This is an accepted manuscript of an article published as
Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Fortunato, S., Adamski, M., Mendivil Ramos, O., Leininger, S., Liu, J., Ferrier, D. E. K. & Adamska, M. 30 Oct 2014 In : Nature. 514, p. 620-623 doi: 10.1038/nature13881

Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes

Sofia A.V. Fortunato[1, 2], Marcin Adamski[1], Olivia Mendivil Ramos[3, †], Sven Leininger[1, ◊], Jing Liu[1], David E.K. Ferrier[3] and Maja Adamska[1§]

[1]Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, 5008 Bergen, Norway

[2]Department of Biology, University of Bergen, Thormøhlensgate 55, 5008 Bergen, Norway

[3] The Scottish Oceans Institute, Gatty Marine Laboratory, School of Biology, University of St Andrews, East Sands, St Andrews, Fife KY16 8LB, UK.

[†] Current address: Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

[◊] Current address: Institute of Marine Research, Nordnesgaten 50, 5005 Bergen, Norway

[§]Corresponding author

Email address: maja.adamska@sars.uib.no

Summary

Sponges are simple animals with few cell types, but their genomes paradoxically contain a wide variety of developmental transcription factors[1-4], including homeobox genes belonging to the *Antennapedia* (ANTP)class[5,6], which in bilaterians encompass *Hox*, *ParaHox* and *NK* genes. In the genome of the demosponge *Amphimedon queenslandica*, no *Hox* or *ParaHox* genes are present, but *NK* genes are linked in a tight cluster similar to the NK clusters of bilaterians[5]. It has been proposed that *Hox* and *ParaHox* genes originated from *NK* cluster genes after divergence of sponges from the lineage leading to cnidarians and bilaterians[5,7]. On the other hand, synteny analysis gives support to the notion that absence of *Hox* and *ParaHox* genes in *Amphimedon* is a result of secondary loss (the ghost locus hypothesis)[8]. In this study, we analyzed complete suites of *Antp* class homeoboxes in two calcareous sponges, *Sycon ciliatum* and *Leucosolenia complicata*. Our phylogenetic analyses demonstrate that these calcisponges possess orthologues of bilaterian *NK* genes (*Hex*, *Hmx* and *Msx*), a varying number of additional *NK* genes and one *ParaHox* gene, *Cdx*. Despite generation of scaffolds spanning multiple genes, we find no evidence of clustering of *Sycon NK* genes. All *Sycon Antp*-class genes are developmentally expressed, with patterns suggesting involvement in cell type specification in embryos and adults, metamorphosis and body plan patterning. The present study demonstrates that *ParaHox* genes predate the origin of sponges, thus confirming the ghost locus hypothesis[8], and highlights the need to analyze genomes of multiple sponge lineages in order to obtain a complete picture of the ancestral composition of the first animal genome.

Sponges (Porifera) are strong candidates for being the earliest extant lineage(s) of animals[9]. The genome sequence of the demosponge *Amphimedon queenslandica* provided rich material for comparative studies on the origins of metazoan developmental genes, cell types and body plan[1]. Among others, it has fuelled hypotheses about the origin of one of the most widely studied groups of developmental genes: the *Antennapedia* (ANTP)class homeoboxes, including *Hox*, *ParaHox* and *NK* genes[5,7,8,10,11]. *Antp* genes have been found in all animals and are involved in multiple developmental processes, including body plan patterning and neurogenesis[12]. *Hox*, *ParaHox* and *NK* genes are often found in clusters[5,13], and in some animals their expression is temporally or spatially correlated to their position within the cluster (this being known as colinearity)[12]. The *Amphimedon* genome contains eight *NK* genes, but neither *Hox* nor *ParaHox* genes are present[5]. Six *NK* genes are linked in a tight cluster, and their simple embryonic and larval expression patterns are not consistent with colinearity[5,6]. Lack of *Hox* and *ParaHox* genes in *Amphimedon*, and also in the ctenophore *Mnemiopsis leidyi*[7], has previously been interpreted as reflecting the ancestral condition, and gave rise to the ParaHoxozoa hypothesis, in which all animal lineages apart from poriferans and ctenophores are collectively known as the ParaHoxozoa. Others[10] have interpreted the phylogenetic evidence differently, suggesting that both *Hox* and *ParaHox* genes were originally present in sponges, but have subsequently been lost. This view has been recently revived by identification of *Hox* and *ParaHox* "ghost loci" (regions that display synteny to bilaterian *Hox* and *ParaHox* loci, but lack the *Hox/ParaHox* genes themselves) in the genome of *Amphimedon*[8].

We expected that expanding the range of sequenced sponge genomes would provide new information about the evolutionary history of genes important for the origin and

65 evolution of the animal kingdom. Calcisponges form a poriferan lineage which has been
66 separated from the demosponges for at least 600 million years[9]. We have recently
67 started analysis of the developmental toolkits of two calcisponges, *Sycon ciliatum* and
68 *Leucosolenia complicata*[2,4,14]. Here, we have searched for *Antp*-class homeobox genes in
69 transcriptomic and genomic assemblies of these species.

70 We retrieved ten *Antp*-class homeodomains in *Sycon*, constituting nine transcripts (one
71 with two homeoboxes), and twelve *Antp*-class homeodomains in *Leucosolenia*,
72 constituting nine transcripts (one with four homeoboxes) (Supplementary dataset 1).
73 Our phylogenetic analyses demonstrate that the repertoire of *Antp*-class genes is similar
74 between the two calcisponges, but strikingly different than in the demosponge
75 *Amphimedon*. Calcisponges and demosponges have clear orthologues of the bilaterian
76 genes *Hex* and *Msx;* calcisponges also have a clear *NK5* (*Hmx)* orthologue, which seems
77 to be lacking in *Amphimedon*. In contrast, this demosponge has possible *Bsh, BarH/BarI*
78 and *Tlx* genes, which are not recognizable in calcisponges. While in *Amphimedon* there is
79 a single gene associated with the bilaterian NK2/3/4 clade[15], several paralogues are
80 present in the two calcisponges. They contain multi-homeobox genes with non-
81 orthologous relationships between *Sycon* and *Leucosolenia*, and other genes containing
82 single homeoboxes in the NK2/3/4 clade. Affiliation of *Sycon* and *Leucosolenia NKB* and
83 *NK*G and the *LcoNKF* genes with a particular bilaterian NK family is not clear. No Hox
84 genes were found; however, a pair of the calcisponge genes showing affinities with the
85 ParaHox Cdx subfamily given the concordance of Neighbour-Joining (NJ) and Maximum
86 Likelihood (ML) analyses (Fig. 1, Extended Data Fig. 1).

