the FEBS Journal

The transcriptome of the colonial marine hydroid *Hydractinia echinata*

Jorge Soza-Ried[1], Agnes Hotz-Wagenblatt[2], Karl-Heinz Glatting[2], Coral del Val[2,3], Kurt Fellenberg[1,4], Hans R. Bode[5], Uri Frank[6], Jörg D. Hoheisel[1] and Marcus Frohme[1,7]

1 Division of Functional Genome Analysis, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany
2 Division of Molecular Biophysics, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany
3 Department of Computer Science and Artificial Intelligence, E.T.S.I. Informatics, Universidad de Granada, Spain
4 Bioanalytics Group, Technical University Munich, Freising, Germany
5 Developmental Biology Center and Developmental and Cell Biology Department, University of California at Irvine, CA, USA
6 School of Natural Sciences and Martin Ryan Marine Science Institute, National University of Ireland, Galway, Ireland
7 Molecular Biology and Functional Genome Analysis, University of Applied Sciences Wildau, Germany

Keywords
Cnidaria; database; EST; *Hydractinia*; transcriptome

Correspondence
J. Soza-Ried, Division of Functional Genome Analysis, Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 580, 69121 Heidelberg, Germany
Fax: +49 6221 424687
Tel: +49 6221 424678
E-mail: j.sozaried@dkfz-heidelberg.de

(Received 11 August 2009, revised 1 November 2009, accepted 3 November 2009)

doi:10.1111/j.1742-4658.2009.07474.x

An increasing amount of expressed sequence tag (EST) and genomic data, predominantly for the cnidarians *Acropora*, *Hydra* and *Nematostella*, reveals that cnidarians have a high genomic complexity, despite being one of the morphologically simplest multicellular animals. Considering the diversity of cnidarians, we performed an EST project on the hydroid *Hydractinia echinata*, to contribute towards a broader coverage of this phylum. After random sequencing of almost 9000 clones, EST characterization revealed a broad diversity in gene content. Corroborating observations in other cnidarians, *Hydractinia* sequences exhibited a higher sequence similarity to vertebrates than to ecdysozoan invertebrates. A significant number of sequences were hitherto undescribed in metazoans, suggesting that these may be either cnidarian innovations or ancient genes lost in the bilaterian genomes analysed so far. However, we cannot rule out some degree of contamination from commensal bacteria. The identification of unique *Hydractinia* sequences emphasizes that the acquired genomic information generated so far is not large enough to be representative of the highly diverse cnidarian phylum. Finally, a database was created to store all the acquired information (http://www.mchips.org/hydractinia_echinata.html).

Introduction

Cnidarians are considered to be among the most basal of living multicellular animals. Despite being characterized as morphologically simple organisms, recent cnidarian sequencing projects revealed a high complexity at the genetic level [1–5]. Several genes and signalling pathways associated with patterning and developmental processes in bilaterians are present in cnidarians. These include components of the wingless, transforming growth factor-β and fibroblast growth factor signalling pathways [1,6]. Additionally, many genes absent from invertebrate model systems, and therefore previously thought to be vertebrate innovations, have been identified in cnidarians. Members of the *wingless* gene subfamilies [1–4,6–9] are an example. Moreover, the genomic organization of cnidarians in terms of intron richness and degree of synteny resembles that of vertebrates rather than that of ecdysozoan invertebrates [1,10]. Sequencing data have also revealed a

Abbreviations
ASW, artificial seawater; EST, expressed sequence tag; FAS, fragment assembly system; GO, gene ontology; HUSAR, Heidelberg Unix Sequence Analysis Resource; NCBI, National Center for Biotechnology Information.

significant number of cnidarian protein-coding sequences that have not been detected in other animals, indicating that they might be either cnidarian innovations or ancient genes lost in the bilaterian genomes analysed so far [1,3].

The combination of the characteristics of the cnidarian genomes coupled with its phylogenetic position allows them to be used as a model system for deciphering the gene content of the last common eumetazoan ancestor. It also extends the understanding of the functional evolution of genes. Indeed, these experimental models are being used for medical research, providing new insights into the genetic and molecular mechanisms underlying human diseases [8,11].

One of the commonly used approaches for direct access to the transcribed genetic information is the sequencing of cDNA clones, resulting in expressed sequence tags (ESTs) [12]. To date, EST databases in cnidarians are predominantly based on the coral *Acropora millepora*, the solitary polyp *Hydra magnipapillata* and the sea anemone *Nematostella vectensis* [2,3]. Furthermore, the Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) recently released the assembled genome of *Nematostella* [1].

However, the phylum Cnidaria is a highly diverse group of animals. Some live as simple solitary or colonial polyps, such as the anthozoans, including *Nematostella* and *Acropora*, and some hydrozoans, such as *Hydra* and *Hydractinia*. Others have a life cycle characterized by alternating generations of polyps and a more complex form, the medusa (jellyfish), as most hydrozoans, scyphozoans and cubozoans [13]. Although the transcribed data of anthozoans are well represented by the model organisms *Nematostella* and *Acropora*, *Hydra* – as a freshwater solitary polyp – is not a typical representative of the class Hydrozoa, as most of its members are colonial and marine. Therefore, we analysed the transcriptome of a more typical member of this class, the colonial marine hydroid *Hydractinia echinata*. This animal offers attractive features of a good model organism. For example, many molecular techniques, including transgenic technology, are already available. Indeed, for decades *Hydractinia* has been a model system to study embryogenesis, metamorphosis, pattern formation and immunity [14–18].

In order to identify a large fraction of the genes represented in the *Hydractinia* transcriptome, we made use of pooled RNA preparations for the cDNA library construction that were collected from various stages of the animal's life cycle. Furthermore, we extended the pool with RNA obtained from several induction experiments. For the sequence analysis of each EST, we assigned it to a taxonomic homology group, as well as carrying out a detailed functional annotation. In particular, we considered nonmetazoan homologues, as growing evidence points to an unexpected role of such homologues in lower metazoans. These genes could be ancestral, belong to symbiotic or epiphytic organisms, or be the result of lateral gene transfer events [3,19–22]. The *Hydractinia* sequences were compared with the *Hydra*, *Acropora* and *Nematostella* DNA datasets in order to identify unique *Hydractinia* transcripts, as well as genes that might be related to the marine or colonial characteristics of *Hydractinia*. All acquired information is being stored in a relational database, which aims to provide easy access and handling of the existing *Hydractinia* data.

Results

Generation of the *Hydractinia echinata* ESTs

To generate a representative EST dataset of the *Hydractinia* transcriptome, we created a size-selected cDNA library, consisting of 21 120 clones. Quality analyses revealed cDNA inserts with a length between 0.4 and 5 kb and an average value of $\sim$ 1.8 kb (data not shown). From the randomly selected clones, 8151 sequences were generated from 5′-ends and 827 sequences from 3′-ends. The ESTs had an average and median length of 409 and 419 bp, respectively. The first clustering was made by physically merging sequence reads derived from clones that were sequenced from both ends. Finally, 8212 sequences were analysed as described in the methods section. The sequences were grouped into 3808 EST clusters, including 2625 singletons and 1183 clusters of two or more clones comprising 5587 ESTs (Fig. 1). Finally, we generated consensus sequences with an average length of 439 bp representing each EST cluster, which were used in the subsequent analyses.

