Demosponge EST Sequencing Reveals a Complex Genetic Toolkit of the Simplest Metazoans

Matija Harcet,*[,1] Maša Roller,[2] Helena Ćetković,[1] Drago Perina,[1] Matthias Wiens,[3] Werner E.G. Müller,[3] and Kristian Vlahovíček*[,2,4]

[1]Department of Molecular Biology, Rudjer Boskovic Institute, Zagreb, Croatia

[2]Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

[3]Institute for Physiological Chemistry and Pathobiochemistry, Johannes Gutenberg University, Medical School, Mainz, Germany

[4]Department of Informatics, University of Oslo, Oslo, Norway

***Corresponding author:** E-mail: mharcet@irb.hr; kristian@bioinfo.hr.

Associate editor: Billie Swalla

Research article

Abstract

Sponges (Porifera) are among the simplest living and the earliest branching metazoans. They hold a pivotal role for studying genome evolution of the entire metazoan branch, both as an outgroup to Eumetazoa and as the closest branching phylum to the common ancestor of all multicellular animals (Urmetazoa). In order to assess the transcription inventory of sponges, we sequenced expressed sequence tag libraries of two demosponge species, *Suberites domuncula* and *Lubomirskia baicalensis*, and systematically analyzed the assembled sponge transcripts against their homologs from complete proteomes of six well-characterized metazoans—*Nematostella vectensis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, and *Homo sapiens*. We show that even the earliest metazoan species already have strikingly complex genomes in terms of gene content and functional repertoire and that the rich gene repertoire existed even before the emergence of true tissues, therefore further emphasizing the importance of gene loss and spatio-temporal changes in regulation of gene expression in shaping the metazoan genomes. Our findings further indicate that sponge and human genes generally show similarity levels higher than expected from their respective positions in metazoan phylogeny, providing direct evidence for slow rate of evolution in both "basal" and "apical" metazoan genome lineages. We propose that the ancestor of all metazoans had already had an unusually complex genome, thereby shifting the origins of genome complexity from Urbilateria to Urmetazoa.

Key words: metazoan evolution, comparative genomics, genome complexity, *Suberites domuncula*, *Lubomirskia baicalensis*.

Introduction

Some of the fundamental points of interest in animal evolution are the historical and phylogenetic origins of genome complexity, genetic origins of germ layers, and the relation of the species' morphological characteristics to the amount and variability of genetic information. The view that simple animals have simple genomes and that genome complexity should increase proportionally with phenotypic complexity is rapidly fading with insights gained from sequence data of basal metazoan species (Steele 2005). Some of the earlier work on cnidarians (Kortschak et al. 2003; Kusserow et al. 2005; Miller et al. 2005; Matus et al. 2006) offered glimpses into the unexpectedly diverse gene pool of the simplest eumetazoans—animals defined by the presence of true tissues usually originating from all three germ layers. Complete genome sequence of the starlet sea anemone *Nematostella vectensis* further showed that much of the genomic complexity in terms of gene content and structure was already present in the common ancestor of all Eumetazoa (Putnam et al. 2007; Hui et al. 2008). One of the few branches of multicellular animals that does not belong to Eumetazoa and is located at the base of the monophyletic tree (Wainright et al. 1993; Muller 1995) of the kingdom Animalia is the phylum Porifera—sponges (fig. 1).

Sponges are, by all standards, living fossils. They are the simplest extant and probably the earliest branching metazoan phylum with a known fossil record dating back at least 580 My, prior to the Cambrian explosion (Li et al. 1998). Their ancient origin and basal position in the animal kingdom make them an important subject for metazoan genome evolution studies. Sponges are one of the two phyla within the Parazoa group, characterized by the lack of true tissues, organs or organic systems, and with simple embryonic development (Ereskovsky and Dondua 2006). However, despite their simple morphology and basal position in the metazoan phylogeny, indications exist that sponges harbor a number of genes found in deuterostomes but missing in protostomes. For example, in our previous analyses, we found evidence of protein kinases (BtkSD) and a GTPase, previously thought to exist only in deuterostomes (Cetkovic, Muller et al. 2004; Harcet et al. 2005; Cetkovic et al. 2007). Several gene families that demonstrate ancient duplications and diversifications have also been documented in sponges (Hoshiyama et al. 1998;

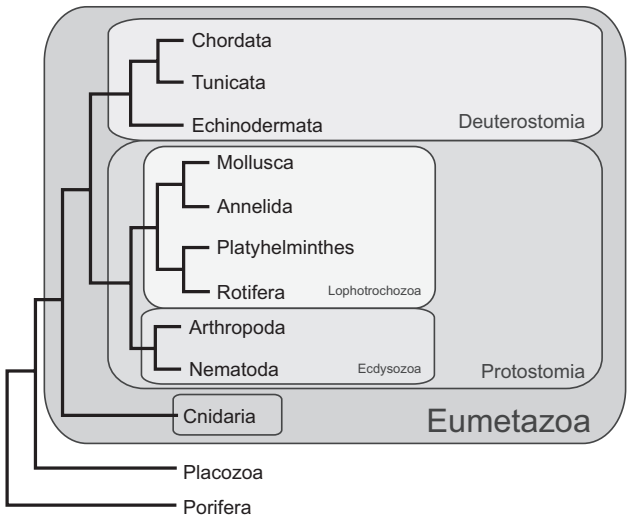

Fig. 1. Phylogenetic relationships within the metazoan kingdom.

Ono et al. 1999; Suga, Koyanagi et al. 1999; Suga, Ono et al. 1999; Nichols et al. 2006; Suga et al. 2008). The recent availability of raw sequencing reads from the *Amphimedon queenslandica* sequencing project provided evidence for the existence of even more genes in sponges—most notably the homeobox (Wiens, Batel et al. 2003; Wiens, Mangoni et al. 2003; Larroux et al. 2007), Wnt (Adamska et al. 2007; Lapebie et al. 2009), and several other transcription factors (Larroux et al. 2008), pushing the origin of key metazoan developmental genes and pathways back to the very root of Metazoa (Tessmar-Raible and Arendt 2005; Arendt 2008; Philippe et al. 2009). Albeit shown on a limited set of sequenced genes, sponge proteins were predominantly found to be more similar, in terms of sequence similarity and gene architecture, to their vertebrate than worm (Gamulin et al. 2000) and fruit fly orthologs (Perina et al. 2006; Cetkovic et al. 2007). However, as of yet no systematic analysis of sponge gene inventory has been performed. In order to evaluate genetic complexity of sponges on a larger scale, we employed the random expressed sequence tags (ESTs) sequencing approach on two demosponge species from different habitats—the marine *Suberites domuncula* and the freshwater *Lubomirskia baicalensis*. Our objective was to determine the presence, as well as the degree of similarity and functional characteristics, of the assembled sponge transcript homologs in complete genomes of six well-characterized metazoan organisms.