87 Given the importance of this potential assignment, we performed further phylogenetic
88 analyses of these putative Cdx orthologues. In addition to the ELEKEF motif which is

shared by many Hox and ParaHox, but not NK-type homeodomains, the Cdx family has some distinctive residues in its homeodomain, most notably the YIT motif present only in a small number of other ANTP class homeodomains (Supplementary note 1 and Extended Data Fig. 2). Phylogenetic analyses focused on these few families in addition to families represented in sponges, based on greater taxon sampling than in the overall classification, produced a significantly supported clustering of *SciCdx* and *LcoCdx* with *Cdx* genes from other species in NJ, ML and Bayesian analyses (Extended Data Figs. 3-5).

We have also investigated the genomic neighbourhood of *SciCdx* to help resolve the identity of this homeobox gene (Fig. 2; Supplementary note 2 and Supplementary Table 1). From the 14 genes on the *SciCdx* scaffold that have clear human orthologues (Supplementary Table 1), four are orthologues of genes linked to ParaHox loci in humans. One of these, SAR1A/B also has a conserved neighbouring relationship with the ParaHox cluster in the cnidarian, *Nematostella vectensis* (Fig. 2a, b). Although these gene numbers are insufficient to reach statistical significance, the neighbour relationships are consistent with the identification of *SciCdx* as a ParaHox gene. Furthermore, as one would expect from the ghost locus hypothesis and the identification of *SciCdx* as a bona fide *ParaHox* gene, we also find clustering of *Sycon* orthologues of ParaHox and Hox neighbour genes into two distinct groups in the *Sycon* genome to statistically significant levels (Fig. 2c-e). Altogether the evidence is consistent with the identification of *SciCdx* and *LcoCdx* as the first examples of sponge *ParaHox* genes.

All of the *Sycon NK* genes are found on separate scaffolds (Extended data fig. 6) with multiple additional genes surrounding the homeobox genes. We interpret this as the ancient *NK* cluster having been broken apart in the *Sycon* genome. Alternatively, our current assembly is not sufficient to provide evidence of a cluster with multiple genes

113 inserted between the *NK* genes. It has been previously shown that arrangements of *NK*
114 genes are variable between different species, ranging from intact and conserved *NK*
115 clusters[5,16,17] to clusters that are partially broken[15,18].

116 We studied the expression of *Antp*-class genes in *Sycon* using a combination of *in situ*
117 hybridization with quantitative transcriptome analysis (Fig. 3, Supplementary note 3
118 and Extended Data Figs 7-9). For all *Antp*-class genes, except *SciHex*, expression can be
119 detected in oocytes and during cleavage (Fig. 3a and Extended Data Fig. 7a-g). During
120 embryogenesis, the most striking expression domain of the majority of the identified
121 genes is the cruciform cells, which are putative larval sensory cells[2,14]. Beginning at the
122 four-cell stage, stronger expression of *SciNKA* marks the cytoplasm destined to become
123 partitioned into the cruciform cells (Fig. 3d and Extended Data Fig. 7h-q) and expression
124 of *SciHmx* is also markedly elevated in these cells (Fig. 3e). *SciNKC* and *NKD* are uniquely
125 and strongly expressed in the cruciform cells of more advanced (pre-inversion stage)
126 embryos (Fig. 3f-g). *SciNKA* is additionally detected in macromeres of embryos and
127 larvae, and along with *SciNKG* and *SciNKB* domains, forms a set of adjacent stripes along
128 the larval anterior-posterior axis (Fig. 3h-p). This pattern is reminiscent of "striped"
129 patterns reported for *NK* genes in bilaterians[19], and might be indicative of roles for the
130 calcisponge *NK* genes in axial patterning of the larval body plan or in cell type
131 determination with cells destined for specific fates distributed along the larval axis. For
132 example, the macromeres give rise to the pinacocytes of the outer cell layer[20], and the
133 *SciNKG*-positive micromeres are good candidates for future sclerocytes (spicule
134 producing cells) given co-expression of *SciNKG* and sclerocyte-specific carbonic
135 anhydrases (*scl-CA1* and *scl-CA2*)[21]. All of the *Antp*-class genes except *SciNKC* and *SciNKD*
136 are expressed during metamorphosis (Fig. 3b, Extended Data Fig. 9) in sub-populations

of cells in all three cell layers (Extended Data Fig. 7). The clear expression of *SciCdx* in the inner cell mass during formation of the choanocyte chamber (Fig .3q) is particularly striking in light of the recently revived notion of homology of the sponge choanoderm with bilaterian endoderm[14], as ParaHox expression in bilaterians is often associated with the developing gut. In adults, most of the *Antp*-class genes display differential expression along the body axis (Fig. 3b and Extended Data Table 1). *SciNKG* and *SciNKA* are strongly expressed in sclerocytes, while *SciMsx* and *SciHmx* transcripts are predominantly detected within and around the oscular sphincter (Fig. 3r-w).

In summary, analysis of *Antp*-class genes in a previously understudied lineage of sponges allowed us to demonstrate pre-poriferan ancestry of *ParaHox* genes, thus confirming the ghost locus hypothesis and rejecting the ParaHoxozoa hypothesis of *Hox/ParaHox* gene origins. Expression patterns of the identified genes indicate that developmental functions of *Antp*-class genes also predate poriferans, with probable involvement in specification of potentially homologous structures (choanoderm/endoderm and cross cells/sensory cells) as well as morphological novelties (calcareous spicules). Differences in *Antp*-class gene repertoires between the demosponge *Amphimedon* and the two calcisponges, *Sycon* and *Leucosolenia*, are striking, and the fact that both classes of sponges share a subset of genes with bilaterians indicates independent gene loss events in the two poriferan lineages.

