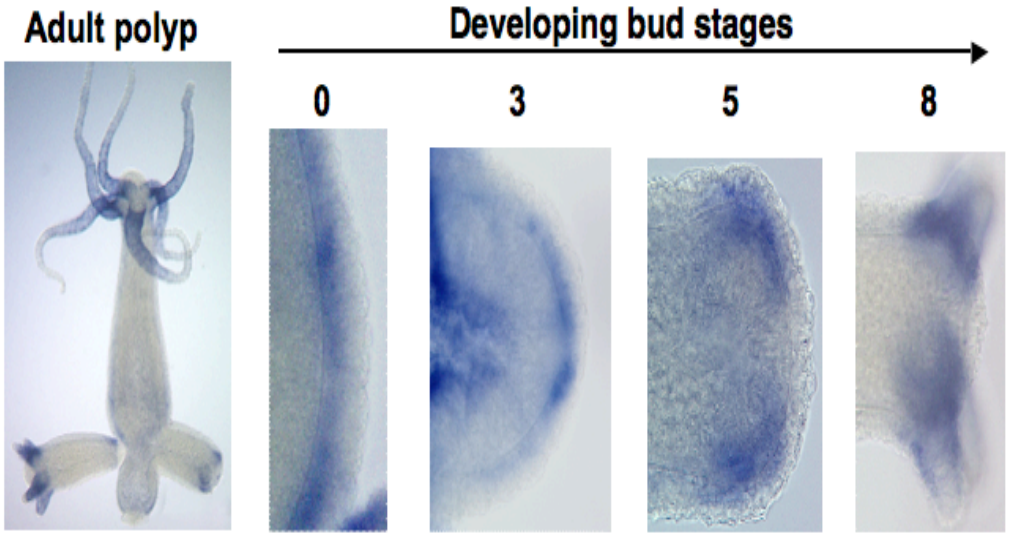
FIGURE S16: RT-PCR Analysis of Nicotinic Acetylcholine Receptor Gene Expression in *Hydra*



Gene	105	Epi	whole	head	gastric region	budding region	foot
HynAChR-1	++	++	++	++	+	+	+
HynAChR-2	++	++	++	++	+	+	+
HynAChR-3	++	+	++	++	+	+	+
HynAChR-4	++	++	++	++	+	+	+
HynAChR-5	++	++	++	++	+	+	+
HyChT-1	++	++	++	++	+	+	+

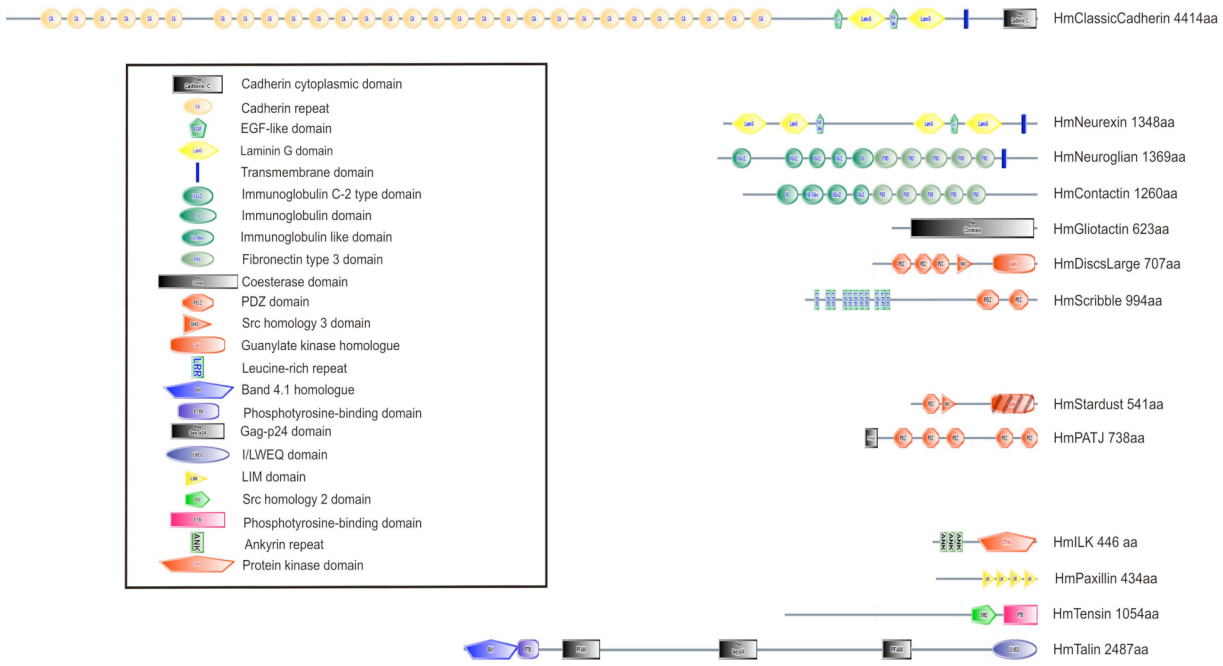
RT-PCR analysis of expression of nicotinic acetylcholine receptor genes and a choline transporter gene in non-budding *Hydra*. Upper panel shows expression of HynAChR-1 together with EF-1α. Lower panel shows expression of five HynAChRs and HyChT-1.

