LETTER

https://doi.org/10.1038/s41586-019-1290-4

Pluripotency and the origin of animal multicellularity

Shunsuke Sogabe[1,3,8], William L. Hatleberg[1,4,8], Kevin M. Kocot[1,2], Tahsha E. Say[1], Daniel Stoupin[1,5], Kathrein E. Roper[1,6], Selene L. Fernandez-Valverde[1,7], Sandie M. Degnan[1]* & Bernard M. Degnan[1]*

A widely held—but rarely tested—hypothesis for the origin of animals is that they evolved from a unicellular ancestor, with an apical cilium surrounded by a microvillar collar, that structurally resembled modern sponge choanocytes and choanoflagellates[1–4]. Here we test this view of animal origins by comparing the transcriptomes, fates and behaviours of the three primary sponge cell types—choanocytes, pluripotent mesenchymal archaeocytes and epithelial pinacocytes—with choanoflagellates and other unicellular holozoans. Unexpectedly, we find that the transcriptome of sponge choanocytes is the least similar to the transcriptomes of choanoflagellates and is significantly enriched in genes unique to either animals or sponges alone. By contrast, pluripotent archaeocytes upregulate genes that control cell proliferation and gene expression, as in other metazoan stem cells and in the proliferating stages of two unicellular holozoans, including a colonial choanoflagellate. Choanocytes in the sponge *Amphimedon queenslandica* exist in a transient metastable state and readily transdifferentiate into archaeocytes, which can differentiate into a range of other cell types. These sponge cell-type conversions are similar to the temporal cell-state changes that occur in unicellular holozoans[5]. Together, these analyses argue against homology of sponge choanocytes and choanoflagellates, and the view that the first multicellular animals were simple balls of cells with limited capacity to differentiate. Instead, our results are consistent with the first animal cell being able to transition between multiple states in a manner similar to modern transdifferentiating and stem cells.

The last common ancestor of all living animals appears to have minimally possessed epithelial and mesenchymal cell types that could transdifferentiate within an ontogenetic life cycle[1,4]. This life cycle required an ability to regulate spatial and temporal gene expression, and included a diversified set of signalling pathways, transcription factors, enhancers, promoters and non-coding RNAs[5–9] (Fig. 1). Recent analyses reveal that unicellular holozoans use similar gene regulatory mechanisms to transit through the different cell states that comprise their life cycles[2,5,6,10–12]. These observations suggest that early stem metazoans were more complex than has generally been thought[1,3,4].

To test whether extant choanocytes and choanoflagellates accurately reflect the ancestral animal cell type, we first compared cell-type-specific transcriptomes[13] from the sponge *A. queenslandica* with transcriptomes expressed during the life cycles of the choanoflagellate *Salpingoeca rosetta*, the filasterean *Capsaspora owczarzaki* and the ichthyosporean *Creolimax fragrantissima*[10–12] (Fig. 1). We chose three sponge somatic cell types that are hypothesized to be homologous to cells present in the last common ancestor of contemporary metazoans, choanozoans or holozoans: (i) choanocytes, which are internal epithelial feeding cells that capture food by pumping water through the sponge; (ii) epithelial cells called pinacocytes, which line internal canals and the outside of the sponge; and (iii) mesenchymal pluripotent stem cells called archaeocytes, which inhabit the middle collagenous layer and have a range of other functions[2,14–16] (Extended Data Fig. 1 and Supplementary Video 1). These three cell types were manually collected and frozen within 15 min of *A. queenslandica* being dissociated (Supplementary Video 2). Their transcriptomes were sequenced using CEL-Seq2[17] and mapped to the Aqu2.1 annotated genome[18]. This approach enabled visual verification of the three cell types, minimized the time for transcriptional changes to occur after cell dissociation and enabled deep sequencing of cell-type transcriptomes (Extended Data Table 1 and Supplementary Files 1, 2).

Principal component analysis and sparse partial least squares discriminant analysis (sPLS-DA)[19] reveal that the transcriptomes of the three *A. queenslandica* cell types are unique, with choanocytes being the most distinct (Fig. 2a and Extended Data Fig. 1). Of 44,719 protein-coding genes, 11,013 genes were identified as significantly differentially expressed in at least one cell type from pairwise comparisons between the three cell types using DESeq2[20] (Fig. 2b and Supplementary File 3). Significant differences between cell types were independently corroborated by sPLS-DA, which highlighted a subset of 110 genes that explains 15% of the variance in the dataset and clearly discriminates the choanocytes from the other two cell types (Extended Data Fig. 1). This subset includes numerous putative immunity genes that typically

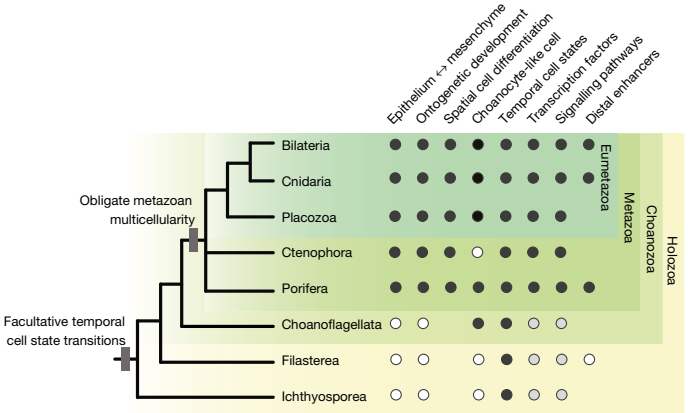


Fig. 1 | Cellular and regulatory traits in metazoans and unicellular holozoans. A phylogenetic tree showing holozoan relationships. Black dots, trait present; white dots, trait absent; grey dots, trait present but to a lesser extent than in animals; blank, trait undetermined. Facultative, environmentally induced gene regulation, which can lead to cell-state changes, appears to be an ancestral holozoan trait. Endogenous spatiotemporal gene regulation is obligatory for multicellular animals.