Methods

Genome and transcriptome assemblies will be described in detail elsewhere (Adamski, Leininger and Adamska, unpublished results). Briefly, the high quality draft genome assembly of *S. ciliatum* was generated using two (360bp and 530bp) paired-end libraries and several mate-pair libraries ranging from 2.0 to 9.0 kb and the preliminary draft

assembly of *L. complicata* was generated from a single 295 bp paired-end library, all prepared and sequenced by Illumina technology. Assembly was performed using SOAPdenovo2[22] and scaffolding using SSPACE v2.2[23], and resulted in N50 = 150kbp and 450bp for *S. ciliatum* and *L. complicata*, respectively. Transcriptomes were assembled using Trinity[24]. For *S. ciliatum*, genomic scaffolds and transcripts of sponge origin (as opposed to those derived from associated organisms) were identified by aligning the resulting assembly to reads from an Illumina sequenced library obtained from laboratory grown, eukaryotic- contamination free juveniles. The calcisponge *Antp*-class sequences were retrieved from these assemblies using TBLASTN with representative query homeodomain sequences from *Amphimedon queenslandica*, *Mus musculus*, *Tribolium castaneum* and *Branchiostoma floridae*. For phylogenetic analysis, we selected *B. floridae* and *T. castaneum* to provide a framework for the classification of the sponge sequences, as these species have been shown to collectively contain homologues of all major bilaterian *Antp*-class genes[25]. Their homeodomain sequences were extracted from HomeoDB[26]. Prottest3.0[27] and Modelgenerator v0.85[28] were used to determine the best suitable model of sequence evolution (LG+G). Phylogenetic analyses were based on Neighbour-Joining (Phylip v3.69), Maximum Likelihood (PhyML v3.0) and Bayesian inference (MrBayes v3.1.2) methods. Gene expression was studied using available packages[29,30]. *S. ciliatum Antp*-class gene amplification, cloning, sequencing, probe production and single *in situ* hybridization were performed as described previously[2]. In the double *in situ* experiment, samples were hybridized simultaneously with digoxigenin-labelled *SciNKB* probe and fluorescein-labelled *SciNKG* probe. After detection of the digoxigenin-labelled probe with NBT/BCIP substrate, the anti-digoxigenin antibody was removed by two 5-minute washes in 0.1 M glycine/HCl, pH 2.2/0.1% Tween 20 followed by three additional maleic acid buffer washes. A second

round of pre-blocking, antibody incubation and post-antibody washes were as in the single probe protocol with the exception that anti-Fluo-AP antibody was used and the colour developed using Fast Red tablets (Roche) according to manufacturer's instructions. Photographs demonstrating gene expression are representative of multiple individual specimens, with following replicates: oocytes and embryos: 3-4 small pieces of adult sponge, each containing tens to hundreds of oocytes or embryos of a given developmental stage; young syconoid sponges: at least 5 individual specimens; juveniles: small petri dishes or wells of multi-well plates containing at least 10 juveniles. At least two independent experiments were carried for each probe.

References:

1 Srivastava, M. *et al.* The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**, 720-726, doi:http://www.nature.com/nature/journal/v466/n7307/abs/nature09201.html#supplementary-information (2010).

2 Fortunato, S. *et al.* Genome-wide analysis of the sox family in the calcareous sponge Sycon ciliatum: multiple genes with unique expression patterns. *EvoDevo* **3**, 14 (2012).

3 Larroux, C. *et al.* Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* **25**, 980 - 996 (2008).

4 Fortunato, S., Leininger, S. & Adamska, M. Evolution of the Pax-Six-Eya-Dach network: the calcisponge case study. *EvoDevo* **5**, 23 (2014).

5 Larroux, C. *et al.* The NK Homeobox Gene Cluster Predates the Origin of Hox Genes. *Current biology : CB* **17**, 706-710 (2007).

6 Fahey, B., Larroux, C., Woodcroft, B. J. & Degnan, B. M. Does the High Gene Density in the Sponge NK Homeobox Gene Cluster Reflect Limited Regulatory Capacity? *The Biological Bulletin* **214**, 205-217 (2008).

7 Ryan, J. *et al.* The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *EvoDevo* **1**, 9, doi:10.1186/2041-9139-1-9 (2010).

8 Mendivil Ramos, O., Barker, D. & Ferrier, David E. K. Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. *Current biology : CB* **22**, 1951-1956 (2012).

9 Wörheide, G. *et al.* in *Advances in Marine Biology* Vol. Volume 61 (eds Maria J. Uriz Manuel Maldonado Mikel A. Becerro & Turon Xavier) 1-78 (Academic Press, 2012).

10 Peterson, K. J. & Sperling, E. A. Poriferan ANTP genes: primitively simple or secondarily reduced? *Evolution & Development* **9**, 405-408, doi:10.1111/j.1525-142X.2007.00179.x (2007).

11 Ferrier, D. E. K. in *Hox Genes: Studies from the 20th to the 21st Century* Vol. 689 *Advances in Experimental Medicine and Biology* (ed J. S. Deutsch) 91-100 (Springer-Verlag Berlin, 2010).