We performed comparative genomics analysis on two separate sets of 4,646 unique *S. domuncula* and 1,335 unique *L. baicalensis* transcripts, assembled from two independent single-pass random EST sequencing runs. Apart from different habitats where the two demosponge species were collected, we sampled cells in different developmental stages, further extending the range of transcribed genes included in the final EST library. We searched for sponge protein homologs within a comprehensive nonredundant proteome database of six metazoan organisms with available complete genomes: cnidarian *N. vectensis* (starlet sea anemone), nematode *Caenorhabditis elegans* (worm), arthropod *Drosophila melanogaster* (fruit fly), echinoderm *Strongylocentrotus purpuratus* (purple sea urchin), urochordate *Ciona intestinalis* (sea squirt), and vertebrate *Homo sapiens* (human). Results obtained with the *L. baicalensis* data set, although on a smaller sample, reiterate findings drawn from the *S. domuncula* analysis and are, for the purpose of brevity, presented in the supplementary supporting information (SI) (Supplementary Material online).

This paper presents the first step toward the systematic elucidation of the transcriptional inventory of sponges, which will in turn help infer the complexity of the Urmetazoa genome, and provide an indication of genome dynamics across the entire metazoan lineage.

Methods

Background information on sponges, sequencing protocols and the outline of the analysis with detailed description of methods and procedures, as well as the full description of the analysis pipeline are described in the supplementary SI (Supplementary Material online). Here, we briefly outline the key steps in EST sequencing and bioinformatic analysis.

Both sponge cDNA libraries were randomly sequenced (see supplementary SI, Supplementary Material online) resulting in 13,384 *S. domuncula* and 2,573 *L. baicalensis* EST transcript sequences, respectively. Reads were organized into separate databases and processed independently. ESTs were cleaned from sequence contaminants (e.g., vectors) and from poly-A and poly-T tails and assembled using the CAP3 Sequence Assembly Program (Huang and Madan 1999) for a final yield of 4,646 *S. domuncula* and 1,335 *L. baicalensis* assembled transcripts longer than 100 bp.

Sponge transcripts were compared using BlastX (no sequence filtering and a default *E* value cutoff of 10) against the STRING extended ortholog database v6.3 (von Mering et al. 2003) and assigned a COG/KOG category based on three-nearest neighbor consensus rule (category is assigned if the three best matches [smallest *E* value] for each query sequence originate from the same orthologous group, i.e., have the same COG ID).

We constructed a proteome database of six metazoan species with complete genomes by acquiring Ensembl proteomes of nematode, fruit fly, sea squirt, and human. Starlet sea anemone and sea urchin proteomes were obtained from NCBI GenBank. Additionally, we obtained from NCBI nematode, fruit fly, sea squirt, and human proteins not found in Ensembl data sets. Final database contained a total of 176,973 nonredundant protein sequences.

We searched the proteome database with *S. domuncula* and *L. baicalensis* ESTs by BlastX with cutoff levels at 1×10^{-5} and 1×10^{-40}. For each query sponge transcript, single best match per proteome was selected (up to six subject sequences) and multiply aligned using Muscle together with the translated sponge sequence. Ortologies were confirmed with reciprocal BlastT hits at the same cutoff.

Pathway reconstitution was performed by running a pairwise sequence search against the KEGG-curated set

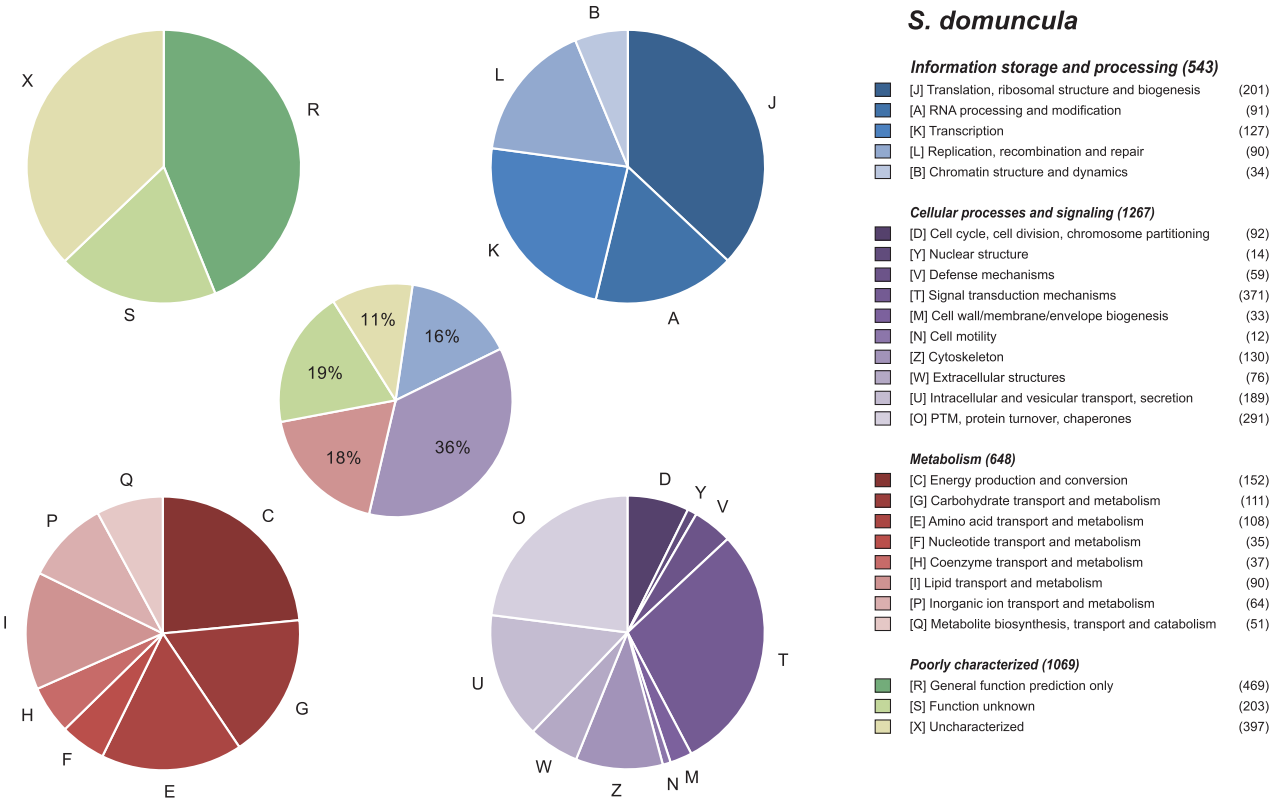

Fig. 2. Functional characterization of *Suberites domuncula* transcripts. A total of 3,077 transcripts were classified into KOG/COG categories giving rise to a total of 3,130 class assignments (some COGs belong to more than one class). Distribution over functional classes is given in the central pie chart, with each super category slice broken down into separate pie charts in the corners (Poorly characterized and uncharacterized function categories [R and S] are combined with uncharacterized category [X] in a separate pie chart the top left corner). Overall class distribution follows that of other metazoan genomes, with most abundant functions in signal transduction, protein turnover, translation, and transcription. Functional categorization for *Lubomirskia baicalensis* is shown in the supplementary SI, Supplementary Material online.

of human proteins (Kanehisa et al. 2008) and mapping the percent identity of the alignment to KEGG metabolic and signaling pathways with MADNet (Segota et al. 2008).