[1]School of Biological Sciences, University of Queensland, Brisbane, Queensland, Australia. [2]Department of Biological Sciences and Alabama Museum of Natural History, The University of Alabama, Tuscaloosa, AL, USA. [3]Present address: The Scottish Oceans Institute, Gatty Marine Laboratory, School of Biology, University of St Andrews, St Andrews, UK. [4]Present address: Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA. [5]Present address: BioQuest Studios, Port Douglas, Queensland, Australia. [6]Present address: Centre for Clinical Research, Faculty of Medicine, University of Queensland, Herston, Queensland, Australia. [7]Present address: CONACYT, Unidad de Genómica Avanzada, Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, Mexico. [8]These authors contributed equally: Shunsuke Sogabe, William L. Hatleberg. *e-mail: s.degnan@uq.edu.au; b.degnan@uq.edu.au

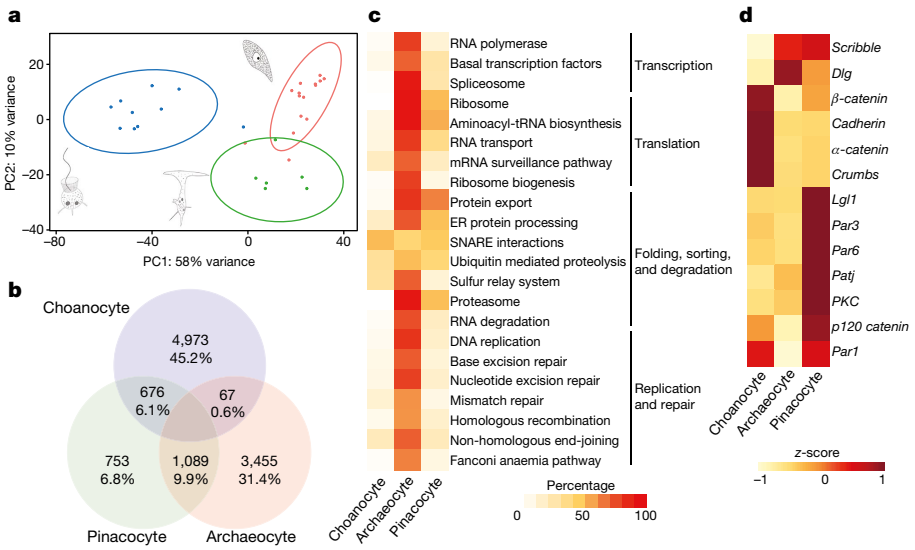


Fig. 2 | Comparison of choanocyte, archaeocyte and pinacocyte transcriptomes. a, Principal component (PC) analysis plot of CEL-Seq2 transcriptomes with 95% confidence level ellipse plots. Blue, choanocytes ($n = 10$); red, archaeocytes ($n = 15$); green, pinacocytes ($n = 6$). **b**, Venn diagram summarizing the number of significantly upregulated genes based on pairwise comparisons between each of the three cell types using a negative binomial distribution in DESeq2 with a false discovery rate <0.05. The percentages are of the total genes differentially upregulated in all cell types. **c**, Percentage of KEGG (Kyoto Encyclopedia of Genes and Genomes) 'genetic information processing' genes present in each cell type, corresponding to the number of components that make up each identified KEGG category. **d**, Scaled heat map illustrating the expression (*z*-score) of *A. queenslandica* epithelial cell polarity, junction and basal lamina genes in each cell type. Expression based on collapsed count values using the variance stabilizing transformation, which was blind to the experimental design.

encode multiple domains in unique configurations, including scavenger-receptor cysteine-rich, tetratricopeptide repeat and epidermal growth factor domains (Supplementary File 4).

We find that archaeocytes significantly upregulate genes involved in the control of cell proliferation, transcription and translation, consistent with their function as pluripotent stem cells (Fig. 2c and Supplementary File 5). By contrast, choanocyte and pinacocyte transcriptomes are enriched for suites of genes that are involved in cell adhesion, signalling and polarity, consistent with their role as epithelial cells (Fig. 2d, Extended Data Fig. 2 and Supplementary File 5).

The evolutionary age of all protein-coding genes in the *A. queenslandica* genome, and specifically of genes significantly and uniquely upregulated in each cell-type-specific transcriptome, was determined using phylostratigraphy, which is based on sequence similarity with genes in other organisms with a defined phylogenetic distance[21]. *A. queenslandica* genes were classified as having evolved (i) before or (ii) after the divergence of metazoan and choanoflagellate lineages (these are called pre-metazoan and metazoan genes, respectively), or (iii) after the divergence of the sponge lineage from all other animals (sponge-specific genes). The *A. queenslandica* genome comprises 28% pre-metazoan, 26% metazoan and 46% sponge-specific protein-coding