ESTs functional annotation

BLASTX analysis showed that 1797 *Hydractinia* sequences (47.5%) with an acceptance cut-off E-value $< 10^{-6}$ matched entries in protein databases. A high percentage of ESTs (38.5%, 1468 sequences) exhibited no significant similarity to any known sequence, whereas 543 sequences (14%) presented an uninformative, i.e. hypothetical, probable, putative or chromosomal, annotation (Fig. 2A). In order to characterize these ESTs, we searched for known protein domain architectures within the sequences. This allowed the assignment of 267 new functional annotations (Table S1).

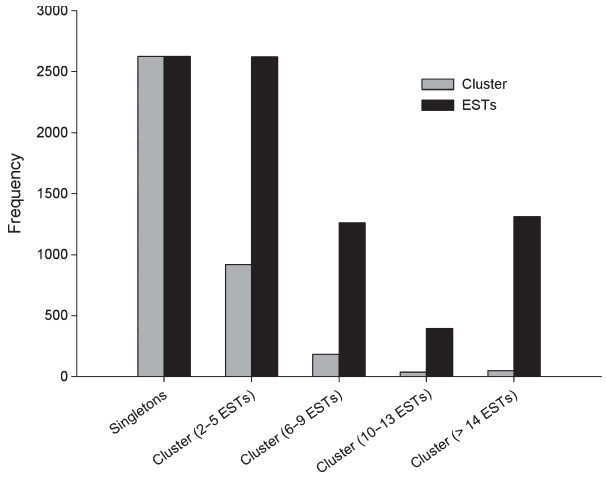

Fig. 1. Histogram of the size distribution of the EST clusters with their corresponding EST frequency. The *x*-axis shows the cluster size. The *y*-axis represents the frequency of each cluster size group and the abundance of ESTs. The *Hydractinia* ESTs were grouped into 3808 clusters, indicating a 2.2-fold normalization. One-third of the ESTs (2625) were represented only once (singletons) in the dataset, whereas 2622 ESTs were grouped into 919 clusters of 2–5 ESTs; 1261 ESTs were grouped into 182 clusters of 6–9 ESTs; 393 ESTs were grouped into 36 clusters of 10–13 ESTs; and 1311 ESTs were grouped into 46 clusters of more than 14 ESTs.

For an overview of all the different functional classes present in our data, we also annotated the sequences with gene ontology (GO) terms. In the category 'molecular function', the *Hydractinia* sequences were associated with different GO functions, including mainly hydrolase, transferase and binding activities. In the category 'biological process', the majority of the GO term predictions appeared to be related to metabolism (e.g. biosynthetic and catabolic processes), cell communication and biogenesis, as well as transport and regulation of biological processes (Fig. 2B).

Nonmetazoan hits

In the BLASTX analysis, 22% (844 sequences) of the *Hydractinia* proteins showed a nonmetazoan prokaryotic hit, of which 263 and 491 sequences had homologies to bacteria from the beta- and gamma-proteobacteria classes, respectively. Among the former, homologies to *Bordetella* spp. and *Burkholderia* spp. accounted for the majority of hits, whereas in the latter class, 425 sequences presented homology to *Pseudomonas* spp. To analyse if we were observing a common feature within cnidarians, we compared the *Hydractinia* sequences using the TBLASTX algorithm with the *Acropora*, *Hydra* and *Nematostella* EST datasets, as well as the recently annotated *Nematostella* genome. We observed that with an E-value acceptance threshold $< 10^{-3}$, 58% (487 sequences) of the prokaryotic protein sequences are represented at least in one of the mentioned datasets, including 331 sequences with a hit on the DNA of *Nematostella*. Analysis at the nucleotide level using BLASTN with the same significance criterion revealed that 201 of these sequences (24%) are common within cnidarians.

The GC content of the sequences classified as nonmetazoan was significantly different from the GC profile observed in sequences with a metazoan hit (Fig. 3). With average and median GC values of 43% and 40%, respectively, the GC profile of unknown sequences tended to be similar to the one of sequences with a metazoan match. In contrast, the GC content of sequences with uninformative hits showed a similar profile to the one of nonmetazoan sequences (Fig. 3). Comparing the GC composition among several organisms, we observed that the *Hydractinia* metazoan sequences clustered in the range of 39–42% of GC content with the GC profiles of the *Hydra* and *Nematostella* EST datasets as well as with the *Caenorhabditis elegans* cDNAs. In contrast, among *Hydractinia*'s nonmetazoan consensus sequences, the GC content extended from the 39–42% range to include the GC percentage observed in bacteria such as *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis* [23–26] (Fig. S1).

Characteristics of the *Hydractinia* transcriptome

Using TBLASTX, the translated *Hydractinia* sequences were compared with the translated cDNAs of different vertebrate and invertebrate model organisms. We observed that 153 consensus sequences were by a factor of 10^{10} more closely related to their vertebrate orthologues than to their invertebrate orthologues. In contrast, only 18 sequences appeared to be more similar to invertebrate sequences using the same criteria (Fig. 4). Indeed, we detected 28 consensus sequences with a vertebrate homologue but without any hit in the invertebrate datasets, whereas four *Hydractinia* sequences were found only in invertebrates (Table S2).

Unique sequences of *Hydractinia*

In an attempt to detect genes present in the *Hydractinia* transcriptome but absent in other cnidarians, we compared the *Hydractinia* sequences using TBLASTX with the sets of ESTs of *Acropora millepora*, *Hydra* spp. and *Nematostella vectensis*, as well as the genomic DNA data of *Nematostella*. With an E-value $< 10^{-3}$ and excluding all ESTs related to a nonmetazoan