Results and Discussion

Functional Characterization

We successfully classified 3,077 (66%) *S. domuncula* and 814 (61%) *L. baicalensis* transcripts using the STRING database (von Mering et al. 2003) and a stringent assignment process (discussed in the Methods section). The graphical distribution of functional classes for *S. domuncula* is given in figure 2; *L. baicalensis* functional characterization is presented in supplementary table S1 (Supplementary Material online). Distribution of functional classes is consistent with that of both human and fruit fly complete proteomes (Tatusov et al. 2003), with most abundant categories in processes of signal transduction (T), translation (J), and protein turnover (O), indicating the adequate coverage of the sequenced EST libraries, even in the case of *L. baicalensis*.

Presence of Homologs

In order to minimize false-positive matches, all similarity searches were performed at two *E* value cutoff levels—less and more stringent (1×10^{-5} and 1×10^{-40}, respectively). With the less stringent cutoff, of 4,646 unique *S. domuncula* transcripts, 3,290 (~71%) showed a positive match to proteins from one or more species in our database (tabulated results for each transcript are presented in the supplementary SI, supplementary table S3, Supplementary Material online). *Lubomirskia baicalensis* results had slightly lower hit count—791 of 1,335 (~60%; supplementary table S4, Supplementary Material online). Most sponge transcript homologs originate from the sea anemone and, surprisingly, human proteomes. The sea urchin is ranked third by the number of hits, whereas the urochordate *Ciona* has significantly fewer hits. Both protostomes, the fruit fly and particularly nematode, also have far fewer hits than the human and sea anemone and are ranked last.

The exclusive matches (i.e., sponge homologs present in only one of six proteomes) follow the same trend of hit counts: the sea anemone followed by human and sea urchin. The sea squirt, fruit fly, and nematode have drastically less exclusive matches. The general tendency of homolog presence across lineages is even more apparent if we group exclusive *S. domuncula* homologs into higher order taxonomies, shown in figure 3. Apart from the sea anemone—the single diploblast representative with the highest number of exclusive hits—an unexpectedly high number of sponge gene homologs are found only in the

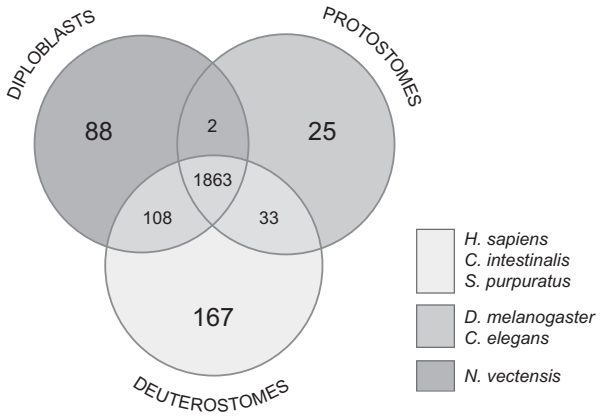

Fig. 3. Venn diagram of *Suberites domuncula* transcript homologs across taxonomic groups—diploblasts (*Nematostella vectensis*), protostomes (*Drosophila Melanogaster* and *Caenorhabditis elegans*), and deuterostomes (*Strongylocentrotus purpuratus*, *Ciona intestinalis*, and *Homo sapiens*). Most sponge transcripts (1,863) were found in all three groups. However, the largest set of exclusive homologies is found within the deuterostomes (167, nearly 10%), demonstrating the breadth of a sponge gene repertoire.

three deuterostomes; the sea urchin, *Ciona*, and human cumulative count is 167, of which 55 (33%) are found exclusively in the human proteome. Figure 3 also shows that 1,863 *S. domuncula* transcripts (~57% of 3,290 genes with at least one homology) are shared within all six phyla. Detailed breakdown of sponge homolog presence across six phyla is presented in supplementary SI and supplementary figure S2 (Supplementary Material online).

The control search results with a more stringent E value cutoff level of $1 \times 10_{-40}$ follow the same trend of homolog presence, with exactly the same ranking of matching organisms. In fact, the results tend to be more robust in terms of the relative increase in the number of exclusive hits to the human proteome (33 of 1,424 vs. 55 of 3,290).

The closest relatives of metazoans are unicellular choanoflagellates. A recent report on the sequenced genome of choanoflagellate *Monosiga brevicollis* estimates the gene count at ~9,200 (King et al. 2008). Although the *Monosiga* genome shows evidence of cell adhesion and signaling protein domains needed for transition to multicellularity, previously thought to be exclusively metazoan, the total gene count amounts to only half the number found in the sea anemone genome. Moreover, a preliminary scan against the *S. domuncula* EST data set reveals 1,140 genes, mostly involved in the signaling processes, present in sponge, and either missing or significantly divergent in the *M. brevicollis* genome (supplementary SI and supplementary table S5, Supplementary Material online) leading to a conclusion that the choanoflagellate genome is not nearly as complex as any known metazoan genome neither in terms of gene number nor repertoire. If we use the missing gene count to assess the size of the *S. domuncula* transcriptome, we can arrive at a conservative estimate of ~12,000 genes—again suggesting that a large gene and module explosion event occurred in the metazoan ancestor. This is in turn consistent with the characteristics of the recently sequenced *Trichoplax* genome (~12,000 genes), postulated to have branched off after the sponges (Srivastava et al. 2008).

It could be argued that our homolog presence results may be biased by differences in quality of annotation and completeness of the compared organisms' genomes/proteomes. However, there is no correlation between the protein count per species in our database and the number of best Blast hits to sponge proteins per compared organism (supplementary SI and supplementary fig. S2, Supplementary Material online), especially in the protostome domain where extensive gene loss has been previously documented (Ogura et al. 2005; Hui et al. 2009). This signifies that Blast hits largely are true homologs. Moreover, the apparent overrepresentation of human proteome in the entire data set originates primarily in the fact that many proteins are present with several (highly redundant but not identical) transcript variants, whereas only a single variant was selected as the best match.