12 Garcia-Fernandez, J. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6**, 881-892 (2005).
13 Chourrout, D. *et al.* Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* **442**, 684-687, doi:http://www.nature.com/nature/journal/v442/n7103/suppinfo/nature04863_S1.html (2006).
14 Leininger, S. *et al.* Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. *Nat Commun* **5**, doi:10.1038/ncomms4905 (2014).
15 Luke, G. N. *et al.* Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5292-5295, doi:10.1073/pnas.0836141100 (2003).
16 Hui, J. H. L. *et al.* Extensive Chordate and Annelid Macrosynteny Reveals Ancestral Homeobox Gene Organization. *Mol. Biol. Evol.* **29**, 157-165, doi:10.1093/molbev/msr175 (2012).
17 Hui, J. H. L., Holland, P. W. H. & Ferrier, D. E. K. Do cnidarians have a ParaHox cluster? Analysis of synteny around a Nematostella homeobox gene cluster. *Evolution & Development* **10**, 725-730, doi:10.1111/j.1525-142X.2008.00286.x (2008).
18 Seo, H.-C. *et al.* Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. *Nature* **431**, 67-71, doi:http://www.nature.com/nature/journal/v431/n7004/suppinfo/nature02709_S1.html (2004).
19 Saudemont, A. *et al.* Complementary striped expression patterns of NK homeobox genes during segment formation in the annelid Platynereis. *Dev. Biol.* **317**, 430-443, doi:http://dx.doi.org/10.1016/j.ydbio.2008.02.013 (2008).
20 Amano, S. & Hori, I. Metamorphosis of Calcareous Sponges .2. Cell Rearrangement and Differentiation in Metamorphosis. *Invertebrate Reproduction & Development* **24**, 13-26 (1993).
21 Voigt, C., Adamski, M., Sluzek, M. & Adamska, M. Genome-wide screening identifies two carbonic anhydrases involved in biomineralization in calcareous sponges. *BMC Evol. Biol. pending minor revisions* (2014).
22 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).
23 Boetzer, M., Henkel, C., Jansen, H., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).
24 Grabherr, M. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652, doi:10.1038/nbt.1883 (2011).
25 Takatori, N. *et al.* Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. *Dev. Genes Evol.* **218**, 579-590, doi:10.1007/s00427-008-0245-9 (2008).
26 Zhong, Y.-F. & Holland, P. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & development* **13**, 567-568, doi:10.1111/j.1525-142X.2011.00513.x (2011).
27 Darriba, D., Taboada, G., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **27**, 1164-1165, doi:10.1093/bioinformatics/btr088 (2011).
28 Keane, T., Creevey, C., Pentony, M., Naughton, T. & Mclnerney, J. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology* **6**, 29 (2006).
29 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010).
30 Li, B. & Dewey, C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

Acknowledgements

This study was funded by the Sars Centre core budget to MA. Sequencing was performed at Norwegian High Throughput Sequencing Centre funded by the Norwegian Research Council. OMR and DEKF acknowledge support from the BBSRC and the School of Biology, University of St Andrews. We thank Brith Bergum for collecting samples in the 2011 season and Rita Holdhus from the Genomics Core Facility at the University of Bergen for help with DNA shearing.

Author Contributions

SAVF carried out gene identification and cloning, analysed gene expression by in situ hybridization, and participated in the phylogenetic analyses and manuscript writing. Mar.A performed sequence assembly, annotation, quantification of gene expression and participated in sample collection, phylogenetic analyses and manuscript writing. OMR performed the synteny analyses, participated in phylogenetic analyses, manuscript writing and design of the research approach for synteny and phylogenetic analyses. SL isolated samples for sequencing of genomes, generated MP libraries, and participated in sample collection. JL generated samples for sequencing of *S. ciliatum* metamorphosis stages. DEKF participated in design of the research approach for synteny and phylogenetic analyses and writing of the manuscript. Maj.A conceived the study, participated in data analysis, sample collection and writing of the manuscript.

Author Information

Genome assembly of *Sycon ciliatum* and the coding sequences and their translations from transcriptome assemblies of *S. ciliatum* and *Leucosolenia complicata* used in this study can be accessed through http://compagen.zoologie.uni-kiel.de/ and are also deposited at http://datadryad.org/ (doi:10.5061/dryad.tn0f3). The RNA-Seq data are

deposited at www.ebi.ac.uk/arrayexpress (E-MTAB-2430, 2431, 2890), and the cloned coding sequences of *S. ciliatum Antp*-class genes are deposited at European Nucleotide Archive under accession codes HGXXXXXX to YYYYYY.

Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M. Adamska (maja.adamska@sars.uib.no).

Figure Legends

Figure 1. Phylogenetic tree of the *Antp*-class homeodomains. A neighbour joining (NJ) tree is displayed**.** Three support values are shown: left/black value is NJ bootstrap support, middle/blue is Maximum Likelihood bootstrap support and right/red is posterior probability from Bayesian analysis. Bootstrap values below 10% and posterior probability values below 0.5 are not shown except for associations of calcisponge sequences. The root was determined by using selected *Prd*-class genes as an outgroup. Acronyms of the species are: *Amphimedon queenslandica,* Aqu; *Leucosolenia complicata*, Lco; *Sycon ciliatum*, Sci; *Branchiostoma floridae*, Bfl; and *Tribolium castaneum*, Tca. Scale bar indicates number of aminoacid substitutions per site.

Figure 2. *SciCdx* synteny and ghost loci simulations. (**a, b**) Genomic neighborhoods of *SciCdx* gene and *N. vectensis* ParaHox cluster[17], colors indicate orthologous relationship to human genes with following chromosomal location: yellow – ParaHox neighbours, orange – not linked to Hox/ParaHox loci, yellow-orange – mix of ParaHox and non-Hox/ParaHox neighbours, grey – no orthology. Green lines highlight the conserved *Sar1*-

Cdx linkage. (**c-e**) Monte Carlo simulations of human Hox and ParaHox neighbour orthologue distributions and their overlap across *S. ciliatum* scaffolds. Arrows indicate numbers of scaffolds with Hox and ParaHox neighbour orthologues and their co-localization in *S. ciliatum;* the observed distributions imply distinct Hox and ParaHox loci.

Figure 3. Expression of *S.ciliatum Antp*-class genes. a, Adult specimen. **b**, expression "heat map", * indicate statistically higher[29,30] expression in the apical/top region in comparison to middle (>M) or basal (>B) parts. **c**, oocytes. **d**, **e**, cleavage and **f-k**, pre-inversion stage embryos: arrows/cc/rainbow colouring – forming cross cells expressing multiple *Antp*-class genes, */mac/blue – macromeres expressing *SciNKA*; mic/red – equatorial micromeres expressing *SciNKG* and *SciNKB*. **l-p,** post-inversion embryos and larvae; **q,** post-larva; **r-w**, top parts of sponges: *SciNKG* and *SciNKA* in sclerocytes, *SciMsx* and *SciHmx* in cells of the oscular sphincter (arrows). Scale bars represent 10μm, except **l-o**: 25μm, **q**, **r**, **t**, **v**, **w** – 50μm.