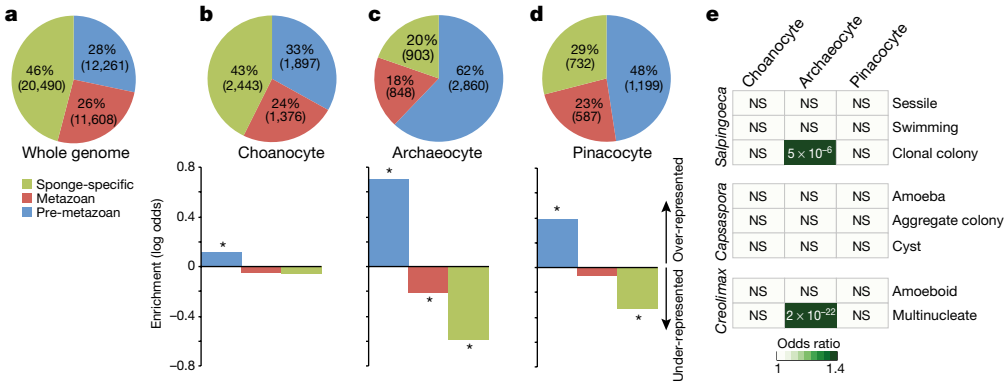


Fig. 3 | Analysis of gene age of choanocyte, archaeocyte and pinacocyte transcriptomes. a, Phylostratigraphic estimate of the evolutionary age of coding genes in the *A. queenslandica* genome. **b**–**d**, Estimate of gene age of differentially expressed genes in choanocytes (**b**, top), archaeocytes (**c**, top) and pinacocytes (**d**, top) and the enrichment of phylostrata relative to the whole genome (**b**–**d**, bottom). Asterisks indicate significant difference (two-sided Fisher's exact test $P < 0.001$) from the whole genome. The enrichment values (log-odds ratio) for: choanocytes (**b**; $n = 10$) are sponge-specific (−0.0089, $P = 0.7747$), metazoan (−0.0361, $P = 0.9958$) and premetazoan (0.0439, $P = 0.0004$) genes; archaeocytes (**c**; $n = 15$) are sponge-specific (−0.5634, $P = 1.33 \times 10^{-133}$), metazoan (−0.1923, $P = 1.04 \times 10^{-18}$) and premetazoan (0.6772, $P = 0$); and pinacocytes (**d**; $n = 6$) are sponge-specific (−0.2173, $P = 5.23 \times 10^{-13}$), metazoan (−0.0008, $P = 0.5231$) and premetazoan (0.2359, $P = 3.07 \times 10^{-36}$). **e**, A heat map comparing uniquely upregulated genes in *A. queenslandica* cell types ($n = 18{,}774$) that are orthologous to genes expressed during different life stages of *S. rosetta* ($n = 10{,}350$ orthology groups), *C. owczarzaki* ($n = 9{,}492$ orthology groups) and *C. fragrantissima* ($n = 11{,}449$ orthology groups). Colour indicates the significance of overlap in transcriptional profiles based on the odds ratio. Values indicate adjusted *P* values and show significant resemblance only between the archaeocyte and the *S. rosetta* colonial stage and the *C. fragrantissima* multinucleate stage transcriptomes. NS, not significant.