FIGURE S17: In Situ Hybridization Analysis of the HynAChR-like 1 Gene



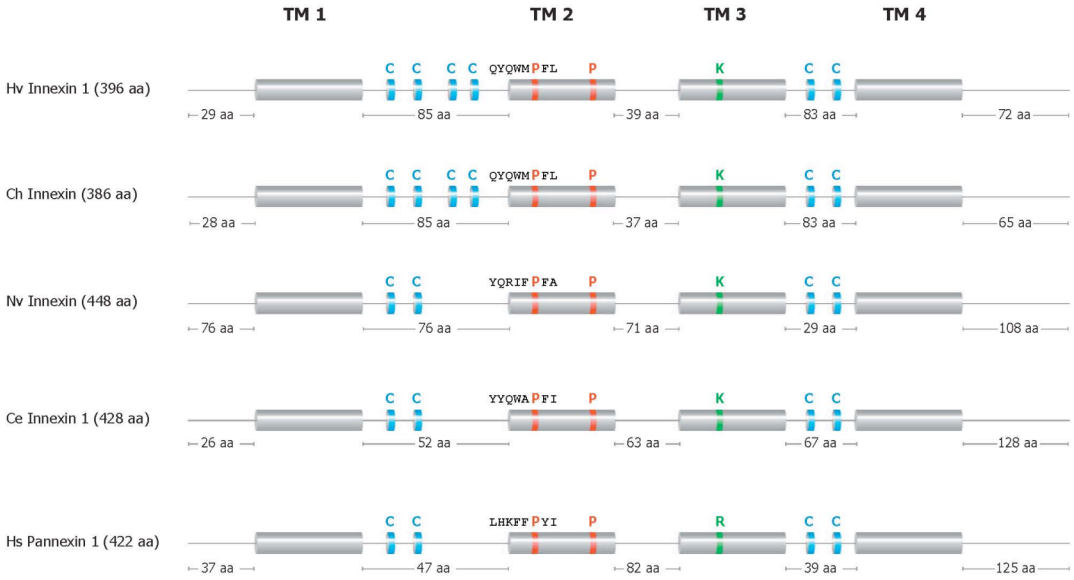
In the adult polyp, the gene is expressed strongly in the tentacle region and weakly throughout the body column, except for the hypostome and the basal disk. During budding, the gene was strongly expressed in the ectoderm of the presumptive budding region (stage 0) and gradually became confined to the apical tip (stage 3), the presumptive tentacle bud (stage 5), and finally to the tentacles (stage 8 and beyond).

FIGURE S18: Domain Structures of Predicted Cell-Cell and Cell-Substrate Contact Proteins in *Hydra*



The protein structures are based on predictions of domains by SMART/Pfam analysis. We show a selected set of major transmembrane proteins and cytoplasmic proteins associated with cell-cell and cell-substrate contact sites.

FIGURE S19: Conserved Sequence Motifs in Innexin and Pannexin Proteins



This figure shows four transmembrane domains and conserved sequence motifs in innexins from *Hydra* (Hv), *Clytia* (Ch), *Nematostella* (Nv), *C. elegans* (Ce), and pannexin from human (Hs).

