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 Vlahoviček⁎,2,4

1Department of Molecular Biology, Rudjer Boskovic Institute, Zagreb, Croatia
2Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia
3Institute for Physiological Chemistry and Pathobiochemistry, Johannes Gutenberg University, Medical School, Mainz, Germany
4Department of Informatics, University of Oslo, Oslo, Norway

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

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 mono-

© The Author(s) 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
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 (Tesmar-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\(^{-60}\). 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
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 ($1 \times 10^{-5}$ and $1 \times 10^{-40}$, 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 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
suggesting that a large gene and module explosion event counts at the count amounts to only half the number found in the sea protein domains needed for transition to multicellularity, pre-either missing or significantly divergent in the involved in the signaling processes, present in sponge, and genome (supplementary SI and supplementary table S5, Supplementary Material online) leading to a conclusion assess the size of the as any known metazoan genome neither in terms of gene that the choanoflagellate genome is not nearly as complex as any known metazoan genome neither in terms of gene number nor repertoire. 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 proteome 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
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. 6A).

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. 6B 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...
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 synten, 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. 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).
Fig. 6. Maximum parsimony phylogenetic tree (A) based on concatenated sequences of proteins involved in signaling pathways. Homologs with \( E \) values of \( 1 \times 10^{-20} \) 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...
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 \textit{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 Maravic Vlahoviček, M. Madan Babu, Bojan Žagrović, Bassem Hassan, James Sharpe, and three reviewers for critical reading and numerous suggestions on improving the manuscript. Petar Glaziš is kindly acknowledged for the help with MADNet pathway mapping.

This work is funded by the European Molecular Biology Organization Young Investigator Program (Registration grant 1431/2006 to K.V.), International Center for Genetic Engineering and Biotechnology collaborative research program grant CRP/CR007-03 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; M.W.: analyzed data; K.V.: designed research, performed research, and wrote paper; 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