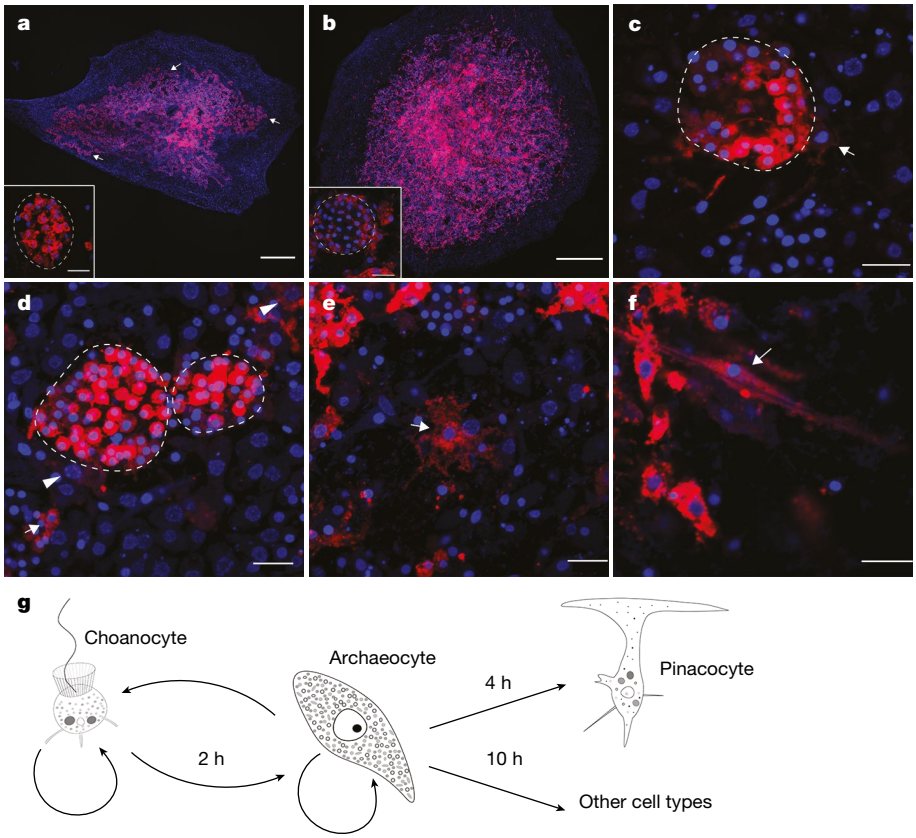


Fig. 4 | Transdifferentiation of choanocytes in *A. queenslandica*. **a**, **b**, Whole-mount views of four-day-old juveniles labelled with CM-DiI, 30 min after CM-DiI labelling (**a**; arrows, representative labelled choanocyte chambers), and 24 h after labelling (**b**). CM-DiI labelling spread from choanocyte chambers at 30 min to throughout the juvenile at 24 h with limited staining still present in choanocyte chambers; inserts, predominantly labelled and unlabelled choanocytes in chambers at 30 min and 24 h, respectively. **c**, **d**, Two hours (**c**) and four hours (**d**) after labelling, labelled cells (arrow) are present outside of choanocyte chambers (dotted lines), some of which have a large nucleus and a nucleolus (arrowheads) characteristic of archaeocytes. **e**, Six hours after labelling, CM-DiI-labelled pinacocytes (arrow) with thin pseudopodia are present. **f**, Twelve hours after initial labelling, labelled sclerocytes (arrow) and other cell types are present. The images presented in **a**–**f** represent the consensus of cell behaviours obtained from 10 independent labelling experiments, each comprising a minimum of 24 juveniles. **g**, Summary diagram of cell-type transition in the *A. queenslandica* juvenile. Scale bars, 200 μm (**a**, **b**), 10 μm (**c**–**f**).

genes (Fig. 3a and Supplementary File 6). We find that 43% of genes significantly upregulated in choanocytes are sponge-specific, which is similar to the proportion of the entire genome that is sponge-specific (Fig. 3b). By contrast, 62% of genes significantly upregulated in the pluripotent archaeocytes belong to the evolutionarily oldest pre-metazoan category, significantly higher than the 28% of genes for the entire genome (Fig. 3c). As with archaeocytes, pinacocytes express significantly more pre-metazoan and fewer sponge-specific genes than would be expected from the whole-genome profile (Fig. 3d). Results supporting this analysis are obtained when we (i) undertake the same phylostratigraphic analysis of all genes expressed in these cell types, taking into account relative transcript abundances (Extended Data Fig. 3 and Supplementary File 7), or (ii) classify gene age using an alternative orthology inference method (homology cluster containing both orthologues and paralogues)[22] among unicellular holozoan, yeast and *Arabidopsis* coding sequences (Extended Data Fig. 4).

Comparison of *A. queenslandica* cell-type transcriptomes with stage-specific transcriptomes from the choanoflagellate *S. rosetta*[10], the filasterean *C. owczarzaki*[11] and the ichthyosporean *C. fragrantissima*[12] reveals that archaeocytes have a transcriptome that is significantly similar to that of the colonial stage of the choanoflagellate and the multinucleate stage of the ichthyosporean (Fig. 3e). Consistent with this result, the significantly upregulated genes in the colonial or multinucleate stages of all three unicellular holozoans share the highest proportion of orthogroups with genes significantly upregulated in archaeocytes (Extended Data Fig. 5). By contrast, choanocyte and pinacocyte transcriptomes have no significant similarity to any of the examined unicellular holozoan transcriptomes, and share a lower proportion of orthogroups with unicellular holozoans compared to archaeocytes (Fig. 3e and Extended Data Fig. 5a).

When we compare the 94 differentially upregulated transcription factor genes in *A. queenslandica* choanocytes, pinacocytes and archaeocytes, we find no marked difference in their phylostratigraphic age, which suggests that the gene regulatory networks in these cells are of a similar evolutionary age (Extended Data Fig. 6 and Supplementary File 8). We detected 20, 25 and 21 orthologues of the 43 evolutionarily oldest (that is, pre-metazoan) transcription factor genes expressed in the *A. queenslandica* cells in the genomes of *S. rosetta*, *C. owczarzaki* and *C. fragrantissima*, respectively, with 9 of these being present in all species (Supplementary File 8). Comparison of the expression profiles of the transcription factor genes shared among these unicellular holozoans and *A. queenslandica* revealed no evidence of a conserved, co-expressed gene regulatory network (Extended Data Fig. 7 and Supplementary File 8). However, the proto-oncogene *Myc* and its heterodimeric partner *Max* are upregulated in *A. queenslandica* archaeocytes (Extended Data Fig. 6), as observed in other metazoan self-renewing pluripotent stem cells[23]. *Myc* and *Max* are also present in choanoflagellates, filastereans and ichthyosporeans, in which they heterodimerize and bind to E-boxes just as they do in animals[10–12,24]. *Myc* is expressed in the proliferative stage of *Capsaspora*, in which it regulates genes associated with ribosome biogenesis and translation[6]. Sponge archaeocytes also have enriched expression of genes involved in translation, transcription and DNA replication (Fig. 2c). This suggests that the role of *Myc* in regulating proliferation and differentiation predates its role in bilaterian

stem cells and cancer[23,25], and was probably a cardinal feature of the first metazoan cell.