References

1 Campbell, R. D. in *Hydra: Research Methods* (ed H.M. Lenhoff) 19-28 (Plenum Press, 1983).

2 Hemmrich, G., Anokhin, B., Zacharias, H. & Bosch, T. C. Molecular phylogenetics in Hydra, a classical model in evolutionary developmental biology. *Mol. Phylogenet. Evol.* **44**, 281-290, (2007).

3 Holstein, T. W., Campbell, R. D. & Tardent, P. Identity crisis. *Nature* **346**, 21-22, (1990).

4 Fujisawa, T. *Hydra* peptide project 1993-2007. *Dev. Growth Differ.* **50 Suppl 1**, S257-268, (2008).

5 Takahashi, T. *et al.* Systematic isolation of peptide signal molecules regulating development in hydra: LWamide and PW families. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1241-1246, (1997).

6 Sugiyama, T. in *Hydra: Research Methods* (ed H.M. Lenhoff) 211-221 (Plenum Press, 1983).

7 Zacharias, H., Anokhin, B., Khalturin, K. & Bosch, T. C. G. Genome sizes and chromosomes in the basal metazoan *Hydra*. *Zoology* **107**, 219-227, (2004).

8 Wittlieb, J., Khalturin, K., Lohmann, J. U., Anton-Erxleben, F. & Bosch, T. C. G. Transgenic *Hydra* allow in vivo tracking of individual stem cells during morphogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6208-6211, (2006).

9 Ito, T. A new fresh-water polyp, *Hydra magnipapillata*, n. sp. from Japan. *Science Reports of the Tohoku University, Fourth Series (Biology)* **18**, 6-10, (1947).

10 Lenhoff, H. M. & Brown, R. D. Mass culture of *Hydra*: an improved method and its application to other aquatic invertebrates. *Lab. Anim.* **4**, 139-154, (1970).

11 Glöckner, G. Large scale sequencing and analysis of AT rich eukaryotic genomes. *Curr. Genomics* **1**, 289-299, (2000).

12 Vinson, J. P. *et al.* Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.* **15**, 1127-1135, (2005).

13 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071, (2008).

14 Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254, (2007).

15 Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196-2204, (2000).

16 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351., (2001).

17 Wheeler, W. C. & Gladstein, D. S. MALIGN: a multiple sequence alignment program. *J. Heredity* **85**, 417-418, (1994).

18 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664, (2002).

19 Nordborg, M. in *Handbook of Statistical Genetics* eds D.J. Balding, M.J. Bishop, & C. Cannings) 179-212 (John Wiley & Sons, 2003).

20 Hardie, D. C., Gregory, T. R. & Hebert, P. D. From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. *J. Histochem. Cytochem.* **50**, 735-749, (2002).

21 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716, (2004).

22 Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195, (2000).

23 Broun, M., Gee, L., Reinhardt, B. & Bode, H. R. Formation of the head organizer in hydra involves the canonical Wnt pathway. *Development* **132**, 2907-2916, (2005).

24 Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803-816, (2001).

25 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucl. Acids Res.* **31**, 5654-5666, (2003).

26 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225, (2003).

27 Stein, L. D. *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**, 1599-1610, (2002).

28 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792-1797, (2004).

29 Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166, (1989).

30 Li, R. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.* **1**, e43, (2005).

31 Kapitonov, V. V. & Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **9**, 411-412; author reply 414, (2008).

32 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10, (2004).

33 Stover, N. A. & Steele, R. E. Trans-spliced leader addition to mRNAs in a cnidarian. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5693-5698, (2001).

34 Hampson, S., Kibler, D. & Baldi, P. Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics* **18**, 513-528, (2002).

35 Steele, R. E., Hampson, S. E., Stover, N. A., Kibler, D. F. & Bode, H. R. Probable horizontal transfer of a gene between a protist and a cnidarian. *Curr. Biol.* **14**, R298-R299, (2004).

36 Blumenthal, T. *et al.* A global analysis of *Caenorhabditis elegans* operons. *Nature* **417**, 851-854, (2002).

37 Davis, R. E. & Hodgson, S. Gene linkage and steady state RNAs suggest *trans*-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* **89**, 25-39, (1997).

38 Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D. & Thompson, E. M. Spliced-leader RNA *trans* splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol. Cell.*

Biol. **24**, 7795-7805, (2004).