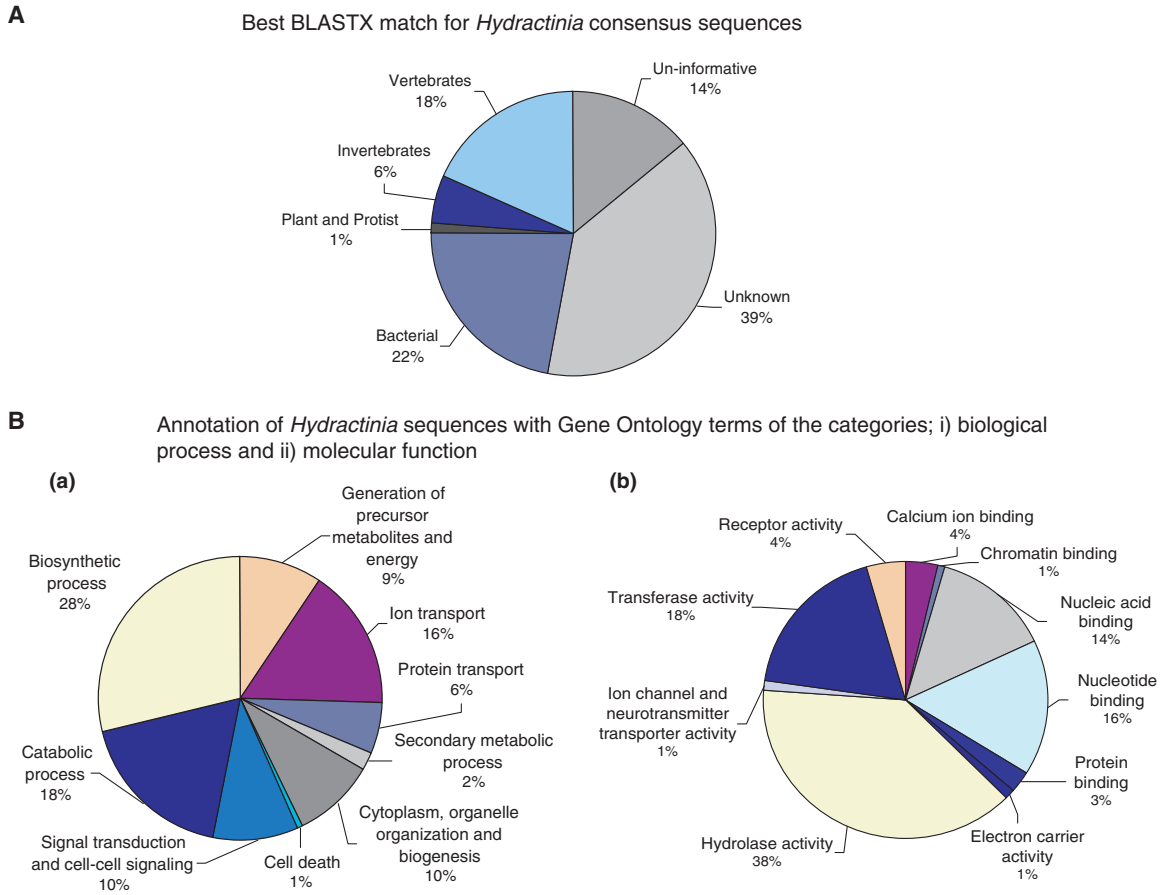

Fig. 2. Diversity of the *Hydractinia* ESTs. (A) Distribution of the *Hydractinia* ESTs according to their best matches to specific organism groups, together with the percentage of sequences without any significant hit. (B) Distribution of the ESTs into the GO functional categories (a) biological process and (b) molecular function.

sequence, we detected 23 unique *Hydractinia* sequences with a known protein or protein domain hit (Table 1). Some sequences pointed to the same protein domain hit. However, analysis by specialized BLAST algorithms, such as BL2SEQ (data not shown), revealed that these sequences do not have a significant sequence similarity with one another. This is supported by the fact that they were not clustered in the sequence analysis pipeline. With regard to consensus sequences that have a nonmetazoan match, 393 sequences were uniquely present in the *Hydractinia* dataset, and 36 of them were annotated by protein domain analyses.

The few cnidarians that are being used as model systems differ markedly in many aspects of their biology, morphology and life history. Cnidarians are solitary or colonial species, living in a freshwater environment or are marine organisms. In addition, these species have different stem cell systems, reproduce asexually or sexually, and inhabit different ecological niches. Taking as working examples marine versus freshwater cnidarians and solitary versus colonial cnidarians, we analysed the cnidarian datasets to find genes that are unique to two different combinations of cnidarians as follows: (a) *Hydra* and *Nematostella* are solitary polyps, whereas *Acropora* and *Hydractinia* are colonial; (b) *Hydra* is a freshwater organism, whereas *Hydractinia*, *Nematostella* and *Acropora* are marine animals. In order to identify genes linked to these traits, using the TBLASTX algorithm we extracted all *Hydractinia* sequences shared with *Acropora* and *Nematostella* but not with *Hydra*, as well as all sequences present in *Hydractinia* and *Acropora* but missing in the *Hydra* and *Nematostella* datasets. Using the same significance criteria as above (E-values $< 10^{-3}$), 11 *Hydractinia* sequences, shared by *Acropora* and *Nematostella*, were absent in *Hydra*. The sequences are mainly related to metabolism, including catalytic activities, protein modification, protein-mediated transport, physiological processes and

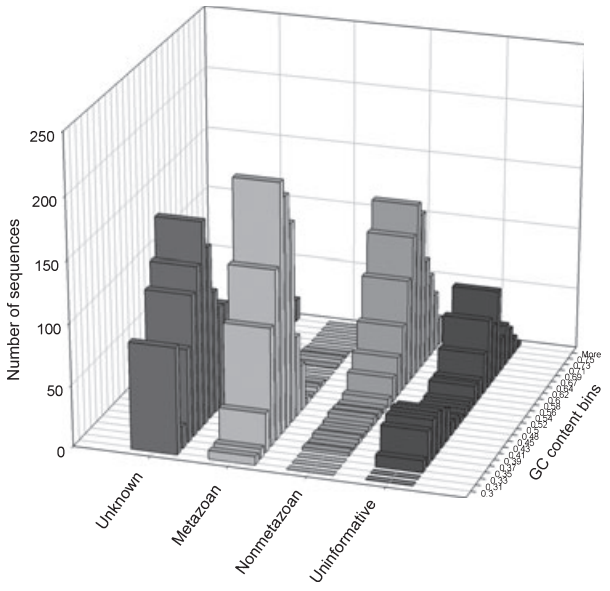

Fig. 3. Histogram of the GC profile of the *Hydractinia* consensus sequences. Only sequences with more than 100 bp were considered for the analysis. The ESTs were subclustered with BLASTX into metazoan, nonmetazoan, uninformative and unknown group of sequences. Their GC content was calculated using the software COMPOSITION. The GC content of metazoan sequences (median GC value 39%) was significantly ($P < 0.05$) different from that of nonmetazoan sequences (median GC value 63%). Unknown and uninformative sequences presented median GC values of 40% and 60%, respectively.

signal transduction (Table 2, Table S3). In the second analysis, 15 sequences were uniquely found in *Hydractinia* and *Acropora*. These sequences are associated with metabolism, nucleotide binding and signal transduction functions, and one was related to an intracellular non-membrane-bound organelle (Table 2, Table S3).