Given that *A. queenslandica* choanocytes and archaeocytes express the largest number of derived and ancient transcriptomes, respectively, we investigated the developmental role of these cell types. In *Amphimedon* and most other demosponges, archaeocytes form during embryogenesis to populate the inner cell mass of the larva and are the most prevalent cell type during early metamorphosis[15,16,26]. As metamorphosis progresses, *Amphimedon* archaeocytes differentiate into other cell types that populate the juvenile body plan, including pinacocytes and choanocytes[16,26]. To understand the stability of choanocytes and their capacity to transdifferentiate, we selectively labelled choanocytes in three-day-old juvenile *A. queenslandica* with CM-DiI (Fig. 4a) and followed their fate over 24 h (Fig. 4b). Within 2 h of labelling, many choanocytes dedifferentiated into archaeocytes (Fig. 4c, d and Supplementary Video 3); this did not require prior cell division (Extended Data Fig. 8). Four hours later, some of these CM-DiI-labelled archaeocytes had differentiated into pinacocytes (Fig. 4e); within 12 h, multiple labelled cell types were present (Fig. 4e, f). Together, these results suggest that archaeocytes are essential in the development and maintenance of the *A. queenslandica* body plan, as appears to be the case in other sponges[15]. Unlike archaeocytes, choanocytes appear late in development and exist in a metastable state, sometimes lasting only a few hours before de-differentiating back into archaeocytes (Fig. 4g and Extended Data Fig. 8).

In conclusion, our analysis of sponge and unicellular holozoan cell transcriptomes, development and behaviour provides no support for the long-standing hypothesis that multicellular animals evolved from an ancestor that was an undifferentiated ball of cells resembling extant choanocytes and choanoflagellates[1–4]. This conclusion is corroborated by recent studies that question the homology of choanocytes and choanoflagellates based on cell structure[27,28]. As an alternative, we posit that the ancestral metazoan cell type had the capacity to exist in and transition between multiple cell states in a manner similar to modern transdifferentiating and stem cells. Recent analyses of unicellular holozoan genomes support this proposition, with some of the genomic foundations of pluripotency being established deep in a unicellular past[6,24]. Genomic innovations unique to metazoans—including the origin and expansion of key signalling pathway and transcription factor families, and regulatory DNA and RNA classes[7,9,29]—may have conferred the ability of this ancestral pluripotent cell to evolve a regulatory system whereby it could co-exist in multiple states of differentiation, giving rise to the first multicellular animal.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1290-4.

Received: 27 October 2017; Accepted: 16 May 2019;
Published online 12 June 2019.

1. Cavalier-Smith, T. Origin of animal multicellularity: precursors, causes, consequences-the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Phil. Trans. R. Soc. Lond. B* **372**, 20150476 (2017).
2. Brunet, T. & King, N. The origin of animal multicellularity and cell differentiation. *Dev. Cell* **43**, 124–140 (2017).
3. Arendt, D., Benito-Gutierrez, E., Brunet, T. & Marlow, H. Gastric pouches and the mucociliary sole: setting the stage for nervous system evolution. *Phil. Trans. R. Soc. Lond. B* **370**, 20150286 (2015).
4. Nielsen, C. Six major steps in animal evolution: are we derived sponge larvae? *Evol. Dev.* **10**, 241–257 (2008).
5. Sebé-Pedrós, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).
6. Sebé-Pedrós, A. et al. The dynamic regulatory genome of *Capsaspora* and the origin of animal multicellularity. *Cell* **165**, 1224–1237 (2016).
7. Gaiti, F. et al. Landscape of histone modifications in a sponge reveals the origin of animal *cis*-regulatory complexity. *eLife* **6**, e22194 (2017).
8. Gaiti, F., Calcino, A. D., Tanurdžić, M. & Degnan, B. M. Origin and evolution of the metazoan non-coding regulatory genome. *Dev. Biol.* **427**, 193–202 (2017).
9. Babonis, L. S. & Martindale, M. Q. Phylogenetic evidence for the modular evolution of metazoan signalling pathways. *Phil. Trans. R. Soc. Lond. B* **372**, 20150477 (2017).
10. Fairclough, S. R. et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol.* **14**, R15 (2013).
11. Sebé-Pedrós, A. et al. Regulated aggregative multicellularity in a close unicellular relative of metazoa. *eLife* **2**, e01287 (2013).
12. de Mendoza, A., Suga, H., Permanyer, J., Irimia, M. & Ruiz-Trillo, I. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* **4**, e08904 (2015).
13. Arendt, D. et al. The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
14. Maldonado, M. Choanoflagellates, choanocytes, and animal multicellularity. *Invertebr. Biol.* **123**, 1–22 (2004).
15. Ereskovsky, A. *The Comparative Embryology of Sponges* (Springer, 2010).
16. Nakanishi, N., Sogabe, S. & Degnan, B. M. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biol.* **12**, 26 (2014).
17. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
18. Fernandez-Valverde, S. L., Calcino, A. D. & Degnan, B. M. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*. *BMC Genomics* **16**, 387 (2015).
19. Lê Cao, K. A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).
20. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
21. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
22. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
23. Fagnocchi, L. & Zippo, A. Multiple roles of MYC in integrating regulatory networks of pluripotent stem cells. *Front. Cell Dev. Biol.* **5**, 7 (2017).
24. Young, S. L. et al. Premetazoan ancestry of the Myc–Max network. *Mol. Biol. Evol.* **28**, 2961–2971 (2011).
25. Kress, T. R., Sabò, A. & Amati, B. MYC: connecting selective transcriptional control to global RNA production. *Nat. Rev. Cancer* **15**, 593–607 (2015).
26. Sogabe, S., Nakanishi, N. & Degnan, B. M. The ontogeny of choanocyte chambers during metamorphosis in the demosponge *Amphimedon queenslandica*. *Evodevo* **7**, 6 (2016).
27. Mah, J. L., Christensen-Dalsgaard, K. K. & Leys, S. P. Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. *Evol. Dev.* **16**, 25–37 (2014).
28. Pozdnyakov, I., Sokolova, A., Ereskovsky, A. & Karpov, S. Kinetid structure of choanoflagellates and choanocytes of sponges does not support their close relationship. *Protistology* **11**, 248–264 (2017).
29. Srivastava, M. et al. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell isolation. Three random adult *A. queenslandica* were collected from Heron Island Reef, Great Barrier Reef and transferred to a closed aquarium facility where they were housed for no more than three days before being cut into approximately 1-cm^3 cubes. These cubes were randomly selected and mechanically dissociated by squeezing through a 20-μm mesh. The resultant cell suspension was diluted with 0.22-μm-filtered seawater (FSW) and the target cell types were identified microscopically on the basis of morphology. Archaeocytes are much larger than the other cells and possess a highly visible nucleolus. Choanocytes remain in intact choanocyte chambers after dissociation. Pinacocytes, unlike the other cell types, are translucent and maintain protruding cytoplasmic processes after dissociation. This approach avoided misidentification of dissociated cell types, but could not determine whether these cells are in the process of dividing or differentiating. Individual cells or choanocyte chambers were randomly collected under an inverted microscope (Nikon Eclipse Ti microscope) using a micropipette mounted on a micromanipulator (MN-4, Narishige) connected to CellTram Oil (Eppendorf) (Supplementary Video 2), flash-frozen and stored at −80 °C. All cells were frozen within 15 min of dissociation. Samples used in CEL-Seq2 comprised pools of either five to six archaeocytes or pinacocytes, or a single choanocyte chamber (~40–60 cells) (Extended Data Table 1). On the basis of differences in cell size, we estimated that these pools have similar amounts of total RNA. Three pinacocyte, and five archaeocyte and choanocyte samples were randomly collected from each of three sponges (Supplementary File 2).