39 Satou, Y. *et al.* Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biol* **9**, R152, (2008).

40 Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucl. Acids Res.* **36**, D250-254, (2008).

41 Walter, M. C. *et al.* PEDANT covers all complete RefSeq genomes. *Nucl. Acids Res.* **37**, D408-411, (2009).

42 Rattei, T. *et al.* SIMAP--structuring the network of protein similarities. *Nucl. Acids Res.* **36**, D289-292, (2008).

43 Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucl. Acids Res.* **37**, D141-145, (2009).

44 Ding, L. & Yokota, A. Proposals of *Curvibacter gracilis* gen. nov., sp. nov. and *Herbaspirillum putei* sp. nov. for bacterial strains isolated from well water and reclassification of [*Pseudomonas*] *huttiensis*, [*Pseudomonas*] *lanceolata*, [*Aquaspirillum*] *delicatum* and [*Aquaspirillum*] *autotrophicum* as *Herbaspirillum huttiense* comb. nov., *Curvibacter lanceolatus* comb. nov., *Curvibacter delicatus* comb. nov. and *Herbaspirillum autotrophicum* comb. nov. *Int. J. Syst. Evol. Microbiol.* **54**, 2223-2230, (2004).

45 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27-30, (2000).

46 Gloor, G. B. *et al.* Type I repressors of P element mobility. *Genetics* **135**, 81-95, (1993).

47 Giegerich, R., Meyer, F. & Schleiermacher, C. GeneFisher-software support for detection of postulated genes. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 68-77, (1996).

48 Fraune, S. & Bosch, T. C. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13146-13151, (2007).

49 Bosch, T. C. G. & David, C. N. Stem cells of *Hydra magnipapillata* can differentiate into somatic cells and germ line cells. *Devel. Biol.* **121**, 182-191, (1987).

50 Buss, L. W. *The Evolution of Individuality*. (Princeton University Press, 1987).

51 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, (2006).

52 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128, (2007).

53 Denker, E., Bapteste, E., Le Guyader, H., Manuel, M. & Rabet, N. Horizontal gene transfer and the evolution of cnidarian stinging cells. *Curr. Biol.* **18**, R858-859, (2008).

54 Hwang, J. S. *et al.* The evolutionary emergence of cell type-specific genes inferred from the gene expression analysis of *Hydra*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14735-14740, (2007).

55 Milde, S. *et al.* Characterization of taxonomically restricted genes in a phylum-restricted cell type. *Genome Biol.* **10**, R8, (2009).

56 David, C. N. *et al.* Evolution of complex structures: minicollagens shape the cnidarian nematocyst. *Trends Genet.* **24**, 431-438, (2008).