Hydractinia database

A database was created in order to optimize the handling of all generated data, including the physical information of each EST clone, the results of the EST clustering, the representative consensus sequences and the BLAST programs. Searches within the database can be carried out using GenBank identification numbers, clones or consensus sequence names, etc. It is possible to query simultaneously different fields by combining search criteria with 'AND' and 'OR'. Query results are listed on screen, with direct links to the detailed clone or sequence information, which can be easily extracted for further analysis. The *Hydractinia* EST database can be accessed at http://www.mchips.org/hydractinia_echinata.html

Discussion

The quality of EST collections depends on the selection of the RNA sources employed for the generation

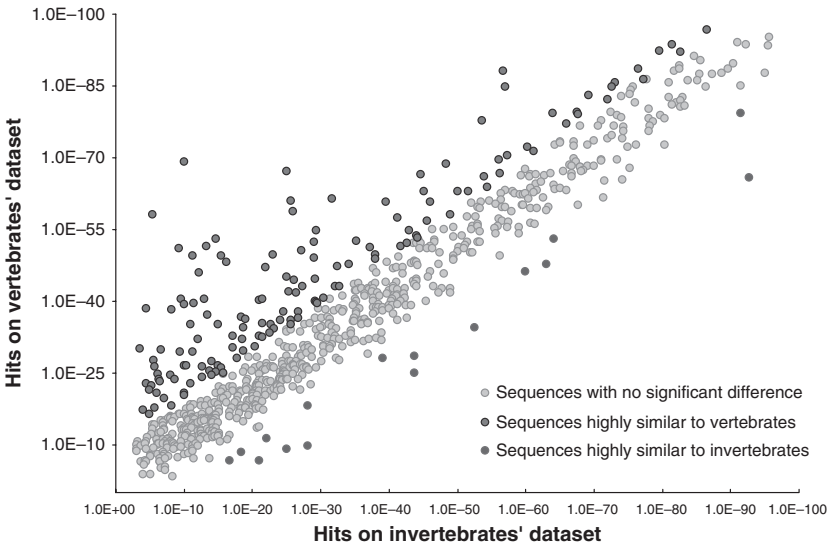

Fig. 4. *Hydractinia* consensus sequence best hits on the invertebrate and vertebrate cDNA datasets. Only sequences showing a TBLASTX hit with a confidence E-value between 10^{-3} and 10^{-100} were included in the plot. Sequence comparisons were made against the vertebrate cDNA datasets of *Macaca mulatta*, *Canis familiaris*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Xenopus tropicalis* and the invertebrate cDNA datasets of *Aedes aegypti*, *Anopheles gambiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. The difference between the E-values was considered significant when sequences exhibited 10^{10}-fold more similarity to one of the datasets. Sequences with only a vertebrate or invertebrate homologue, as well as those with lower E-values ($< 10^{-100}$) are not shown.

Table 1. *Hydractinia echinata* unique sequences with known annotation. Sequence annotation was carried out with BLAST or DOMAINSWEEP using the Swiss-Prot/TrEMBL and InterPro domain databases, respectively.

Clone name	Sequence GenBank identification	Protein match identification number at GenBank/InterPro	Sequence/domain annotation
HEAB-0027M01	68411965	IPR008412	Bone sialoprotein II
HEAB-0034N17	74135604	IPR002952	Eggshell protein
HEAB-0036J11	74132951	IPR001876	Zinc finger, RanBP2-type
HEAB-0038D19	74134674	IPR005649	Chorion 2
HEAB-0038H17	74134662	IPR006706	Extensin-like region
HEAB-0039H23	74134110	IPR005649	Chorion 2
HEAB-0040M05	74134400	IPR003908	Galanin 3 receptor
HEAB-0042M23	74134684	IPR001841	Zinc finger, RING-type
tah96a10	49453351	IPR006706	Extensin-like region
tah98e04	49451948	IPR002952	Eggshell protein
tah99a03	49453544	IPR007087	Zinc finger, C2H2-type
tai01f07	50347174	gi: 62510506	CHCH5_HUMAN
tai01g09	50347183	IPR006706	Extensin-like region
tai08h10	50351274	IPR000637	HMG-I and HMG-Y, DNA binding
tai10f09	50348080	IPR007087	Zinc finger, C2H2-type
tai21h03	50351781	IPR005649	Chorion 2
tai32e08	50351456	IPR001152	Thymosin beta-4
tai35e09	50352319	IPR010800	Glycine-rich
tai46c12	50697716	IPR007223	Peroxin 13, N-terminal
tam53h06	59829660	IPR007718	SRP40, C-terminal
tam54c10	59829689	IPR002952	Eggshell protein
tam55f08	59829784	IPR006706	Extensin-like region
tam57a05	59829876	IPR007223	Peroxin 13, N-terminal

of the cDNA library. In standard libraries, it is difficult to discover rarely expressed genes. The yield in gene discovery can be increased by in-depth sequencing or by broadening the diversity of source materials [27,28]. In the case of *Hydractinia*, its complex life cycle provides a broad spectrum of temporally and spatially regulated genes. To obtain a more complete representation of the transcriptome, as well as access to *Hydractinia*-specific genes, RNA extracted from different developmental stages and induction experiments was pooled and used for the construction of the cDNA library. Using this approach, the information related to gene expression at any particular stage was lost, but all life stages were covered and the chance to include particular transcripts in the library was increased. Despite having a nonnormalized library, EST clustering resulted in 60% of the ESTs being singletons or grouped in clusters of two to five sequences (Fig. 1). Only relatively few ESTs were highly redundant. They mainly correspond to housekeeping genes. The 3808 consensus sequences generated by the fragment assembly system (FAS) may be considered as an overestimation of the real number of unique transcripts isolated. EST end-sequencing does not usually retrieve the complete cDNA sequence of a clone. This complicates assembly and clustering, which may result in different consensus sequences (contigs) representing the same gene.

On the other hand, it is also possible to have an under-representation of the real number of unique sequences because of members of closely related gene families [28]. With the availability of genome data, it might be possible to test and improve the EST assembly, but this information has not been generated as yet for *Hydractinia* [29]. However, the quality of the assembly was assessed in two different ways. At the nucleotide level, a BLASTN comparison of the consensus sequences to all *Hydractinia* ESTs corroborated the physical clustering carried out by the FAS programs (data not shown). At the protein level, a BLASTX comparison to different protein databases revealed a redundancy of 1.6% in all consensus sequences with a significant hit. These sequences represent different parts of genes and therefore could not be clustered by FAS because of a lack of overlapping sequences. Most of these genes encode ribosomal, actin and lectin proteins, or proteins involved in an enzymatic activity.

As expected, a significant number of sequences could not be annotated and were considered to be unknown or with an inconsistent description (Fig. 2A). Analyses of these sequences revealed a low average sequence length of $\sim$ 300 bp, with a median at 160 bp. Thus, it

Table 2. *Hydractinia* sequences compared with those of other cnidarians model organisms. Sequences were annotated with BLAST and DOMAINSWEEP using the Swiss-Prot/TrEMBL and InterPro domain databases. In addition, the sequences were annotated with GO terms from the two main categories: biological process and molecular function. For a detailed description of the GO terms, see Table S3. Not applicable (n/a) was considered when sequences had no significant match to domain, Swiss-Prot/TrEMBL or GO databases.