CEL-Seq2 sample preparation, sequencing and analysis. Samples were prepared according to the CEL-Seq2 protocol[17] and sequenced on two lanes of Illumina HiSeq2500 on rapid mode using HiSeq Rapid SBS v.2 reagents (Illumina); CEL-Seq2 libraries were randomized in relation to cell type and source adult sponge in these two lanes. CEL-Seq2 reads were processed using a publicly available pipeline (https://github.com/yanailab/CEL-Seq-pipeline; see additional supplementary data on Dryad: /CEL-Seq pipeline; https://doi.org/10.5061/dryad.hp2fr73). Read counts were obtained from demultiplexed reads mapped to *A. queenslandica* Aqu2.1 gene models[18]; the *A. queenslandica* genome sequence can be accessed at http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index. Samples with read counts less than 10^6 were removed and not included in subsequent analyses (Supplementary File 2). For the samples included in the final analysis, approximately 60% of the reads successfully mapped to the genome (Extended Data Table 1), as per other studies using CEL-Seq[30].

Analysis of differentially expressed genes. The mapped read counts were analysed for differential gene expression using the bioconductor package DESeq2[20,31] (see additional supplementary data on Dryad: /DESeq2; https://doi.org/10.5061/dryad.hp2fr73). Genes that had read counts with a row sum of zero were removed. Principal component analyses were performed on blind variance stabilizing transformation (VST) counts obtained using DESeq2 and were visualized using the ggplot2 package[32]. Pairwise comparisons were conducted between each of the three cell types to generate a differentially expressed gene (DEG) list for each cell type using a false discovery rate <0.05. Venn diagrams were generated using VENNY (http://bioinfogp.cnb.csic.es/tools/venny) to visualize and compare the list of DEGs between each cell type. Heat maps were generated using the R packages pheatmap[33] and RColorBrewer[34] to visualize the expression patterns between the cell types using VST counts, which were scaled into *z*-score values ranging from −1 (low expression) to 1 (high expression).

All protein-coding genes were annotated using blastp (*e*-value cutoff = 1×10^{-3}) and InterProScan (default settings), which were merged in Blast2GO[35,36]. KEGG annotations were obtained using the online tool BlastKOALA[37] (see additional supplementary data on Dryad: /KEGG annotation; https://doi.org/10.5061/dryad.hp2fr73). Pathway analyses were performed using the annotations on the KEGG Mapper Reconstruct Pathway tool[38]. Complete DEG lists with BLAST2GO, InterPro, Pfam and phylostrata ID can be found in Supplementary File 3, as well as KEGG pathway enrichments in Supplementary File 5.

To identify the genes that best explain differences among cell-type transcriptomes, we adopted the multivariate sPLS-DA[19], implemented in the mixOmics package[39] in R v.3.3.1 (see additional supplementary data on Dryad: /sPLS-DA/README.txt; https://doi.org/10.5061/dryad.hp2fr73). This is a supervised analysis that uses the sample information (cell type) to identify the most predictive genes for classifying the samples according to cell type. The optimized numbers of genes per component were obtained by training and correctly evaluating the performance of the predictive model using fivefold cross-validation, repeated 100 times. A sample plot was used to visualize the similarities between samples for the final sPLS-DA model with 95% confidence ellipses using the plotIndiv function in R. A heat map was used to visualize relative expression levels of the selected gene models for the two components, using VST counts and the package pheatmap[33] in R. Venn diagrams were generated using VENNY to visualize and compare the DEGs generated by DESeq2 and sPLS-DA.