57 Wheeler, B. M. *et al.* The deep evolution of metazoan microRNAs. *Evol. Dev.* **11**, 50-68,

(2009).
58 Zuker, M., Mathews, D. H. & D.H. Turner, D. H. in *RNA Biochemistry and Biotechnology NATO ASI Series* eds J. Barciszewski & B.F.C. Clark) 11-43 (Kluwer Academic Publishers, 1999).
59 Ambros, V. *et al.* A uniform system for microRNA annotation. *RNA* **9**, 277-279, (2003).
60 Friedlander, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407-415, (2008).
61 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucl. Acids Res.* **36**, D154-158, (2008).
62 Washietl, S. Prediction of structural noncoding RNAs with RNAz. *Methods Mol. Biol.* **395**, 503-526, (2007).
63 Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271-290, (1999).
64 Vize, P. D. Transcriptome analysis of the circadian regulatory network in the coral Acropora millepora. *Biol. Bull.* **216**, 131-137, (2009).
65 Fritzenwanker, J. H. & Technau, U. Induction of gametogenesis in the basal cnidarian *Nematostella vectensis* (Anthozoa). *Dev. Genes Evol.* **212**, 99-103, (2002).
66 Shagin, D. A. *et al.* GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Mol. Biol. Evol.* **21**, 841-850, (2004).
67 Bode, H. R. The interstitial cell lineage of hydra: a stem cell system that arose early in evolution. *J. Cell Sci.* **109**, 1155-1164, (1996).
68 Bosch, T. C. Hydra and the evolution of stem cells. *Bioessays* **31**, 478-486, (2009).
69 David, C. N. & Murphy, S. Characterization of interstitial stem cells in hydra by cloning. *Dev. Biol.* **58**, 372-383, (1977).
70 Feng, B., Ng, J. H., Heng, J. C. & Ng, H. H. Molecules that promote or enhance reprogramming of somatic cells to induced pluripotent stem cells. *Cell Stem Cell* **4**, 301-312, (2009).
71 Hartl, M. *et al.* Stem cell-specific activiation of an ancestral *myc* protooncogene with conserved basic functions in the early metazoan *Hydra*. *Proc. Natl. Acad. Sci. U. S. A.* **in press**, (2010).
72 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).
73 Phochanukul, N. & Russell, S. No backbone but lots of Sox: Invertebrate Sox genes. *Int. J. Biochem. Cell Biol.*, (2009).
74 Jager, M., Queinnec, E., Houliston, E. & Manuel, M. Expansion of the SOX gene family predated the emergence of the Bilateria. *Mol. Phylogenet. Evol.* **39**, 468-477, (2006).
75 Martin, V. & Archer, W. Migration of interstitial cells and their derivatives in a hydrozoan planula. *Dev. Biol.* **116**, 486-496, (1986).
76 Muller, W. A., Teo, R. & Frank, U. Totipotent migratory stem cells in a hydroid. *Dev. Biol.* **275**, 215-224, (2004).
77 Rebscher, N., Volk, C., Teo, R. & Plickert, G. The germ plasm component Vasa allows tracing of the interstitial stem cells in the cnidarian *Hydractinia echinata*. *Dev. Dyn.* **237**, 1736-1745, (2008).

78 Denker, E., Manuel, M., Leclere, L., Le Guyader, H. & Rabet, N. Ordered progression of nematogenesis from stem cells through differentiation stages in the tentacle bulb of *Clytia hemisphaerica* (Hydrozoa, Cnidaria). *Dev. Biol.* **315**, 99-113, (2008).

79 Marlow, H. Q., Srivastava, M., Matus, D. Q., Rokhsar, D. & Martindale, M. Q. Anatomy and development of the nervous system of *Nematostella vectensis*, an anthozoan cnidarian. *Dev. Neurobiol.* **69**, 235-254, (2009).

80 Nakanishi, N., Yuan, D., Jacobs, D. K. & Hartenstein, V. Early development, pattern, and reorganization of the planula nervous system in *Aurelia* (Cnidaria, Scyphozoa). *Dev. Genes Evol.* **218**, 511-524, (2008).

81 Stangl, K., v. Salvini-Plawen, L. & Holstein, T. W. Staging and induction of medusa metamorphosis in *Carybdea marsupialis* (Cnidaria, Cubozoa). *Vie et Milieu* **52**, 131-140, (2002).

82 Schaller, H. C. & Bodenmüller, H. Isolation and amino acid sequence of a morphogenetic peptide from hydra. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7000-7004, (1981).

83 Schaller, H. C. A neurohormone from hydra is also present in the rat brain. *J. Neurochem.* **25**, 187-188, (1975).

84 Hampe, W. *et al.* A head-activator binding protein is present in hydra in a soluble and a membrane-anchored form. *Development* **126**, 4077-4086., (1999).

85 Browne, E. N. The production of new hydranths in hydra by the insertion of small grafts. *J. Exp. Zool.* **7**, 1-37, (1909).

86 Goldstein, B. & Freeman, G. Axis specification in animal development. *Bioessays* **19**, 105-116, (1997).

87 Hobmayer, B. *et al.* WNT signalling molecules act in axis formation in the diploblastic metazoan Hydra. *Nature (London)* **407**, 186-189., (2000).

88 Broun, M. & Bode, H. R. Characterization of the head organizer in hydra. *Development* **129**, 875-884, (2002).

89 Lengfeld, T. *et al.* Multiple Wnts are involved in Hydra organizer formation and regeneration. *Dev Biol* **330**, 186-199, (2009).

90 Philipp, I. *et al.* Wnt/beta-catenin and noncanonical Wnt signaling interact in tissue evagination in the simple eumetazoan Hydra. *Proc Natl Acad Sci U S A* **106**, 4290-4295, (2009).

91 Wikramanayake, A. H. *et al.* An ancient role for nuclear beta-catenin in the evolution of axial polarity and germ layer segregation. *Nature* **426**, 446-450, (2003).