No lophotrochozoan complete genomes were, to date, available for inclusion into our database. However, we have compared our EST sequences with several incompletely sequenced or insufficiently annotated Lophotrochozoan genomes or EST data sets. The results, albeit must be considered inconclusive, are in accordance with our findings regarding the richness of the sponge genome repertoire (supplementary table S6, Supplementary Material online).

Our findings not only support previous conclusions about genome complexity dynamics across metazoan lineages (Dehal et al. 2002; Kortschak et al. 2003; Sodergren et al. 2006) but also more importantly show that sponges, the simplest and oldest extant animal phylum, also have highly complex genomes with gene content similar to that of cnidarians and vertebrates. This in turn demonstrates that there is low correlation between gene repertoire and morphological complexity even without considering the emergence of true tissues and a variety of cell types—rather, we place the origins of genome complexity to a gene accumulation process at the base of the metazoan tree of life.

Sequence Conservation

In order to compare the rates of sequence change between different metazoan lineages, we determined the extent of sponge transcript similarity to their respective homologs in six metazoan species. Distributions of sponge transcripts according to the count of the highest similarity homologs are shown in figure 4. The majority of sponge proteins most closely match the sea anemone proteome, whereas only slightly fewer are, again surprisingly, most similar to human proteins. The sea urchin is ranked third, whereas the sea squirt, fruit fly, and especially nematode are drastically underrepresented in terms of best-matching homologs. The results further support our finding that besides gene repertoire, the sequence divergence (i.e., the sequence distance) is also highest in lineages leading to the nematode, fruit fly, and sea squirt. A detailed demonstration of how sponge

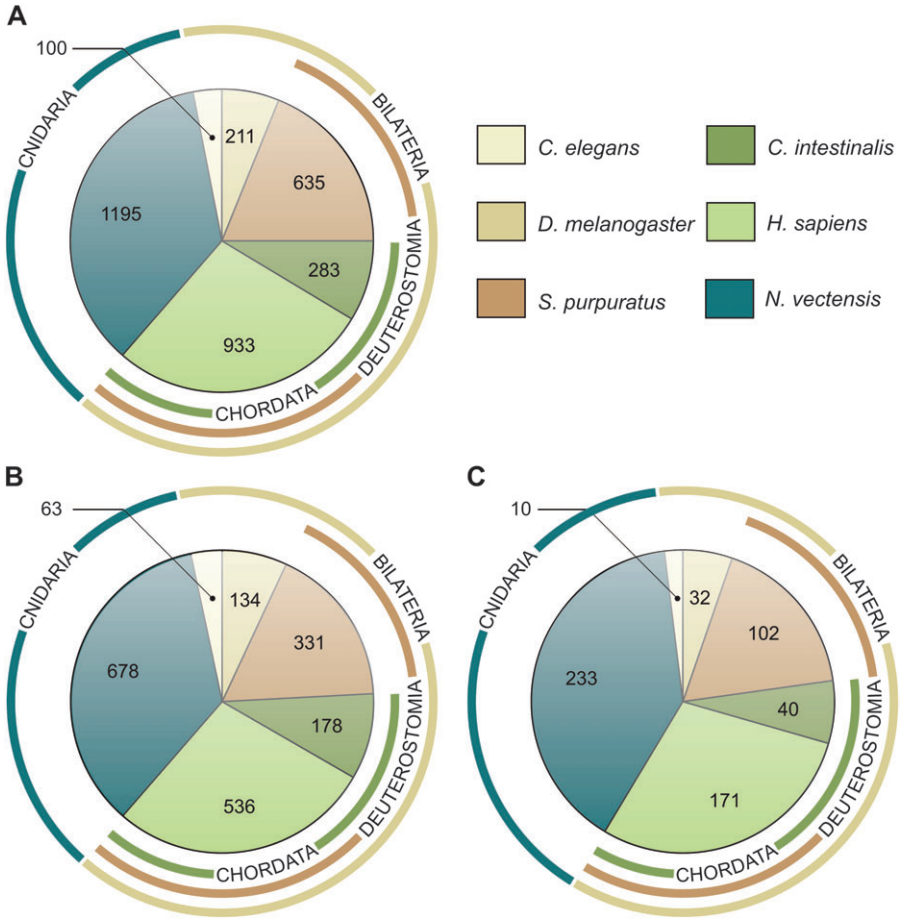

Fig. 4. Distribution of *Suberites domuncula* transcripts according to the closest matching homolog across six phyla on (*A*) a total set of found homologies (i.e., any one of six proteomes matched a sponge transcript); (*B*) with hits present in all six phyla at less stringent cutoff; and (*C*) more stringent cutoff values. Surprisingly, regardless of the comparison method and threshold, human proteins show highest similarity to almost 30% sponge transcripts. This ranks human similarities second best, right next to sea anemone.

proteins are related to the six proteomes is shown in figure 5, where we quantified the relative sequence distance between each sponge transcript and a corresponding set of homologs from three species in our database. If we consider that some of these homologs are not whole transcript matches but rather domain or fragment similarities, by using the multiple alignment approach (see Methods and supplementary SI, Supplementary Material online), it is still evident that even at the level of protein modules there is an unusual degree of similarity between sponge and human coding sequences. This implicates a slow evolutionary rate in both sponge and human genomes that cannot fully be attributed either to possible long generation time in sponges or the low population count in humans (fig. 6*A*). As a consequence, we can speculate that the two genomes generally may be very similar (at least at the level of protein-coding sequence) to a metazoan ancestor.

Enrichment of Clade-Specific Functions

We subsequently performed the analysis of functional gene category (according to the STRING/COG classification) enrichment across six phyla based on similarity to sponge transcripts. Sponge transcripts were subdivided within each functional class according to the organism where the best hit is found (Table 1), and count frequencies were tested for statistically significant deviation patterns from the overall functional distribution. Interestingly, the signal transduction category (T) showed high bias toward human homologs and away from the cnidarians, suggesting that the signaling machinery conservatively propagated throughout the entire metazoan lineage, sharing most features (i.e., the 'metazoan signaling toolkit' [Erwin 2009]) with higher vertebrates, whereas cnidarians significantly diverged either by gene loss or by sequence divergence. On the other hand, the translation and ribosome biogenesis processes show the opposite trend, with increasing divergence from lower to higher metazoans.