Phylostratigraphy. To estimate the evolutionary age of genes upregulated in each cell type, phylostratigraphy analyses[21] were performed using blastp and an *e*-value cutoff of 0.001 on a custom database containing 1,757 genomes and transcriptomes[40] that was modified to account for the phylogenetic position of *A. queenslandica* (that is, all eumetazoan and bilaterian taxa were moved into the metazoan phylostratum, and three phylostrata—poriferan, demosponge and haplosclerid—were added to increase the representation of poriferan transcriptomes; Supplementary File 6, additional supplementary data on Dryad: /Phylostratigraphy annotations; https://doi.org/10.5061/dryad.hp2fr73). Every gene model in *A. queenslandica* was blasted against each sequence in the database, and its age of gene origin was inferred based on the oldest blast hit relative to a predetermined phylogenetic tree (see additional supplementary data on Dryad: /Phylostratigraphy annotations; https://doi.org/10.5061/dryad.hp2fr73).

Phylostrata enrichments were performed using the Fisher's exact test[41] in the BioConductor package, GeneOverlap[42] in R, to identify significant differences in gene age of the cell-type DEG lists relative to the genome (see additional supplementary data on Dryad: /Fig. 3b–d and /ED_Fig. 3_files; https://doi.org/10.5061/dryad.hp2fr73). Enrichment (log odds ratio value above 0) and under-representation (log odds ratio value below 0) of each phylostrata found in the cell type DEG lists relative to the genome, were visualized using the R packages pheatmap[33] and RColorBrewer[34].

Orthology analyses. Orthology analyses were performed using FastOrtho[43] from a custom 'all-vs-all' blastp database of coding sequences from the genomes of *Saccharomyces cerevisiae*[44], *Arabidopsis thaliana*[45], *C. fragrantissima*[12], *Sphaeroforma arctica*[46], *C. owczarzaki*[47], *Monosiga brevicollis*[48] and *S. rosetta*[10], using the following configuration settings: pv_cutoff = 1e-5; pi_cutoff = 0.0; pmatch_cutoff = 0.0; maximum_weight = 316.0; inflation = 1.5; blast_e = 1e-5 (see additional supplementary data on Dryad: /FastOrtho; https://doi.org/10.5061/dryad.hp2fr73). FastOrtho classifies all of the genes present in each genome into orthology groups, which contain all orthologous and paralogous genes from each species. Genes that do not have any orthologues in other species or paralogues within the same genome were not included in any orthogroups. To compare the gene lists between species in all downstream analyses, species-specific gene names were changed to the common orthogroup identifier.

Orthology analyses between *A. queenslandica* and *S. rosetta*, *C. fragrantissima*, and *C. owczarzaki* cell types were performed using the cell-type-specific DEG lists obtained from previous studies on *S. rosetta*[10], *C. fragrantissima*[12] and *C. owczarzaki*[11]. The BioConductor package, GeneOverlap[42], was used to identify (1) the number of overlapping orthogroups between species and cell type, and (2) the statistical significance of that overlap based on list size and total number of orthogroups (see additional supplementary data on Dryad: /Fig. 3e; https://doi.org/10.5061/dryad.hp2fr73). This function provided the odds ratio between the orthogroup lists, for which the null hypothesis was no significant overlap (odds ratio value of 1 or smaller) and the alternative was a significant overlap detected between the lists (odds ratio value over 1), as well as a *P* value calculated for odds ratio values over 1.

To supplement phylostratigraphy analyses of *A. queenslandica* cell-type-specific gene lists (Fig. 3 and Extended Data Fig. 3), the BioConductor package, GeneOverlap[42] was used to identify the number and percentage of orthogroups that are also present in the genomes of *A. thaliana, S. cerevisiae, C. fragrantissima, S. arctica, C. owczarzaki, M. brevicollis* and *S. rosetta* (Extended Data Figs. 4, 5; see additional supplementary data on Dryad: /ED_Fig. 4 and ED_Fig. 5; https://doi.org/10.5061/dryad.hp2fr73).

Classification of gene-expression levels into quartiles. In addition to differential gene-expression analyses for *A. queenslandica* transcriptomes, the relative gene-expression levels for all cell types were assigned to one of four expression quartiles based on the number of reads that mapped to a given Aqu2.1 gene model (Extended Data Fig. 3). All zero read counts were discarded and the mean expression value of the non-transformed normalized count values of all samples (from all cell types) was used to calculate the quartile values. These values (Q_1: 2.30, Q_2: 6.06, Q_3: 15.83) were used to classify the expression of all of genes in each cell type into four groups based on transcript abundance, ranging from lowest (Q1) to highest (Q4).

Phylostrata enrichments for the different quartile value thresholds were performed as described above for the cell-type DEG lists; heat maps were generated using pheatmaps[33] in R (see additional supplementary data on Dryad: /ED_Fig. 3_files; https://doi.org/10.5061/dryad.hp2fr73). All downstream analyses used the median value (Q_2: 6.06) as a cut-off value to obtain a list of expressed genes. Orthology analyses using FastOrtho were performed as described above, and the percentage of genes with shared orthologous group in each gene list was calculated (see additional supplementary data on Dryad: /ED_Fig. 4_files and ED_Fig. 5_files; https://doi.org/10.5061/dryad.hp2fr73). In these analyses,

exclusive lists refer to all of the regions in the Venn diagram being treated as a separate list (for example, archaeocyte only, common between archaeocyte and choanocyte, common between archaeocyte and pinacocyte, and so on), whereas non-exclusive lists collapse all of the lists containing a given cell type into one list (for example, archaeocyte non-exclusive DEG list includes, archaeocyte DEGs + (archaeocyte + pinacocyte DEGs) + (archaeocyte + choanocyte DEGs).