92 Kusserow, A. *et al.* Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156-160, (2005).

93 Plickert, G., Jacoby, V., Frank, U., Muller, W. A. & Mokady, O. Wnt signaling in hydroid development: Formation of the primary body axis in embryogenesis and its subsequent patterning. *Dev Biol*, (2006).

94 Momose, T., Derelle, R. & Houliston, E. A maternally localised Wnt ligand required for axial patterning in the cnidarian Clytia hemisphaerica. *Development* **135**, 2105-2113, (2008).

95 Gerhart, J. Evolution of the organizer and the chordate body plan. *Int J Dev Biol* **45**, 133-153, (2001).

96 Kourakis, M. J. & Smith, W. C. Did the first chordates organize without the organizer? *Trends Genet* **21**, 506-510, (2005).

97 Yu, J. K. *et al.* Axial patterning in cephalochordates and the evolution of the organizer. *Nature* **445**, 613-617, (2007).

98 De Robertis, E. M. Spemann's organizer and self-regulation in amphibian embryos. *Nat Rev Mol Cell Biol* **7**, 296-302, (2006).

99 Kass-Simon, G. & Pierobon, P. Cnidarian chemical neurotransmission, an updated overview. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **146**, 9-25, (2007).

100 Govindasamy, L. *et al.* Structural insights and functional implications of choline acetyltransferase. *J. Struct. Biol.* **148**, 226-235, (2004).

101 Jogl, G., Hsiao, Y. S. & Tong, L. Structure and function of carnitine acyltransferases. *Ann. N. Y. Acad. Sci.* **1033**, 17-29, (2004).

102 Cronin, C. N. Redesign of choline acetyltransferase specificity by protein engineering. *J. Biol. Chem.* **273**, 24465-24469, (1998).

103 Candiani, S., Lacalli, T. C., Parodi, M., Oliveri, D. & Pestarino, M. The cholinergic gene locus in amphioxus: molecular characterization and developmental expression patterns. *Dev. Dyn.* **237**, 1399-1411, (2008).

104 Kawashima, K. & Fujii, T. Basic and clinical aspects of non-neuronal acetylcholine: overview of non-neuronal cholinergic systems and their biological significance. *J. Pharmacol. Sci.* **106**, 167-173, (2008).

105 Willmann, R. & Fuhrer, C. Neuromuscular synaptogenesis: clustering of acetylcholine receptors revisited. *Cell. Mol. Life Sci.* **59**, 1296-1316, (2002).

106 Sossin, W. S. Tracing the evolution and function of the Trk superfamily of receptor tyrosine kinases. *Brain Behav. Evol.* **68**, 145-156, (2006).

107 Oishi, I. *et al.* A novel Drosophila receptor tyrosine kinase expressed specifically in the nervous system. Unique structural features and implication in developmental signaling. *J. Biol. Chem.* **272**, 11916-11923, (1997).

108 Silman, I. & Sussman, J. L. Acetylcholinesterase: how is structure related to function? *Chem. Biol. Interact.* **175**, 3-10, (2008).

109 Scholl, F. G. & Scheiffele, P. Making connections: cholinesterase-domain proteins in the CNS. *Trends Neurosci.* **26**, 618-624, (2003).

110 Denker, E., Chatonnet, A. & Rabet, N. Acetylcholinesterase activity in *Clytia hemisphaerica* (Cnidaria). *Chem. Biol. Interact.* **175**, 125-128, (2008).

111 Sperling, E. A., Pisani, D. & Peterson, K. J. in *The Rise and Fall of the Ediacaran Biota* eds P. Vickers-Rich & P. Komarower) 355-368 (Geological Society, 2007).

112 Hyman, L. H. *The Invertebrates*. (1940).

113 Tyler, S. Epithelium--the primary building block of metazoan complexity. *Integr. Comp. Biol.* **43**, 55-63, (2003).

114 Abedin, M. & King, N. The premetazoan ancestry of cadherins. *Science* **319**, 946-948, (2008).

115 Brower, D. L., Brower, S. M., Hayward, D. C. & Ball, E. E. Molecular evolution of integrins: genes encoding integrin beta subunits from a coral and a sponge. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 9182-9187., (1997).

116 King, N., Hittinger, C. T. & Carroll, S. B. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science* **301**, 361-363, (2003).