Signaling Pathway Reconstruction

Another demonstration of the increase in the functional toolkit with the transition to metazoans is the identification of modules required for most metazoan signaling cascades (figs. 6*B* and *C*). By comparing *S. domuncula* transcripts with human proteins involved in signaling pathways and cell adhesion processes, we were able to demonstrate the presence of equivalent functional elements

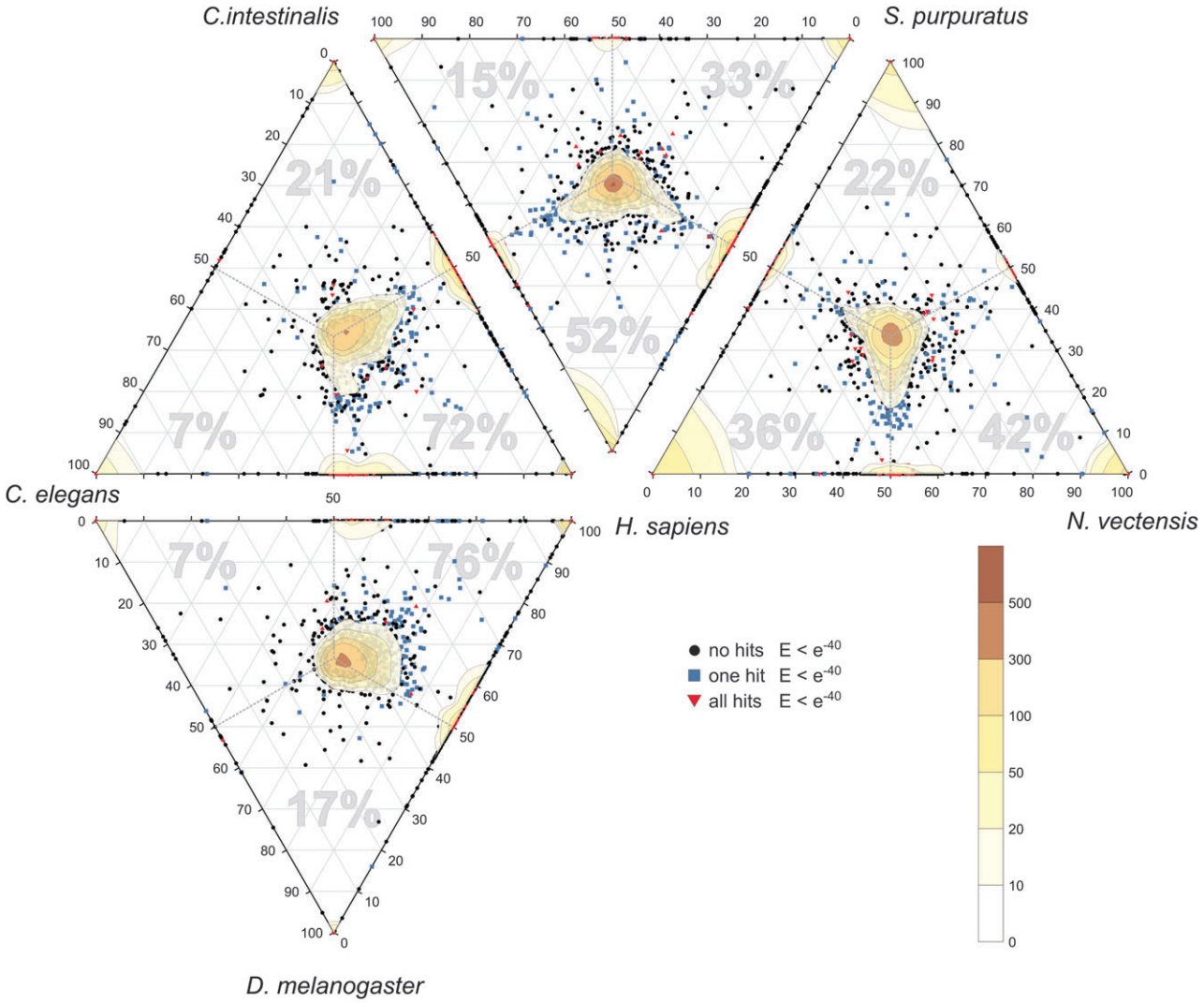

Fig. 5. *Suberites domuncula* gene similarity profiles. Each symbol (dot, square, and triangle) in a ternary plot represents a single sponge transcript, whereas the position of a symbol represents the relative sequence similarity to the three phyla, calculated from normalized pairwise scores derived from a multiple alignment with all available homologs from all six phyla. Transcripts that are equally similar to proteins from all three phyla will tend to move toward the center of the triangle, whereas those found near corners suggest a higher similarity to a single phylum. Symbols in corners represent transcripts that are exclusively found only in a single phylum, whereas symbols on triangle sides denote transcripts with one missing hit (to the phylum in the opposite corner). Overlaying contour map represents symbol density estimate and is provided for clarity. General tendency for genes to migrate toward the human corner is apparent in all four plots (the difference between human and starlet sea anemone is statistically insignificant).

sufficient to reconstitute key processes in signaling and cell adhesion pathways. Some of the domains and modules have been identified with low similarity to their human homologs and will need direct experimental validation of their precise roles and mechanisms. However, we argue that the increased available repertoire of metazoan-only functional modules may have alleviated and increased the combinatorial potential of the domain shuffling processes suggested by King et al. (2008) and have diversified elementary adhesion functions (mostly performed through cadherin domains in *M. brevicollis*) into cellular signaling cascades.

Other Genome Characteristics

Data about other characteristics of the sponge genome, such as gene structure or synteny, are scarce. Published research (Gamulin et al. 1997; Muller et al. 2002; Cetkovic, Grebenjuk et al. 2004) only indicates that sponge genes usually resemble their vertebrate homologs with respect to the intron counts and conserved splice site positions. Similar findings were reported for cnidarians (Putnam et al. 2007), annelids (Raible et al. 2005), and echinoderms (Sodergren et al. 2006). Generally, there seems to be a positive correlation between gene repertoire and other features of genome complexity among metazoans. Therefore, we can

→

the percent identity of the sponge transcript (or a fragment thereof) match to a human protein. Many of the identified similarities, especially in the low-identity range, are shared domains of a human multidomain protein. Legend for panels (*B*) and (*C*) shown at the bottom.