Extended Data Figure 1. Phylogenetic tree of the ANTP class homeodomains including representative bilaterian and non- bilaterian sequences. A Neighbour Joining (NJ) tree using the JTT+G (0.5) (1000 bootstraps) model of protein evolution is displayed. A combination of three support values obtained for three phylogenetic methods is shown: left (black) value is bootstrap (BT) support from NJ, middle (blue) is bootstrap support from Maximum Likelihood (LG+G 0.5) and right (red) is posterior probability from Bayesian analysis (LG+G 0.5). BT values below 10% and PP values below 0.5 are not shown except for associations of calcisponge sequences. The root was determined by using selected *Prd*-class genes as an outgroup. Acronyms of the species used are: *Amphimedon queenslandica,* Aqu (Porifera/demosponges); *Leucosolenia*

complicata, Lco; *Sycon ciliatum*, Sci (Porifera/calcisponges); *Nematostella vectensis*, Nve (Cnidaria); *Trichoplax adhaerens,* Tad (Placozoa); *Mnemiopsis leidyi*, Mle (Ctenophora); *Branchiostoma floridae*, Bfl (Chordata); and *Tribolium castaneum*, Tca (Arthropoda). Scale bar indicates number of aminoacid substitutions per site.

Extended Data Figure 2. Variability of the 'YIT/YIS' homeodomain motif within the Cdx/Cad, En and Dbx families in bilaterians, cnidarians, a placozoan and sponges. Acronyms of the species used are Hsa (*Homo sapiens*), Bfl (*Branchiostoma floridae*), Cte (*Capitella teleta*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis*), Tad (*Trichoplax adhaerens*), Tca (*Tribolium castaneum*), Sci (*Sycon ciliatum*), Lco (*Leucosolenia complicata*), Edi (*Eleutheria dichotoma*), Nv (*Nereis virens*), Pdu (*Platynereis dumerilii*), Pfl (*Ptychodera flava*), Dre (*Danio rerio*), Dme (*Drosophila melanogaster*), Ame (*Apis mellifera*), Gga (*Gallus gallus*), Xla (*Xenopus laevis*), Mmu (*Mus musculus*) and Aqu (*Amphimedon queenslandica*).

Extended Data Figure 3. Phylogenetic tree including ANTP class homeodomain subfamilies represented in sponges and two additional subfamilies characterized by presence of YIT motif (Cdx and En), but excluding divergent *A. queenslandica* sequences (NK5/6/7a/b and BarH). NJ (JTT, 1000) bootstrap support values are in black, ML (LG+G 0.4, 1000 replicates) bootstrap support values are in blue and BY (LG+G 0.4) posterior probabilities values in red. Only bootstrap support values equal to or above 500 are shown. All subfamilies except Cdx are collapsed for clarity. Acronyms of the species used are Hsa (*Homo sapiens*), Bfl (*Branchiostoma floridae*), Cte (*Capitella teleta*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis*), Tad (*Trichoplax adhaerens*), Tca (*Tribolium castaneum*), Sci (*Sycon ciliatum*), Lco (*Leucosolenia complicata*), Edi (*Eleutheria dichotoma*), Nv (*Nereis virens*), Pdu (*Platynereis dumerilii*), Pfl (*Ptychodera*

flava), Dre (*Danio rerio*), Dme (*Drosophila melanogaster*), Ame (*Apis mellifera*), Gga (*Gallus gallus*), Xla (*Xenopus laevis*), Mmu (*Mus musculus*) and Aqu (*Amphimedon queenslandica*). Scale bar indicates number of aminoacid substitutions per site.

Extended Data Figure 4. Phylogenetic tree including ANTP class homeodomain subfamilies represented in sponges and three additional subfamilies characterized by presence of YIT/YIS motifs (Cdx, En and Dbx), but excluding some of divergent *A. queenslandica* sequences (NK5/6/7a/b). NJ (JTT, 1000 replicates) bootstrap support values are in black, ML (LG+G 0.4, 1000 replicates) bootstrap support values are in blue and BY (LG+G 0.4) posterior probabilities values in red. Only bootstrap support values equal to or above 500 are shown. All subfamilies except Cdx are collapsed for clarity. Acronyms of the species used are Hsa (*Homo sapiens*), Bfl. (*Branchiostoma floridae*), Cte (*Capitella teleta*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis*), Tad (*Trichoplax adhaerens*), Tca (*Tribolium castaneum*), Sci (*Sycon ciliatum*), Lco (*Leucosolenia complicata*), Edi (*Eleutheria dichotoma*), Nv (*Nereis virens*), Pdu (*Platynereis dumerilii*), Pfl (*Ptychodera flava*), Dre (*Danio rerio*), Dme (*Drosophila melanogaster*), Ame (*Apis mellifera*), Gga (*Gallus gallus*), Xla (*Xenopus laevis*), Mmu (*Mus musculus*) and Aqu (*Amphimedon queenslandica*). Scale bar indicates number of aminoacid substitutions per site.

Extended Data Figure 5. Phylogenetic tree including *Antp*-class homeodomain subfamilies represented in sponges and three additional subfamilies characterized by presence of YIT/YIS motifs (Cdx, En and Dbx). NJ (JTT, 1000 replicates) bootstrap support values are in black, ML (LG+G 0.4, 1000 replicates) bootstrap support values are in blue and BY (LG+G 0.4) posterior probabilities values in red. Only bootstrap support values equal to or above 500 are shown. All subfamilies

except Cdx are collapsed for clarity. Acronyms of the species used are Hsa (*Homo sapiens*), Bfl (*Branchiostoma floridae*), Cte (*Capitella teleta*), Lgi (*Lottia gigantea*), Nve (*Nematostella vectensis*), Tad (*Trichoplax adhaerens*), Tca (*Tribolium castaneum*), Sci (*Sycon ciliatum*), Lco (*Leucosolenia complicata*), Edi (*Eleutheria dichotoma*), Nv (*Nereis virens*), Pdu (*Platynereis dumerilii*), Pfl (*Ptychodera flava*), Dre (*Danio rerio*), Dme (*Drosophila melanogaster*), Ame (*Apis mellifera*), Gga (*Gallus gallus*), Xla (*Xenopus laevis*), Mmu (*Mus musculus*) and Aqu (*Amphimedon queenslandica*). Scale bar indicates number of aminoacid substitutions per site.