Clone name	GenBank identification	Sequence/domain annotation	E-value	GO: biological process	GO: molecular function
(A) *Hydractinia* protein sequences present in *Acropora* and *Nematostella* but not in *Hydra*					
HEAB-0029E05	74134839	Lanin A-related sequence 1 protein	1E-16	GO:0007582	n/a
HEAB-0029J09	74133868	Nuclear protein 1 (p8)	4E-08	n/a	n/a
HEAB-0038N23	74134624	MKIAA0230 protein (fragment)	1E-41	n/a	GO:0004601
tai09b01	50352378	Guanine nucleotide-binding protein Y-e subunit precursor	2E-09	GO:0008277	GO:0004871
tai11f02	50348136	Malate synthase	1E-91	GO:0008152	GO:0004474
tai11g12	50348149	Lysosomal thioesterase ppt2 precursor	2E-45	GO:0006464	GO:0016787
tai20d03	50351692	AP-4 complex subunit sigma-1	2E-08	GO:0016192	n/a
tai33g08	50352245	Isocitrate lyase	2E-72	GO:0008152	GO:0016829
tam56f07	59829849	Cephalosporin hydroxylase family protein	1E-08	n/a	n/a
HEAB-0023B24	68411515	Unknown function	n/a	GO:0005975	GO:0004033
tam53d11	59829628	Unknown function	n/a	n/a	n/a
(B) *Hydractinia* protein sequences present in *Acropora* but not in *Nematostella* and *Hydra*					
HEAB-0020F05	68411267	2-c-methyl-d-erythritol 4-phosphate cytidylyltransferase	1E-24	n/a	GO:0008299
HEAB-0024D20	68411599	Response regulator receiver protein	6E-09	n/a	GO:0000166
HEAB-0028A08	68334384	Major facilitator superfamily MFS_1	1E-38	n/a	n/a
HEAB-0028B20	68334404	Fatty-acid desaturase. 2/2007	2E-16	n/a	n/a
HEAB-0037F13	74133658	PcaB-like protein. 2/2007	1E-94	n/a	GO:0016829
HEAB-0039G08	74134978	Signal peptidase I precursor (EC)	2E-24	n/a	GO:0000155
HEAB-0042I20	74133750	Glucose-methanol-choline oxidoreductase, N-terminal	n/a	n/a	n/a
HEAB-0020L20	68411323	Unknown function	n/a	n/a	GO:0005884
HEAB-0026O12	68411824	Unknown function	n/a	n/a	n/a
HEAB-0029G01	74134845	Unknown function	n/a	n/a	n/a
HEAB-0036O10	74133537	Unknown function	n/a	GO:0006810	GO:0000166
HEAB-0042L12	74133375	Unknown function	n/a	n/a	n/a
tai07g10	50350972	Unknown function	n/a	n/a	n/a
tai16a08	50352144	Unknown function	n/a	n/a	n/a
tai40g01	50697024	Unknown function	n/a	n/a	n/a

is reasonable to assume that the majority of these sequences do not represent a cDNA insert, but correspond mainly to the 3′ noncoding region of genes [12]. In contrast, sequences with a positive match in the protein databases had an average and median length of 639 and 629 bp, respectively. A better characterization of these sequences was possible as more than 60% of the reads corresponded to ORFs. The inclusion of a protein domain annotation step allowed the characterization of 55% of the *Hydractinia* consensus sequences.

The program GOPET, which can perform an organism-independent GO annotation [30], revealed a broad range of functions and processes in the *Hydractinia* dataset (Fig. 2B). GO classification correlated with the BLAST gene product predictions can be used to assess the accuracy and quality of the sequence annotation. Improvements in the functional annotation of *Hydractinia* genes may be reached with a larger number of EST reads. This may allow the generation of longer consensus sequences that represent nearly the complete coding sequences and provide more accurate annotations [31]. In addition, the ongoing cnidarian sequencing projects, as well as the improvements in the GO annotation of other organisms, will provide better platforms for sequence comparisons [1,3].

One other possible explanation for the sequences without a BLAST hit is that they could be cnidarian or even smaller taxon-specific genes (i.e. absent even from *Hydra* and *Nematostella*). These taxon-specific genes may either be the result of the conservation of an ancient gene, lost in all other animals, or evolutionary novelties. For example, cnidarians possess many unique features, such as their stinging cells, known as nematocytes or cnidocytes, which are not found in any other group of animals. These orphan sequences, and particularly those with an ORF, deserve special attention and further detailed analysis.

A significant fraction of the *Hydractinia* consensus sequences corresponded to nonmetazoan hits in the protein databases (Fig. 2A). The majority are related to bacterial proteins with a GC content that was significantly higher than the amount of GC observed in sequences with a metazoan match (Fig. 3). Therefore, on the basis of GC content, the annotated *Hydractinia* EST dataset seems to contain two physically different kinds of sequence. This was confirmed by comparing the GC profiles of the *Hydractinia* sequences with those observed in other organisms, including bacteria, cnidarians, invertebrates and vertebrates (Fig. S1) [23–26]. In the case of sequences without a functional annotation, the broad range of GC percentage suggests that some of them may have a GC composition characteristic of bacterial sequences. However, for the group of unknown sequences, the majority exhibited a low GC percentage, suggesting a higher relationship to metazoan proteins than to bacterial proteins. In contrast, most of the sequences with uninformative terms seem to have a bacterial GC profile. This is to be expected, as several bacterial annotations on the protein databases contain uninformative terms (Fig. 3).

To obtain the expression profile of *Hydractinia*, the RNA pool used for the cDNA library construction was supplemented with RNA extracted from adult tissues that may have carried commensal micro-organisms. We took every experimental precaution to ensure a low level of contamination in our dataset, including the starvation of the adult organisms before RNA isolation and a two-step poly(dT) nucleic acid purification of the RNA prior to cDNA library construction. Together with the characteristics of the sequencing reads described above, it is possible to suggest that many of these nonmetazoan sequences did not originate from a bacterial contamination. Poly A+ selection and oligo dT priming used for mRNA isolation and cDNA construction, respectively, do not rule out the capture of poly A+ tracts that are not located at the 3′-end of RNA sequences. However, the chance that a large number of bacterial sequences with a high GC content are captured by poly(dT) is relatively low.

Hydractinia sequences with a bacterial hit could be divided into two different groups. The first group consists of 487 sequences, which were also found in the ESTs of the *Acropora*, *Hydra* and/or in the *Nematostella* genome. Approximately two-thirds of them might be present in the genome of *Hydractinia*, as 331 sequences were identified in the genome of *Nematostella*. The presence of these sequences in cnidarians may therefore predate the Anthozoa–Hydrozoa divergence. In accordance with the analyses carried out by Technau *et al.* [3] on *Acropora* and *Nematostella*, we also found nonmetazoan sequences containing introns (data not shown) and sequences with homologues in diverse organisms. This favours the hypothesis of an ancient common origin for the majority of these sequences and argues against recent lateral gene transfer events [3,20,21]. However, almost half of the sequences exhibited a best match to a particular class of bacteria (*Pseudomonas* spp.). Thus, it is possible to speculate that some of the sequences appeared in cnidarians by ancient lateral gene transfer events or that the transferred sequences were subsequently lost in other animal lines. Lateral gene transfer events are difficult to prove, and there is no evidence for large-scale sequence transfers into animal genomes. For a satisfactory explanation, it is necessary to access the genome data of *Hydractinia*.