117 King, N. *et al.* The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**, 783-788, (2008).

118 Magie, C. R. & Martindale, M. Q. Cell-cell adhesion in the cnidaria: insights into the evolution of tissue morphogenesis. *Biol. Bull.* **214**, 218-232, (2008).

119 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**, 12451-12456, (2006).

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

121 Sakarya, O. *et al.* A post-synaptic scaffold at the origin of the animal kingdom. *PLoS ONE* **2**, e506, (2007).

122 Srivastava, M. *et al.* The *Trichoplax* genome and the nature of placozoans. *Nature* **454**, 955-960, (2008).

123 Hortsch, M. & Margolis, B. Septate and paranodal junctions: kissing cousins. *Trends Cell Biol.* **13**, 557-561, (2003).

124 Berrier, A. L. & Yamada, K. M. Cell-matrix adhesion. *J. Cell. Physiol.* **213**, 565-573, (2007).

125 Miyoshi, J. & Takai, Y. Structural and functional associations of apical junctions with cytoskeleton. *Biochim. Biophys. Acta* **1778**, 670-691, (2008).

126 Paris, L., Tonutti, L., Vannini, C. & Bazzoni, G. Structural organization of the tight junctions. *Biochim. Biophys. Acta* **1778**, 646-659, (2008).

127 Hand, A. R. & Gobel, S. The structural organization of the septate and gap junctions of Hydra. *J. Cell Biol.* **52**, 397-408, (1972).

128 Westfall, J. A., Kinnamon, J. C. & Sims, D. E. Neuro-epitheliomuscular cell and neuro-neuronal gap junctions in *Hydra*. *J. Neurocytol.* **9**, 725-732, (1980).

129 Traut, W. *et al.* The telomere repeat motif of basal Metazoa. *Chromosome Res.* **15**, 371-382, (2007).

130 Kuhn, K., Streit, B. & Schierwater, B. Homeobox genes in the cnidarian *Eleutheria dichotoma*: evolutionary implications for the origin of *Antennapedia*-class (HOM/Hox) genes. *Mol. Phylogenet. Evol.* **6**, 30-38, (1996).

131 Davis, R. E. Surprising diversity and distribution of spliced leader RNAs in flatworms. *Mol. Biochem. Parasitol.* **87**, 29-48, (1997).

132 Suga, H., Schmid, V. & Gehring, W. J. Evolution and functional diversity of jellyfish opsins. *Curr. Biol.* **18**, 51-55, (2008).

133 Chourrout, D. *et al.* Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* **442**, 684-687, (2006).

134 Chiori, R. *et al.* Are Hox genes ancestrally involved in axial patterning? Evidence from the hydrozoan *Clytia hemisphaerica* (Cnidaria). *PLoS ONE* **4**, e4231, (2009).

135 Mokady, O., Dick, M. H., Lackschewitz, D., Schierwater, B. & Buss, L. W. Over one-half billion years of head conservation? Expression of an *ems* class gene in *Hydractinia symbiolongicarpus* (Cnidaria: Hydrozoa). *Proc. Natl. Acad. Sci. U. S. A.* **95**, 3673-3678, (1998).

136 Miller, D. J. & Miles, A. Homeobox genes and the zootype. *Nature* **365**, 215-216, (1993).

137 de Jong, D. M. *et al.* Components of both major axial patterning systems of the Bilateria are differentially expressed along the primary axis of a 'radiate' animal, the anthozoan cnidarian *Acropora millepora*. *Dev. Biol.* **298**, 632-643, (2006).

138 Finnerty, J. R. & Martindale, M. Q. Homeoboxes in sea anemones (Cnidaria:Anthozoa): a PCR-based survey of Nematostella vectensis and Metridium senile. *Biol. Bull.* **193**, 62-76, (1997).

139 Lee, P. N., Pang, K., Matus, D. Q. & Martindale, M. Q. A WNT of things to come: evolution of Wnt signaling and polarity in cnidarians. *Semin. Cell Dev. Biol.* **17**, 157-167, (2006).

140 Rentzsch, F., Guder, C., Vocke, D., Hobmayer, B. & Holstein, T. W. An ancient chordin-like gene in organizer formation of *Hydra*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 3249-3254, (2007).