Extended Data Figure 6. *Sycon ciliatum* scaffolds containing NK genes (blue) and *Amphimedon queenslandica* scaffold containing cluster of NK genes (modified after[6]; green). Annotation of the neighboring genes (genes within 50 kbp from the NK gene) in *S. ciliatum* was performed using blastp searches against refseq database.

Extended Data Figure 7. Additional expression patterns of ANTP class homeobox genes in embryonic development and during metamorphosis. All of the investigated genes (except Hex, not shown) are expressed in oocytes (a-g). The expression of *SciNKA* is detectable in all blastomeres of the cleavage stage embryos, but the transcripts are concentrated in the corner-most cytoplasm which becomes gradually partitioned to the cross cells (arrows). This subcellular localization of cross-cells enriched transcripts is also observed for *SciNanos*, expression of which, similarly to *SciNKA*, becomes ultimately restricted to cross cells and macromeres in preinversion stage embryos (**l-q**). In metamorphosing postlarvae, *SciNKA* is expressed in the cells of the outer layer (**r**), *SciNKB* and *SciNKG* in (possibly non-overlapping) fractions of cells in the inner cell mass (**s**, **t**); *SciHex* is weakly expressed throughout the inner cell mass (**u**) and *SciNKC* (**v**) and

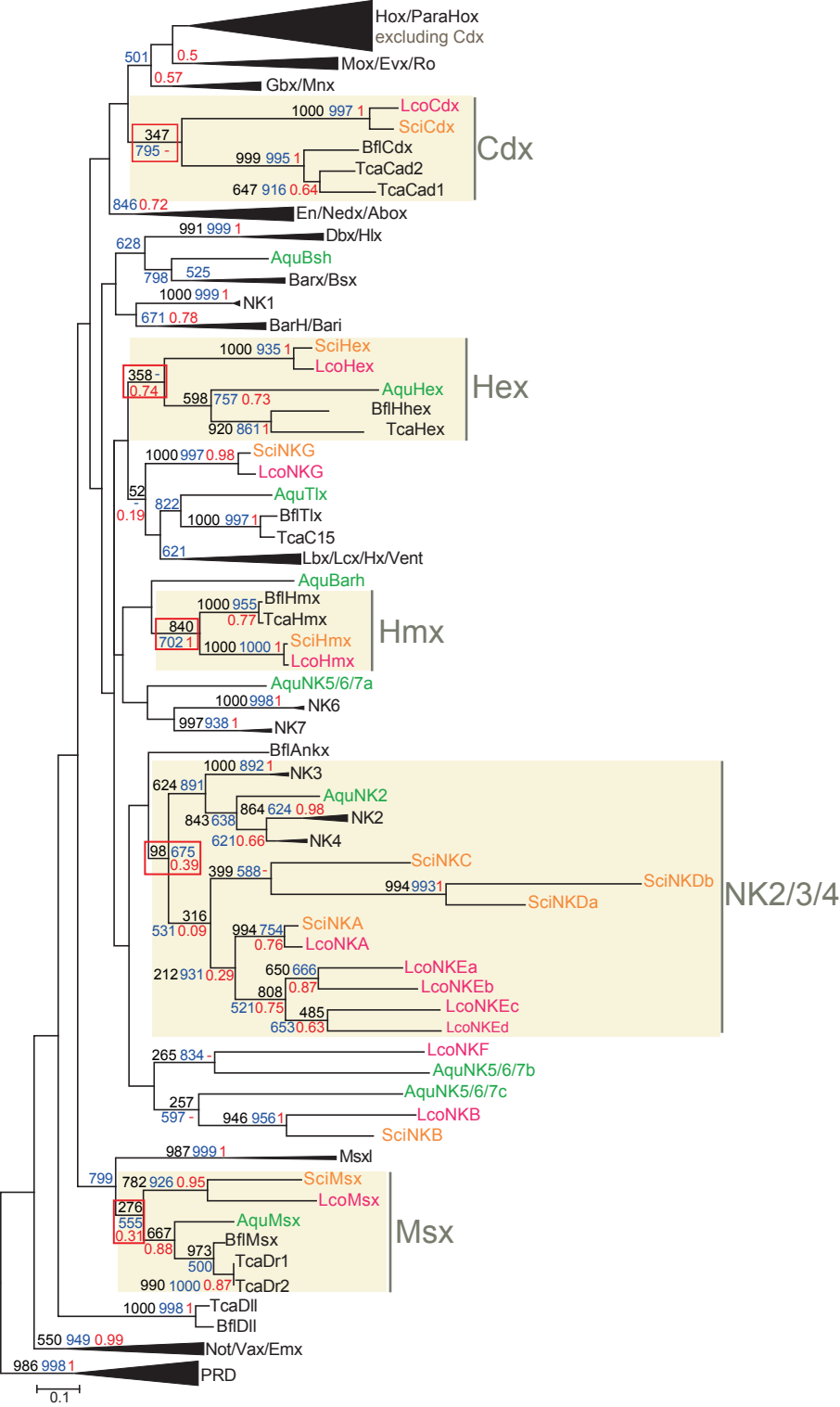
NKD (not shown) are not detectable in the juveniles. Scale bars represent 10μm, except **r-v**: 25μm.