The second group consists of 357 sequences with a bacterial hit and no counterparts in other cnidarians. It is possible to consider them as unique *Hydractinia* sequences, taking into account the suggested substantial variation in gene content within the cnidarians [1]. An alternative explanation might be the inclusion of adult material in the cDNA library. This may have resulted in the discovery of expressed genes related to an adult condition, for example genes related to nutrition or reproduction, which could not be detected in the other EST projects carried out using embryos. The majority of these nonmetazoan sequences were related to enzymatic activities. Nevertheless, for all these *Hydractinia* bacterial-like sequences, especially those without a clear genomic cnidarian representation, the possibility of symbiotic, parasitic or commensal bacterial sources cannot be ruled out. Commensal or epiphytic microbes are common in adult cnidarians as well as in higher metazoans [19,32–34].

Hydractinia homology analyses against 12 different bilaterian model organisms revealed a substantial number of ESTs with a significantly higher sequence similarity to vertebrate sequences rather than to their fly, mosquito or nematode counterparts. This tendency of homology is clearly shown in Fig. 4 for more than 150 sequences. Moreover, we found 28 sequences with only vertebrate homologues. Thus, despite having a small dataset, the *Hydractinia* ESTs do not only corroborate the hypothesis of cnidarian ancestral genetic complexity, but also provide more examples of gene loss or secondary sequence modification in ecdysozoans [1–3,7]. In contrast, fewer sequences had a higher similarity or were even uniquely identified in the invertebrates analysed. Apparently, we are also faced with genes that have been lost or are highly diverged in vertebrates.

One of the objectives of the generation of *Hydractinia* ESTs is to complement the information obtained from other cnidarian genome projects, identifying the

genes maintained or added during cnidarian evolution. Comparing the *Hydractinia* ESTs with all other available cnidarian datasets, we identified a list of 23 unique *Hydractinia* genes with known protein domain architectures (Table 1). Despite the fact that some genes shared protein domains, their sequences did not overlap and were considered unique *Hydractinia* sequences. Examples of these are the six sequences showing a chorion or eggshell protein domain. These families of proteins are associated with a tissue- and temporal-specific gene expression pattern in ovaries, and are highly conserved in evolution [35]. Their presence in our cDNA library may result from the inclusion of sexually mature female colonies in the mRNA pool, rather than being *Hydractinia* specific. Some of the putative proteins identified are unexpected and their functions are hard to interpret at present. For example, we found a sequence homologous to the vertebrate bone sialoprotein, which is associated with bone mineralization and remodelling [36]. Another example is the Galanin receptor. In vertebrates, this receptor is expressed in the peripheral and central nervous system, activating K^+ channels by coupling G proteins [37]. In addition, several sequences without a BLAST hit appeared to be unique to *Hydractinia*, for which there are two possible interpretations. First, as previously described, it is expected that several of these sequences are short ORFs or noncoding sequences, resulting in poor matching by BLAST. This holds true not only for the *Hydractinia* ESTs in question, but also for the other cnidarian EST databases that were used for comparison. Second, we may reconsider that the differences between the transcriptomes of anthozoans and hydrozoans point to extensive divergence of these taxa. This implies large genetic differences and gene family diversity within the Cnidaria [1]. Indeed, there are marked differences in cnidarian morphology and physiology. In an attempt to extract genes that might be related to such differences, a comparison of the databases resulted in a list of sequences that are probably linked to either physiological requirements due to the environment (e.g. sea or freshwater) or the colonial phenotype displayed by *Hydractinia* and *Acropora*. Despite the fact that most of the sequences identified in the first analysis showed an enzymatic (reductase, hydrolase) activity, which may correspond to the regulation of intracellular osmolarity, it is not possible to satisfactorily conclude that there is a direct relationship between these sequences and such physiological functions. The same holds true for the *Hydractinia* sequences shared only with *Acropora*. As most of these sequences are unknown or associated with a diverse functionality, it is not possible to establish a firm link to colonial growth using only the bioinformatics tools currently available. However, we consider such a link a working hypothesis for further analyses towards the characterization of cnidarian diversity and the identification of particular genes involved, for example, in the allogeneic reactions of colonial organisms.

This EST project is the first high-throughput sequencing carried out in a colonial marine hydroid. With the support of a database harbouring all the acquired information, the project provides a platform to promote and facilitate molecular research, not only in *Hydractinia*, but also in other cnidarians. The *Hydractinia* ESTs confirmed the remarkable genetic complexity of cnidarians and reinforces the present view that a substantial number of ancient prokaryotic genes have been maintained in the cnidarian genome but are lost from other metazoans [1–3]. This view may be obscured by some level of contamination, which cannot be ruled out at present. However, the quality measures applied suggest to us that many of the nonmetazoan sequences are genuine. The detection of genes specific to *Hydractinia* or genes that might be associated with the different morphological and physiological conditions offered by cnidarians shows that the cnidarians analysed to date do not represent all the features offered by the phylum. Therefore, a complete picture of the genomic diversity of the Cnidaria will only be possible when sequence data from more basal metazoans are available. In addition, ongoing genome projects in other organisms (e.g. sponges, chaetognath or lophotrochozoans) will help to reconstruct the genetic repertoire of the common metazoan ancestor and provide further insight into the maintenance, loss or divergence of genes in the vertebrates [1,3,9,10,38].

To improve the functional characterization of the *Hydractinia* sequences, the bioinformatics approach will soon be combined with array technology. For this, we have created a microarray comprising the most representative cDNA sequences for each of the 3808 generated EST clusters, as well as 5000 randomly picked, unsequenced cDNAs. Gene expression profiling may provide a straightforward approach for new insights into the functional evolution of ancient genes.

Materials and methods

Animal culture

Hydractinia mature colonies grown on glass slides were cultured as described previously [15]. Fertilized eggs were collected almost daily and maintained in sterile artificial seawater (ASW). Embryos and the subsequent larvae were raised for up to 5 days. Metamorphosis-competent larvae

were induced to metamorphose on glass slides by 3 h incubation at 18 °C with 116 mM CsCl (Sigma-Aldrich, Munich, Germany) in seawater, osmotically corrected to 980 mosmol. Primary polyps were examined regularly under the dissecting microscope, and polyps showing abnormal morphology or slow growth rates were removed.

RNA isolation

RNA was extracted from different developmental stages, as well as organisms subjected to induction experiments. Subsequently, all RNA samples were pooled (Table S4) and used for library construction. Prior to any RNA isolation, animals were starved for up to 2 days. Ten different developmental stages were included: early embryos at 1–5 h postfertilization, gastrulating embryos at 24 h postfertilization, preplanula and planula larvae at 2 and 3 days postfertilization, respectively, metamorphosing animals at 3, 16, 28 and 72 h postmetamorphosis induction with CsCl and finally mature female and male colonies.