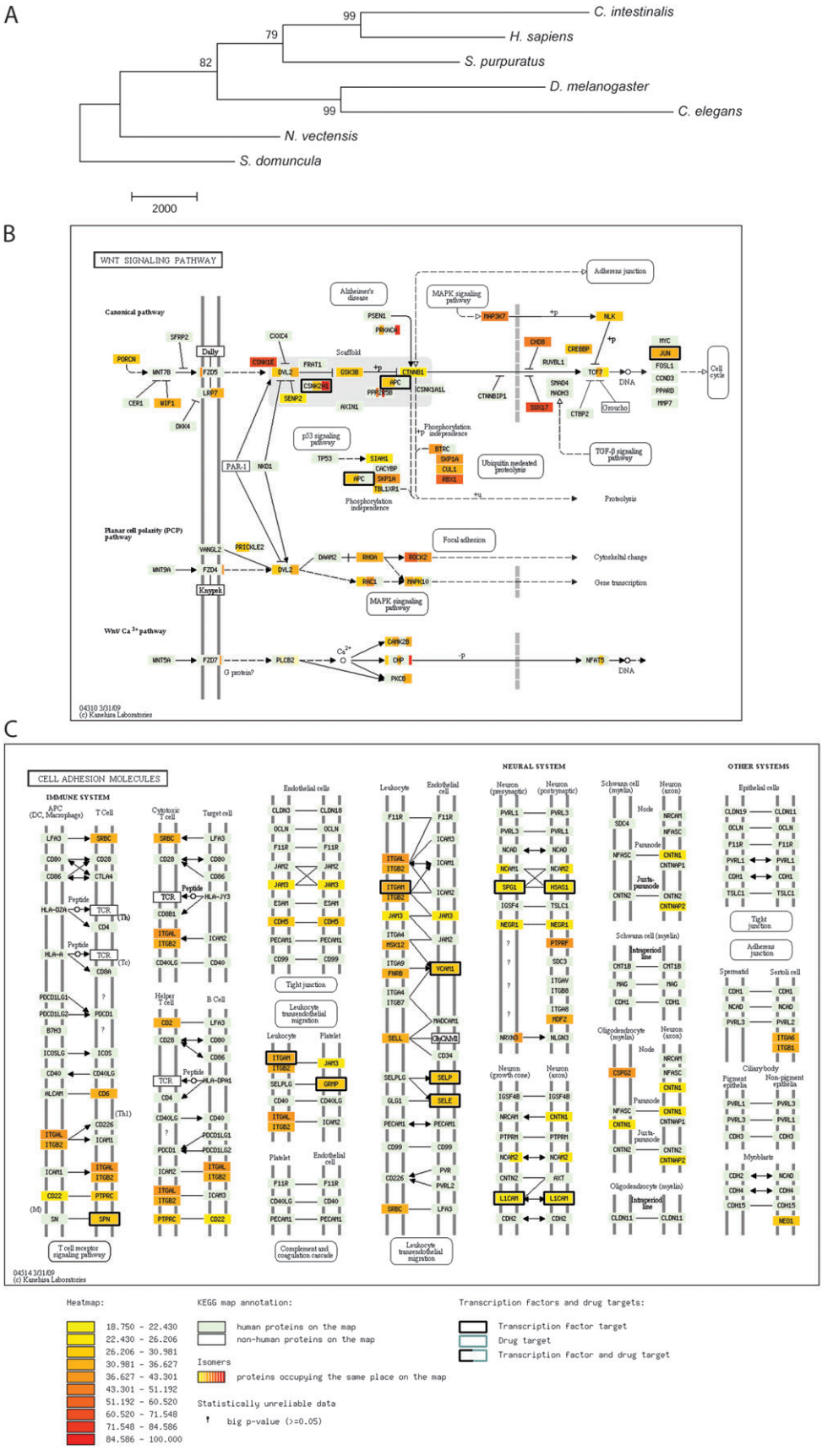

FIG. 6. Maximum parsimony phylogenetic tree (*A*) based on concatenated sequences of proteins involved in signaling pathways. Homologs with *E* values of 1×10^{-20} or less in at least one organism were chosen for the analysis. Bootstrap values based on 1,000 replicates are shown on nodes. Wnt signaling pathway modules (*B*) and cell adhesion modules (*C*) found through sequence similarity with equivalent human homologs and mapped to respective standard KEGG pathways (http://www.genome.jp/kegg/pathway.html). The intensity ranging from yellow to red denotes

Table 1. Enrichment of Functional Categories among Highest Scoring Matches to *Suberites domuncula* Transcripts.

N. vectensis		*C. intestinalis*		*D. melanogaster*		*H. sapiens*		*S. purpuratus*		*C. elegans*		total			COG category
92	8,7%	9	3,7%	9	5,7%	64	6,9%	17	3,1%	2	2,5%	193	6,4%	INFORMATION STORAGE AND PROCESSING	[J] Translation, ribosomal structure and biogenesis
30	2,8%	4	1,6%	6	3,8%	29	3,1%	13	2,4%	3	3,8%	85	2,8%		[A] RNA processing and modification
31	2,9%	10	4,1%	6	3,8%	43	4,6%	20	3,6%	5	6,2%	115	3,8%		[K] Transcription
19	1,8%	7	2,9%	2	1,3%	30	3,2%	19	3,4%	2	2,5%	79	2,6%		[L] Replication, recombination and repair
9	0,8%	4	1,6%	0	0,0%	9	1,0%	7	1,3%	1	1,2%	30	1,0%		[B] Chromatin structure and dynamics
28	2,6%	6	2,5%	6	3,8%	25	2,7%	19	3,4%	1	1,2%	85	2,8%	CELLULAR PROCESSES AND SIGNALING	[D] Cell cycle control, cell division, chromosome partitioning
3	0,3%	0	0,0%	0	0,0%	4	0,4%	1	0,2%	0	0,0%	8	0,3%		[Y] Nuclear structure
12	1,1%	10	4,1%	4	2,5%	12	1,3%	1	0,2%	0	0,0%	39	1,3%		[V] Defense mechanisms
90	8,5%	32	13,1%	18	11,4%	121	13,1%	50	9,1%	10	12,5%	321	10,6%		[T] Signal transduction mechanisms
5	0,5%	1	0,4%	0	0,0%	8	0,9%	10	1,8%	0	0,0%	24	0,8%		[M] Cell wall/membrane/envelope biogenesis
2	0,2%	1	0,4%	0	0,0%	1	0,1%	5	0,9%	0	0,0%	9	0,3%		[N] Cell motility
39	3,7%	12	4,9%	7	4,4%	40	4,3%	21	3,8%	0	0,0%	119	3,9%		[Z] Cytoskeleton
24	2,3%	6	2,5%	3	1,9%	17	1,8%	3	0,5%	3	3,8%	56	1,9%		[W] Extracellular structures
54	5,1%	12	4,9%	11	7,0%	51	5,5%	41	7,4%	7	8,8%	176	5,8%		[U] Intracellular trafficking, secretion, and vesicular transport
87	8,2%	22	9,0%	15	9,5%	75	8,1%	57	10,3%	10	12,5%	266	8,8%		[O] Posttranslational modification, protein turnover, chaperones
62	5,9%	4	1,6%	11	7,0%	35	3,8%	27	4,9%	0	0,0%	139	4,6%	METABOLISM	[C] Energy production and conversion
28	2,6%	11	4,5%	8	5,1%	19	2,1%	21	3,8%	4	5,0%	91	3,0%		[G] Carbohydrate transport and metabolism
36	3,4%	7	2,9%	7	4,4%	21	2,3%	14	2,5%	4	5,0%	89	2,9%		[E] Amino acid transport and metabolism
18	1,7%	2	0,8%	2	1,3%	6	0,6%	4	0,7%	0	0,0%	32	1,1%		[F] Nucleotide transport and metabolism
16	1,5%	1	0,4%	0	0,0%	6	0,6%	6	1,1%	0	0,0%	29	1,0%		[H] Coenzyme transport and metabolism
27	2,5%	10	4,1%	4	2,5%	30	3,2%	11	2,0%	0	0,0%	82	2,7%		[I] Lipid transport and metabolism
14	1,3%	2	0,8%	1	0,6%	16	1,7%	11	2,0%	2	2,5%	46	1,5%		[P] Inorganic ion transport and metabolism
18	1,7%	3	1,2%	0	0,0%	17	1,8%	11	2,0%	0	0,0%	49	1,6%		[Q] Secondary metabolites biosynthesis, transport and catabolism
147	13,9%	33	13,5%	24	15,2%	117	12,6%	76	13,8%	17	21,2%	414	13,7%	POORLY CHARACTERIZED	[R] General function prediction only
73	6,9%	8	3,3%	6	3,8%	56	6,0%	28	5,1%	3	3,8%	174	5,8%		[S] Function unknown
95	9,0%	27	11,1%	8	5,1%	74	8,0%	58	10,5%	6	7,5%	268	8,9%		[X] Uncharacterized
1059		244		158		926		551		80		3018			