141 Matus, D. Q. *et al.* Molecular evidence for deep evolutionary roots of bilaterality in animal development. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11195-11200, (2006).

142 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).

143 Rentzsch, F. *et al.* Asymmetric expression of the BMP antagonists chordin and gremlin in the sea anemone *Nematostella vectensis*: implications for the evolution of axial patterning. *Dev. Biol.* **296**, 375-387, (2006).

144 Böttger, A. *et al.* Genetic screen for signal peptides in *Hydra* reveals novel secreted proteins and evidence for non-classical protein secretion. *Eur. J. Cell Biol.* **85**, 1107-1117, (2006).

145 Reinhardt, B., Broun, M., Blitz, I. L. & Bode, H. R. HyBMP5-8b, a BMP5-8 orthologue, acts during axial patterning and tentacle formation in hydra. *Dev. Biol.* **267**, 43-59, (2004).

146 Finnerty, J. R., Pang, K., Burton, P., Paulson, D. & Martindale, M. Q. Origins of bilateral symmetry: Hox and dpp expression in a sea anemone. *Science* **304**, 1335-1337, (2004).

147 Guder, C. *et al.* An ancient Wnt-Dickkopf antagonism in *Hydra*. *Development* **133**, 901-911, (2006).

148 Broun, M., Sokol, S. & Bode, H. R. Cngsc, a homologue of *goosecoid*, participates in the patterning of the head, and is expressed in the organizer region of *Hydra*. *Development* **126**, 5245-5254, (1999).

149 Martinez, D. E. *et al.* Budhead, a fork head/HNF-3 homologue, is expressed during axis formation and head specification in hydra. *Dev. Biol.* **192**, 523-536, (1997).

150 Fritzenwanker, J. H., Saina, M. & Technau, U. Analysis of forkhead and snail expression reveals epithelial-mesenchymal transitions during embryonic and larval development of *Nematostella vectensis*. *Dev. Biol.* **275**, 389-402, (2004).

151 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**, 618-630, (2005).

152 Martindale, M. Q., Pang, K. & Finnerty, J. R. Investigating the origins of triploblasty:

'mesodermal' gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum, Cnidaria; class, Anthozoa). *Development* **131**, 2463-2474, (2004).

153 Smith, K. M., Gee, L., Blitz, I. L. & Bode, H. R. CnOtx, a member of the Otx gene family, has a role in cell movement in hydra. *Dev. Biol.* **212**, 392-404, (1999).

154 Mazza, M. E., Pang, K., Martindale, M. Q. & Finnerty, J. R. Genomic organization, gene structure, and developmental expression of three clustered otx genes in the sea anemone *Nematostella vectensis*. *J. Exp. Zoolog. B Mol. Dev. Evol.* **308**, 494-506, (2007).

155 Bielen, H. *et al.* Divergent functions of two ancient Hydra Brachyury paralogues suggest specific roles for their C-terminal domains in tissue fate induction. *Development* **134**, 4187-4197, (2007).

156 Technau, U. & Bode, H. R. HyBra1, a Brachyury homologue, acts during head formation in *Hydra*. *Development* **126**, 999-1010, (1999).

157 Scholz, C. B. & Technau, U. The ancestral role of Brachyury: expression of NemBra1 in the basal cnidarian *Nematostella vectensis* (Anthozoa). *Dev. Genes Evol.* **212**, 563-570, (2003).

158 Yasuoka, Y. *et al.* Evolutionary origins of blastoporal expression and organizer activity of the vertebrate gastrula organizer gene lhx1 and its ancient metazoan paralog lhx3. *Development* **136**, 2005-2014, (2009).

159 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462-467, (2005).

160 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, (2007).

161 Hemmrich, G. & Bosch, T. C. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. *Bioessays* **30**, 1010-1018, (2008).

162 Bridge, D. M., Stover, N. A. & Steele, R. E. Expression of a novel receptor tyrosine kinase gene and a paired-like homeobox gene provides evidence of differences in patterning at the oral and aboral ends of hydra. *Dev. Biol.* **220**, 253-262, (2000).

163 Reidling, J. C., Miller, M. A. & Steele, R. E. Sweet Tooth, a novel receptor protein-tyrosine kinase with C-type lectin-like extracellular domains. *J. Biol. Chem.* **275**, 10323-10330., (2000).