Five different types of induction experiment were performed. (a) Heat shock treatment: primary polyps were incubated for 30 min at 30 °C, washed with ASW and incubated for 1 h at 18 °C before RNA isolation. (b) Osmotic shock treatment: mature colonies were incubated for 1 h at a salinity of 1.7%, then washed with ASW and incubated for 1 h at normal salinity (3.5%) before RNA isolation. (c) Regeneration treatment: polyps were cut and incisions were made in the stolon mat of an adult colony. After 3 h of recovery, RNA was isolated. (d) Lipopolysaccharide treatment: animals were exposed to 100 $\mu g \cdot mL^{-1}$ lipopolysaccharide (Sigma-Aldrich) for 1 h and washed several times. RNA extraction was carried out after 1 h of incubation in ASW. (e) Allorecognition experiment: genetically distinct adult animals were allowed to grow into contact with each other. Following the first signs of rejection, RNA was isolated from only the contact area.

In all cases, total RNA was isolated using acid guanidinium thiocyanate [39]. The quality and quantity of the material were assessed by 1.2% formaldehyde (Sigma-Aldrich) agarose gels and spectrophotometer readings.

cDNA library

Poly A+ RNA was isolated from 224 μg of pooled total RNA using the Dynabeads mRNA purification kit (Invitrogen, Karlsruhe, Germany). The oligo-dT-primed cDNA library was constructed from 2.2 μg of poly A+ RNA. For cDNA synthesis, the SuperScript Plasmid System for cDNA and Cloning (Invitrogen) was used following the manufacturer's protocols. The cDNAs of the largest fractions obtained in the fractionation steps were directionally ligated into the plasmid vector pSPORT1 and electroporated into ElectroMAX™ DH10B T1 phage-resistant cells (Invitrogen) using an *Escherichia coli* transporator (BTX Harvard Apparatus, Holliston, MA, USA). After plating on agar, colonies with inserts were picked by the Qpix robot (Genetix, München-Dornach, Germany) and transferred into 384-well microplates (Genetix). Each well had previously been filled with 50 μL 2YT/HFMF freezing media containing 100 $\mu g \cdot mL^{-1}$ carbenicillin (Carl Roth, Karlsruhe, Germany). After overnight incubation at 37 °C, the arrayed library was stored at −80 °C.

EST sequencing and sequence analysis pipeline

Single-pass cDNA sequencing from 5′- and/or 3′-ends was conducted at the Washington University Genome Sequencing Center (http://genome.wustl.edu/). After the removal of vector and ambiguous regions from the raw sequence data, the sequence reads were uploaded to the EST database at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The first step in the sequencing analysis pipeline was a download of the sequences in FASTA format. Subsequently, the Wisconsin GCG package (Accelrys, Cambridge, UK) FAS available at the Heidelberg Unix Sequence Analysis Resource (HUSAR) (http://genome.dkfz-heidelberg.de/) was initialized. Within FAS, the GEL package programs were used, starting with the assembly project (GELSTART), uploading the sequences in GCG format (GELENTER), aligning them into contigs (GELMERGE), editing the assembled contigs (GELASSEMBLE), displaying contig structures (GELVIEW) and finally evaluating the created FAS database with respect to quality and statistics (GELSTATUS and GELANALYZE). The generated consensus sequences were used as a query for BLAST homology searches against GenBank databases [40].

Annotation and subsequent analysis of the *Hydractinia* sequences

At the DNA level, searches were made against the NCBI nonredundant nucleotide database using the BLASTN algorithm with default parameters. In case of insignificant hits, searches were performed against the GenBank EST databases. At the protein level, analyses were carried out using BLASTX against the SwissProtPlus database under the sequence retrieval system [41] at HUSAR, which includes the latest full releases of both Swiss-Prot and TrEMBL [42]. Matches with an E-value acceptance threshold of $< 10^{-6}$ were retrieved from the results page and stored on our local server. Sequences without any significant annotation or with an uninformative hit, e.g. hypothetical, probable, putative or chromosomal annotation, were further analysed using DOMAINSWEEP [43], which allows the identification of domain architectures within a protein sequence. A positive match was only considered when the sequence contained at least two domain hits described in two protein family databases that are members of the same InterPro family/domain, or when there were two blocks or motifs in a correct order

already described in the Prints or Blocks dataset. Further functional annotations were made by adding GO terms to the sequences using GOPET available at HUSAR [30]. Only hits above a confidence threshold of 80% were annotated with GO terms of the two main categories: biological process and molecular function. In subsequent analysis, the consensus sequences were compared with TBLASTX to different databases that were downloaded into HUSAR from NCBI, Ensembl Genome Browser (http://www.ensembl.org/index.html), and from the Joint Genome Institute. For TBLASTX analysis, significant hits were considered when matches presented an E-value acceptance threshold of $< 10^{-3}$. The downloaded databases included: the cnidarian EST databases of *Acropora*, *Hydra* and *Nematostella*, as well as the raw and assembled genome data of *Nematostella*; the new releases of vertebrate cDNA datasets of *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Canis familiaris*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Xenopus tropicalis*; and the ecdysozoan invertebrate cDNA datasets of *Aedes aegypti*, *Anopheles gambiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*.

Hydractinia database

All relevant information about every EST, as well as the information generated in the sequence analysis pipeline, was automatically integrated into a database using in-house scripts. The database is a POSTGRESQL relational database (http://www.postgresql.org/). For an easy-to-use platform, a web interface was created using PERL/CGI. It can be accessed at http://www.mchips.org/hydractinia_echinata.html.

Accession numbers

The ESTs analysed in the current study (8798) have been deposited in GenBank with the following accession numbers: CO370602–CO370702; CO372032–CO372389; CO535098–CO535606; CO538780–CO540606; CO720032–CO720932; DN135280–DN136094; DN602415–DN602701; DN604185–DN604200; DR433357–DR434207 and DT621337–DT624248.

Acknowledgements

We thank Achim Stephan for valuable technical assistance; Yuki Katzukura for help with the library construction; Mareike Weers for support in database design; Werner A. Müller and Lindsay Murrells for reviewing the manuscript; Brahim Mali and Stephan Wiemann for critical discussions. We also thank Sandra Clifton, Rick Wilson and the GSC EST sequencing group at the University of Washington in St Louis for carrying out the sequencing of the ESTs. This work was supported by a grant from the National Science Foundation (IBN-IOB-0120591) to HRB. JSR was supported by the German Cancer Research Center PhD programme.

References

1 Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94.

2 Kortschak RD, Samuel G, Saint R & Miller DJ (2003) EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Curr Biol* **13**, 2190–2195.

3 Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, Grasso LC, Hayward DC, Sensen CW, Saint R, Holstein TW *et al.* (2005) Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. *Trends Genet* **21**, 633–639.