Functionally classified sponge transcripts were assigned to the closest-matching organism (a total of 3,018 COG-assigned transcripts were found with matches to our database). Frequencies of occurrence of hits in each functional category to a specific proteome were then tested for statistically significant overrepresentation and underrepresentation compared with the overall functional distribution in sponge transcripts (column "total") using binomial test and corrected for false discovery rate (FDR) over six organisms. Significant enrichment (FDR corrected P value under 0.025) is marked in red shades, whereas significant depletion is colored blue. The shade indicates relative level of significance (lower P values for richer colors). Arrows pointing upwards and downwards denote enrichment tendency compared with the total (in 1/3 percentiles), whereas arrows pointing to the right indicate no significant tendency in the enrichment. Most interesting find is that the S. domuncula transcripts responsible for signal transduction (T) show elevated levels of similarity to human proteins and decreased similarity to sea anemone proteins, suggesting a high degree of conservation across the entire metazoan lineage

anticipate that the sponge genome is most similar to cnidarian and vertebrate genomes in synteny and intron characteristics (e.g., density and conservation profile). The expected sequence of the first complete sponge genome *A. queenslandica* will eventually serve as final evidence. However, sponges are a diverse group, and data from classes Hexactinellida and Calcarea should also significantly contribute to our understanding of metazoan genome evolution, shedding the final light at the origins of gene complexity that lead to development of multicellular life. There is an ongoing debate on the molecular phylogeny aspects of basal metazoans (Dohrmann et al. 2008; Srivastava et al. 2008; Philippe et al. 2009; Sperling et al. 2009), and although we did not address this issue directly, we hope that the data provided in this paper will provide further evidence for understanding the complex relations between Porifera, Placozoa, and Eumetazoa.

Conclusions

In this systematic analysis of the sponge gene repertoire, we show that the genomic complexity, at least in terms of gene content, was already present at the very beginning of animal evolution, before the appearance of tissue-grade animals or any other complex morphological feature found in all present day Metazoa. Striking similarities between sponge and human protein-coding genes indicate a short distance from both sponge and human genomes to the genome of the metazoan ancestor. Next, according to gene content, sponges are more similar to the sea anemone, human, and sea urchin than to the sea squirt, fruit fly, or nematode. Regarding the latter three, divergence from the sponge/human repertoire seems to serve as a reliable signature of accelerated evolutionary rate in distinct metazoan lineages. This also corroborates the findings that many genes were eliminated from the genomes of analyzed lineages (especially from two invertebrates) and further emphasizes the importance of gene loss in evolutionary processes. Our findings also raise many questions about the roles of numerous genes/proteins in the life of such a simple animal. Finally, the implication that sponges have unusually complex genomes, especially in contrast to unicellular eukaryotes, leads to a conclusion that the ancestor of all metazoans (Urmetazoa) also had a complex genome and strengthens a theory toward a Precambrian "gene explosion" view on metazoan evolution.

Supplementary Material

Supplementary supporting information, figures S2, and tables S1, S2, S3 S4, S5, and S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors wish to dedicate this paper to the late Prof. Vera Gamulin, the initiator of this work who devoted a considerable part of her career to sponge molecular genetics.

We sincerely thank Gordana Maravić Vlahoviček, M. Madan Babu, Bojan Žagrović, Bassem Hassan, James Sharpe, and three reviewers for critical reading and numerous suggestions on improving the manuscript. Petar Glažar is kindly acknowledged for the help with MADNet pathway mapping.

This work is funded by the European Molecular Biology Organization Young Investigator Program (Installation grant 1431/2006 to K.V.), International Center for Genetic Engineering and Biotechnology collaborative research program grant CRP/CRO07-03 to K.V. and Croatian MSES grants 098-0982913-2478 (M.H., H.C., and D.P.) and 119-0982913-1211 (K.V.). W.E.G.M. acknowledges the DFG Mü/14-3 grant. Author contributions: M.H.: designed research, performed research, and wrote paper; M.R.: performed research and analyzed data; H.C.: performed research and analyzed data; D.P.: analyzed data; M.W.: analyzed data; W.E.G.M.: performed sequencing and contributed materials; K.V.: designed research, performed research, and wrote paper.

Data deposition

The EST sequences reported in this paper have been deposited to the dbEST section of GenBank, with accession numbers GH555730-GH558302 for *L. baicalensis* and GH558303-GH571686 for *S. domuncula*.

References

Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, Degnan BM. 2007. Wnt and TGF-beta expression in the sponge Amphimedon queenslandica and the origin of metazoan embryonic patterning. *PLoS One*. 2:e1031.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet*. 9:868–882.

Cetkovic H, Grebenjuk VA, Muller WE, Gamulin V. 2004. Src proteins/src genes: from sponges to mammals. *Gene* 342:251–261.

Cetkovic H, Mikoc A, Muller WE, Gamulin V. 2007. Ras-like small GTPases form a large family of proteins in the marine sponge Suberites domuncula. *J Mol Evol*. 64:332–341.

Cetkovic H, Muller WE, Gamulin V. 2004. Bruton tyrosine kinase-like protein, BtkSD, is present in the marine sponge Suberites domuncula. *Genomics* 83:743–745.

Dehal P, Satou Y, Campbell RK, et al. (87 co-authors). 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* 298:2157–2167.