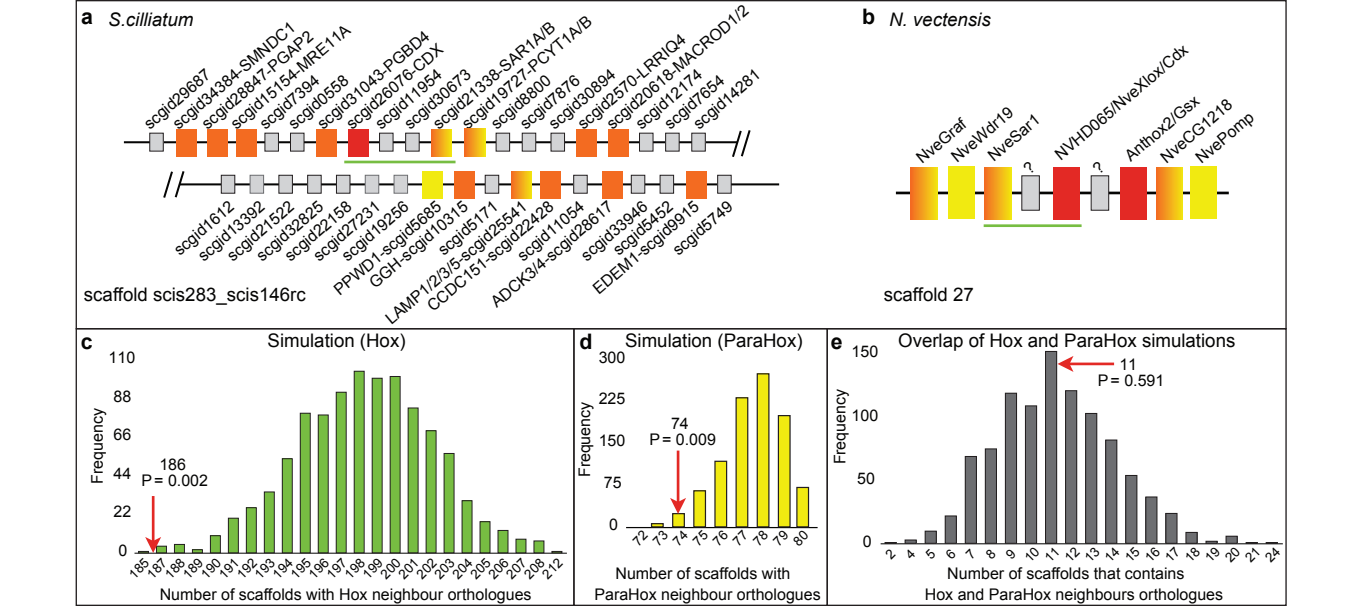
Extended Data Figure 8. **Samples used for quantification of expression**. **a-f**, metamorphosis in *S.ciliatum*, stages are based on[20] with modifications: Stage I, approximately 12 hours post settlement: large flat cells derived from larval macromeres envelop the inner cell mass composed of former micromeres (**a**); Stage II, approximately 24 hours post settlement: single-axis spicules (monaxons) are produced by sclerocytes, which have differentiated from the inner cell mass cells (**b**). Stage III, 2-3 days after settlement: choanocytes which have differentiated from the inner cells mass cells form a single internal chamber (**c**). Stage IV, approximately 4 days after settlement: osculum (exhalant opening) forms at the apical end of the spherical juvenile; first tri-radial spicules become apparent (**d**). Stage V, approximately 10 days after settlement: the juvenile is elongated along the apical-basal axis, long straight spicules form a crown around the osculum (**e**). Young syconoid sponges, approximately 8 weeks after settlement (**f**). (**a-e**) are photographs of live specimens in culture; photographs **a-d** are top (apical) views, cartoon representations of sections and photograph **e** are side views. Scale bars represent 100μm, except **f**: 1mm. (**g**) Details of replicates used for the analysis. Several hundred juveniles were used in each sample. (**h**) Plot demonstrating results of principal component analysis of the metamorphosis series and axial dissection series of non-reproductive adults calculated according to[29] and utilizing information of the top 500 differentially expressed genes as in default parameters. Metamorphosis stages and parts of sponges are colour-coded, with the ovals added manually for easier visualization of similarities and differences between the samples. Progress of development, starting from freshly released larvae and until emergence of adult, but not

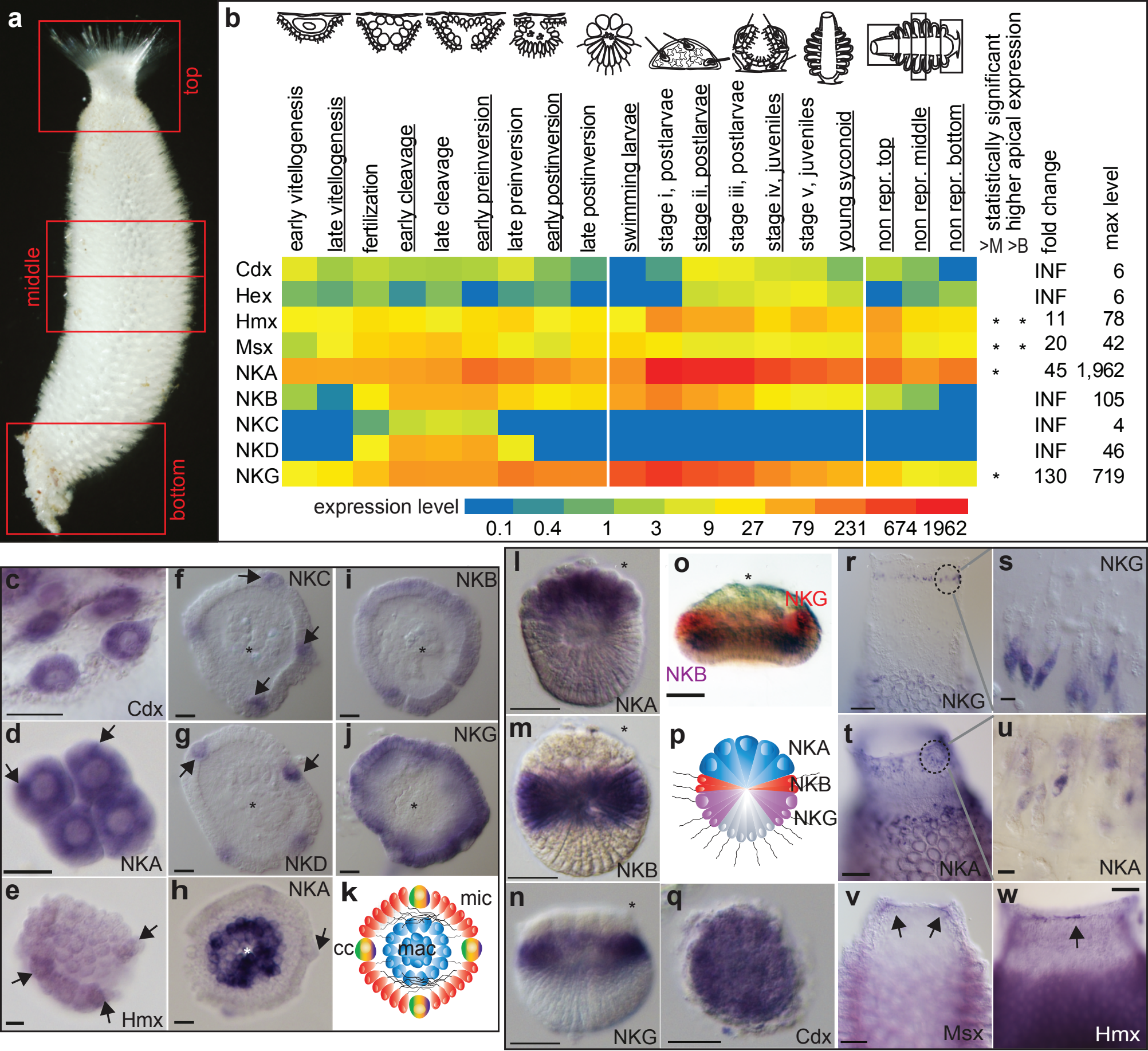
yet reproductive sponges, is indicated by arrows. Note similarities of samples within replicates and with neighbouring stages of the metamorphosis series, and distinctiveness of the top (apical) samples from the basal and middle samples of the adults. (**i**) Heatmap representation of sample-to-sample distances among all samples used in this study, calculated according to[29] and based on expression of all coding genes in *S. ciliatum* (approximately 18K sequences). Note that replicates and neighbouring stages group together, as indicated by highlighting.