4 Miller DJ, Ball EE & Technau U (2005) Cnidarians and ancestral genetic complexity in the animal kingdom. *Trends Genet* **21**, 536–539.

5 Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D & Finnerty JR (2006) StellaBase: the *Nematostella vectensis* genomics database. *Nucleic Acids Res* **34**, D495–D499.

6 Guder C, Philipp I, Lengfeld T, Watanabe H, Hobmayer B & Holstein TW (2006) The Wnt code: cnidarians signal the way. *Oncogene* **25**, 7450–7460.

7 Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ *et al.* (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156–160.

8 Sullivan JC, Reitzel AM & Finnerty JR (2008) Upgrades to StellaBase facilitate medical and genetic studies on the starlet sea anemone, *Nematostella vectensis*. *Nucleic Acids Res* **36**, D607–D611.

9 Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, Agata K & Bosch TC (2007) The innate immune repertoire in cnidaria – ancestral complexity and stochastic gene loss. *Genome Biol* **8**, R59.

10 Miller DJ & Ball EE (2008) Cryptic complexity captured: the *Nematostella* genome reveals its secrets. *Trends Genet* **24**, 1–4.

11 Sullivan JC & Finnerty JR (2007) A surprising abundance of human disease genes in a simple ''basal'' animal, the starlet sea anemone (*Nematostella vectensis*). *Genome* **50**, 689–692.

12 Adams MD, Soares MB, Kerlavage AR, Fields C & Venter JC (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* **4**, 373–380.

13 Galliot B & Schmid V (2002) Cnidarians as a model system for understanding evolution and regeneration. *Int J Dev Biol* **46**, 39–48.
14 Cadavid LF, Powell AE, Nicotra ML, Moreno M & Buss LW (2004) An invertebrate histocompatibility complex. *Genetics* **167**, 357–365.
15 Frank U, Leitz T & Muller WA (2001) The hydroid *Hydractinia*: a versatile, informative cnidarian representative. *BioEssays* **23**, 963–971.
16 Muller WA (2002) Autoaggressive, multi-headed and other mutant phenotypes in *Hydractinia echinata* (Cnidaria: Hydrozoa). *Int J Dev Biol* **46**, 1023–1033.
17 Seipp S, Schmich J, Kehrwald T & Leitz T (2007) Metamorphosis of *Hydractinia echinata* – natural versus artificial induction and developmental plasticity. *Dev Genes Evol* **217**, 385–394.
18 Seipp S, Schmich J & Leitz T (2001) Apoptosis – a death-inducing mechanism tightly linked with morphogenesis in *Hydractina echinata* (Cnidaria, Hydrozoa). *Development* **128**, 4891–4898.
19 Rosenberg E, Koren O, Reshef L, Efrony R & Zilber-Rosenberg I (2007) The role of microorganisms in coral health, disease and evolution. *Nat Rev Microbiol* **5**, 355–362.
20 Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C & Brown JR (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944.
21 Steele RE, Hampson SE, Stover NA, Kibler DF & Bode HR (2004) Probable horizontal transfer of a gene between a protist and a cnidarian. *Curr Biol* **14**, R298–R299.
22 Habetha M & Bosch TC (2005) Symbiotic *Hydra* express a plant-like peroxidase gene during oogenesis. *J Exp Biol* **208**, 2157–2165.
23 Xu HX, Kawamura Y, Li N, Zhao L, Li TM, Li ZY, Shu S & Ezaki T (2000) A rapid method for determining the G+C content of bacterial chromosomes by monitoring fluorescence intensity during DNA denaturation in a capillary tube. *Int J Syst Evol Microbiol* **50** Pt 4, 1463–1469.
24 Wang HC, Badger J, Kearney P & Li M (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. *Mol Biol Evol* **18**, 792–800.
25 Belle EM, Duret L, Galtier N & Eyre-Walker A (2004) The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J Mol Evol* **58**, 653–660.
26 Lobry JR & Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biol* **3**, RESEARCH0058.
27 Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M & Hayashizaki Y (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res* **10**, 1617–1630.
28 Forment J, Gadea J, Huerta L, Abizanda L, Agusti J, Alamar S, Alos E, Andres F, Arribas R, Beltran JP *et al.* (2005) Development of a citrus genome-wide EST collection and cDNA microarray as resources for genomic studies. *Plant Mol Biol* **57**, 375–391.
29 Jain M, Shrager J, Harris EH, Halbrook R, Grossman AR, Hauser C & Vallon O (2007) EST assembly supported by a draft genome sequence: an analysis of the *Chlamydomonas reinhardtii* transcriptome. *Nucleic Acids Res* **35**, 2074–2083.
30 Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S & Konig R (2006) GOPET: a tool for automated predictions of gene ontology terms. *BMC Bioinformatics* **7**, 161.
31 Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB & Robinson GE (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res* **12**, 555–566.
32 Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950–955.
33 Kuo J, Chen MC, Lin CH & Fang LS (2004) Comparative gene expression in the symbiotic and aposymbiotic *Aiptasia pulchella* by expressed sequence tag analysis. *Biochem Biophys Res Commun* **318**, 176–186.
34 Merle PL, Sabourault C, Richier S, Allemand D & Furla P (2007) Catalase characterization and implication in bleaching of a symbiotic sea anemone. *Free Radic Biol Med* **42**, 236–246.
35 Orr-Weaver TL (1991) *Drosophila* chorion genes: cracking the eggshell's secrets. *BioEssays* **13**, 97–105.
36 Sasaguri K, Ganss B, Sodek J & Chen JK (2000) Expression of bone sialoprotein in mineralized tissues of tooth and bone and in buccal-pouch carcinomas of Syrian golden hamsters. *Arch Oral Biol* **45**, 551–562.
37 Branchek TA, Smith KE, Gerald C & Walker MW (2000) Galanin receptor subtypes. *Trends Pharmacol Sci* **21**, 109–117.
38 Marletaz F, Gilles A, Caubit X, Perez Y, Dossat C, Samain S, Gyapay G, Wincker P & Le Parco Y (2008) Chaetognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biol* **9**, R94.
39 Chomczynski P & Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162**, 156–159.
40 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped

BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

41 Etzold T, Ulyanov A & Argos P (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* **266**, 114–128.

42 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370.

43 del Val C, Ernst P, Falkenhahn M, Fladerer C, Glatting KH, Suhai S & Hotz-Wagenblatt A (2007) ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ. *Nucleic Acids Res* **35**, W444–W450.

Supporting information

The following supplementary material is available:

Fig. S1. Comparison of the GC profile of *Hydractinia* ESTs with those of other model organisms.

Table S1. Annotation of *Hydractinia* uninformative and unknown sequences using DOMAINSWEEP.

Table S2. *Hydractinia* sequences shared with either vertebrates or invertebrates.

Table S3. GO annotation of *Hydractinia* sequences shared with other cnidarians.

Table S4. *Hydractinia*'s RNA pooling strategy.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.