Dohrmann M, Janussen D, Reitner J, Collins AG, Worheide G. 2008. Phylogeny and evolution of glass sponges (porifera, hexactinellida). *Syst Biol*. 57:388–405.

Ereskovsky AV, Dondua AK. 2006. The problem of germ layers in sponges (Porifera) and some issues concerning early metazoan evolution [Review]. *Zoologischer Anzeiger*. 245:65–76.

Erwin DH. 2009. Early origin of the bilaterian developmental toolkit. *Philos Trans R Soc Lond B Biol Sci*. 364:2253–2261.

Gamulin V, Muller IM, Muller WEG. 2000. Sponge proteins are more similar to those of Homo sapiens than to Caenorhabditis elegans. *Biol J Linn Soc*. 71:821–828.

Gamulin V, Skorokhod A, Kavsan V, Muller IM, Muller WE. 1997. Experimental indication in favor of the introns-late theory: the receptor tyrosine kinase gene from the sponge Geodia cydonium. *J Mol Evol*. 44:242–252.

Harcet M, Lukic-Bilela L, Cetkovic H, Muller WEG, Gamulin V. 2005. Identification and analysis of cDNAs encoding two nucleoside diphosphate kinases (NDPK/Nm23) from the marine sponge Suberites domuncula. *Croatica Chemica Acta*. 78:343–348.

Hoshiyama D, Suga H, Iwabe N, Koyanagi M, Nikoh N, Kuma K, Matsuda F, Honjo T, Miyata T. 1998. Sponge Pax cDNA related to Pax-2/5/8 and ancient gene duplications in the Pax family. *J Mol Evol*. 47:640–648.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Res*. 9:868–877.

Hui JH, Holland PW, Ferrier DE. 2008. Do cnidarians have a ParaHox cluster? Analysis of synteny around a Nematostella homeobox gene cluster. *Evol Dev*. 10:725–730.

Hui JH, Raible F, Korchagina N, et al. (11 co-authors). 2009. Features of the ancestral bilaterian inferred from Platynereis dumerilii ParaHox genes. *BMC Biol*. 7:43.

Kanehisa M, Araki M, Goto S, et al. (11 co-authors). 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 36:D480–D484.

King N, Westbrook MJ, Young SL, et al. (36 co-authors). 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* 451:783–788.

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.

Kusserow A, Pang K, Sturm C, et al. (11 co-authors). 2005. Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* 433:156–160.

Lapebie P, Gazave E, Ereskovsky A, Derelle R, Bezac C, Renard E, Houliston E, Borchiellini C. 2009. WNT/beta-catenin signalling and epithelial patterning in the homoscleromorph sponge Oscarella. *PLoS One*. 4:e5823.

Larroux C, Fahey B, Degnan SM, Adamski M, Rokhsar DS, Degnan BM. 2007. The NK homeobox gene cluster predates the origin of Hox genes. *Curr Biol*. 17:706–710.

Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. 2008. Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol*. 25:980–996.

Li CW, Chen JY, Hua TE. 1998. Precambrian sponges with cellular structures. *Science* 279:879–882.

Matus DQ, Pang K, Marlow H, Dunn CW, Thomsen GH, Martindale MQ. 2006. Molecular evidence for deep evolutionary roots of bilaterality in animal development. *Proc Natl Acad Sci U S A*. 103:11195–11200.

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

Muller WE. 1995. Molecular phylogeny of Metazoa (animals): monophyletic origin. *Naturwissenschaften*. 82:321–329.

Muller WEG, Bohm M, Grebenjuk VA, Skorokhod A, Muller IM, Gamulin V. 2002. Conservation of the positions of metazoan introns from sponges to humans. *Gene* 295(2 Special Issue):299–309.

Nichols SA, Dirks W, Pearse JS, King N. 2006. Early evolution of animal cell signaling and adhesion genes. *Proc Natl Acad Sci U S A*. 103:12451–12456.

Ogura A, Ikeo K, Gojobori T. 2005. Estimation of ancestral gene set of bilaterian animals and its implication to dynamic change of gene content in bilaterian evolution. *Gene* 345:65–71.

Ono K, Suga H, Iwabe N, Kuma K, Miyata T. 1999. Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan-eumetazoan split. *J Mol Evol*. 48:654–662.

Perina D, Cetkovic H, Harcet M, Premzl M, Lukic-Bilela L, Muller WEG, Gamulin V. 2006. The complete set of ribosomal proteins from the marine sponge Suberites domuncula. *Gene* 366:275–284.

Philippe H, Derelle R, Lopez P, et al. (20 co-authors). 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 19:706–712.

Putnam NH, Srivastava M, Hellsten U, et al. (19 co-authors). 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*. 317:86–94.

Raible F, Tessmar-Raible K, Osoegawa K, et al. (12 co-authors). 2005. Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. *Science* 310:1325–1326.

Segota I, Bartonicek N, Vlahovicek K. 2008. MADNet: microarray database network web server. *Nucleic Acids Res*. 36: W332–W335.

Sodergren E, Weinstock GM, Davidson EH, et al. (228 co-authors). 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science* 314:941–952.

Sperling EA, Peterson KJ, Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol Biol Evol*. 26:2261–2274.

Srivastava M, Begovic E, Chapman J, et al. (21 co-authors). 2008. The Trichoplax genome and the nature of placozoans. *Nature* 454: 955–960.

Steele RE. 2005. Genomics of basal metazoans. *Integr Comp Biol*. 45:639–648.

Suga H, Koyanagi M, Hoshiyama D, Ono K, Iwabe N, Kuma K, Miyata T. 1999. Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. *J Mol Evol*. 48:646–653.

Suga H, Ono K, Miyata T. 1999. Multiple TGF-beta receptor related genes in sponge and ancient gene duplications before the parazoan-eumetazoan split. *FEBS Letters*. 453:346–350.

Suga H, Sasaki G, Kuma KI, Nishiyori H, Hirose N, Su ZH, Iwabe N, Miyata T. 2008. Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Letters*. 582:815–818.

Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 4:41.

Tessmar-Raible K, Arendt D. 2005. New animal models for evolution and development. *Genome Biol*. 6:303.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. 31:258–261.

Wainright PO, Hinkle G, Sogin ML, Stickel SK. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* 260:340–342.

Wiens M, Batel R, Korzhev M, Muller WE. 2003. Retinoid X receptor and retinoic acid response in the marine sponge Suberites domuncula. *J Exp Biol*. 206:3261–3271.

Wiens M, Mangoni A, D'Esposito M, et al. (11 co-authors). 2003. The molecular basis for the evolution of the metazoan bodyplan: extracellular matrix-mediated morphogenesis in marine demosponges. *J Mol Evol*. 57(Suppl 1):S60–S75.