Extended Data Figure 9. Heat-map representation of expression profiles demonstrated in Fig. 3B in the main text, but with data from individual libraries presented separately.

Extended Data Table 1. Quantification of differences in expression levels between top, middle and bottom parts of non-reproductive adult specimens of *S. ciliatum*. 'expression level' was calculated as sum of the posterior probability of each read coming from a given gene over all reads[30] scaled using size factors of the libraries[29]; 'fold change' was calculated between expression levels in middle (middle-top) or bottom (bottom-top) and top part of the sponge; 'adj. p-value' are p-value adjusted for multiple testing with the Benjamini-Hochberg procedure[29]. Values with statistical significance not less than 90% (adj. p-value <= 0.1) and apical expression level higher than in the middle or bottom part of the sponge are indicated by *.


Hox/ParaHox
excluding Cdx
Mox/Evx/Ro
Gbx/Mnx
LcoCdx
SciCdx
BflCdx
TcaCad2
TcaCad1
Cdx
En/Nedx/Abox
Dbx/Hlx
AquBsh
Barx/Bsx
NK1
BarH/Bari
SciHex
LcoHex
AquHex
BflHhex
TcaHex
Hex
SciNKG
LcoNKG
AquTlx
BflTlx
TcaC15
Lbx/Lcx/Hx/Vent
AquBarh
BflHmx
TcaHmx
SciHmx
LcoHmx
Hmx
AquNK5/6/7a
NK6
NK7
BflAnkx
NK3
AquNK2
NK2
NK4
SciNKC
SciNKDb
SciNKDa
SciNKA
LcoNKA
LcoNKEa
LcoNKEb
LcoNKEc
LcoNKEd
NK2/3/4
LcoNKF
AquNK5/6/7b
AquNK5/6/7c
LcoNKB
SciNKB
Msxl
SciMsx
LcoMsx
AquMsx
BflMsx
TcaDr1
TcaDr2
Msx
TcaDll
BflDll
Not/Vax/Emx
PRD
0.1


a
S.cilliatum
scgid29687
scgid34384-SMNDC1
scgid28847-PGAP2
scgid151154-MRE11A
scgid7394
scgid0558
scgid31043-PGBD4
scgid26076-CDX
scgid11954
scgid30673
scgid21338-SAR1A/B
scgid19727-PCYT1A/B
scgid8800
scgid7876
scgid30894
scgid2570-LRRIQ4
scgid20618-MACROD1/2
scgid12174
scgid7654
scgid14281
scgid1612
scgid13392
scgid21522
scgid32825
scgid22158
scgid27231
scgid19256
scgid5685
scgid10315
scgid5171
scgid25541
scgid22428
scgid11054
scgid28617
scgid33946
scgid5452
scgid9915
scgid5749
PPWD1-scgid5685
GGH-scgid10315
LAMP1/2/3/5-scgid25541
CCDC151-scgid22428
ADCK3/4-scgid28617
EDEM1-scgid9915
scaffold scis283_scis146rc
b
N. vectensis
NveGraf
NveWdr19
NveSar1
NVHD065/NveXlox/Cdx
Anthox2/Gsx
NveCG1218
NvePomp
?
?
scaffold 27
c
Simulation (Hox)
Frequency
110
88
66
44
22
0
186
P = 0.002
185 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 212
Number of scaffolds with Hox neighbour orthologues
d
Simulation (ParaHox)
Frequency
300
225
150
75
0
74
P = 0.009
72 73 74 75 76 77 78 79 80
Number of scaffolds with
ParaHox neighbour orthologues
e
Overlap of Hox and ParaHox simulations
Frequency
150
100
50
0
11
P = 0.591
2 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 24
Number of scaffolds that contains
Hox and ParaHox neighbours orthologues


a
top
middle
bottom
b
early vitellogenesis
late vitellogenesis
fertilization
early cleavage
late cleavage
early preinversion
late preinversion
early postinversion
late postinversion
swimming larvae
stage i, postlarvae
stage ii, postlarvae
stage iii, postlarvae
stage iv, juveniles
stage v, juveniles
young syconoid
non repr. top
non repr. middle
non repr. bottom
>M
>B
statistically significant higher apical expression
fold change
max level
Cdx INF 6
Hex INF 6
Hmx * * 11 78
Msx * * 20 42
NKA * 45 1,962
NKB INF 105
NKC INF 4
NKD INF 46
NKG * 130 719
expression level
0.1 0.4 1 3 9 27 79 231 674 1962
c Cdx
f NKC
i NKB
d NKA
g NKD
j NKG
e Hmx
h NKA
k mic cc mac
l NKA
o NKG NKB
r NKG
s NKG
m NKB
p NKA NKB NKG
t NKA
u NKA
n NKG
q Cdx
v Msx
w Hmx