Identification and analysis of expressed *A. queenslandica* transcription factors. A list of *A. queenslandica* transcription factors expressed in the three cell types was obtained using a number of independent methods. First, a non-conservative list of putative *A. queenslandica* transcription factors was obtained using the DNA-binding domain database (DBD: transcription factor prediction database) and the Pfam IDs of sequence specific DBD families, which corresponds to known transcription factor families (www.transcriptionfactor.org)[49]. Second, we collated a list of annotated *A. queenslandica* transcription factors in the literature[7,9,16,47,50–65] (Supplementary File 8). Third, we compared these lists to an unpublished in-house database for *A. queenslandica* and putative transcription factors identified by OrthoMCL. The final list of 173 expressed transcription factor genes used in this study were present in at least 2 of the 3 lists (Supplementary File 8).

The evolutionary age of each of the expressed transcription factors was first assigned based on the DBD contained in the gene model and then manually curated based primarily on literature (Supplementary File 8). From this, each transcription factor was assigned as either originating in sponges after diverging form other animals (sponge-specific), in metazoans after they diverged from choanoflagellates (metazoan) or before metazoans diverged from choanoflagellates (premetazoan).

Analysis of juvenile cell fate and proliferation. Larvae were collected as previously described[66], left in FSW overnight and then placed in sterile 6-well plates with 10 ml of FSW for 1 h in the dark with live coralline algae *Amphiroa fragilissima*. Postlarvae settled on *A. fragilissima* were removed using fine forceps (Dumont no. 5) and resettled on to round coverslips placed in a well with 2 ml FSW in a sterile 24-well plastic plate, with 3 postlarvae placed on each coverslip. Metamorphosis from resettled postlarvae to a functional juvenile takes approximately 72 h[16,67]. For all samples, FSW was changed daily until fixation.

The lipophilic cell tracker CM-DiI (Molecular Probes C7000) was used to label choanocyte chambers in juveniles as previously described[16], with slight modifications in the concentration used and incubation times. *A. queenslandica* juveniles were incubated in 1 μM CM-DiI in FSW for 30 min to 1 h. This minimized the labelling of non-choanocyte cells. Despite this precaution, some non-choanocyte cells would be labelled in some individuals. Hence, all CM-DiI-labelled juveniles were inspected by epifluorescence microscopy (Nikon Eclipse Ti microscope) immediately after CM-DiI was washed out, with juveniles detected with CM-DiI-labelled cells outside of choanocyte chambers discarded from the study. Juveniles were allowed to develop for 0, 2, 4, 6, 12 or 24 h post-incubation (hpi) with CM-DiI, then washed in FSW 3 times for 5 min and fixed[68] without dehydration in ethanol. Fixed juveniles were washed three times in MOPST (1× MOPS buffer + 0.1% Tween). Nuclei were labelled with DAPI (1:1,000, Molecular Probes) for 30 min, washed in MOPST for 5 min and mounted using ProlongGold antifade reagent (Molecular Probes). All samples were observed using the ZEISS LSM 710 META confocal microscope, and image analysis was performed using the software ImageJ.

To visualize cell proliferation, the thymidine analogue EdU (Click-iT EdU Alexa Fluor 488 cell proliferation kit, Molecular Probes C10337) was used as previously described[16,26]. To label S-phase nuclei, juveniles were incubated in FSW containing 200 μM EdU for 6 h, washed in FSW and immediately fixed as described above. Fluorescent labelling of incorporated EdU was conducted according to the manufacturer's recommendations before DAPI labelling and mounting in ProLong Gold antifade reagent as described above.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All cell-type transcriptome data are available in the NCBI SRA database under accession number PRJNA412708. Additional supplementary data are available from the Dryad Digital Respository: https://doi.org/10.5061/dryad.hp2fr73.

30. Levin, M. et al. The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637–641 (2016).
31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
32. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
33. Kolde, R. pheatmap v.1.0.8 https://cran.r-project.org/package=pheatmap (2012).
34. Neuwirth, E. RColorBrewer v.1.1-2 https://cran.r-project.org/package=RColorBrewer (2011).
35. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
36. Götz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
37. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
38. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44** (D1), D457–D462 (2016).
39. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
40. Aguilera, F., McDougall, C. & Degnan, B. M. Co-Option and de novo gene evolution underlie molluscan shell diversity. *Mol. Biol. Evol.* **34**, 779–792 (2017).
41. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
42. Shen, L. GeneOverlap: an R package to test and visualize gene overlaps http://shenlab-sinai.github.io/shenlab-sinai/ (2014).
43. Wattam, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2014).
44. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44** (D1), D710–D716 (2016).
45. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
46. Ruiz-Trillo, I., Lane, C. E., Archibald, J. M. & Roger, A. J. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *J. Eukaryot. Microbiol.* **53**, 379–384 (2006).
47. Suga, H. et al. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat. Commun.* **4**, 2325 (2013).
48. King, N. et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**, 783–788 (2008).
49. Wilson, D., Charoensawan, V., Kummerfeld, S. K. & Teichmann, S. A. DBD-taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* **36**, D88–D92 (2008).
50. Srivastava, M. et al. Early evolution of the LIM homeobox gene family. *BMC Biol.* **8**, 4 (2010).
51. Larroux, C. et al. Genesis and expansion of metazoan transcription factor gene classes. *Mol. Biol. Evol.* **25**, 980–996 (2008).
52. Larroux, C. et al. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evol. Dev.* **8**, 150–173 (2006).
53. Shimeld, S. M., Degnan, B. & Luke, G. N. Evolutionary genomics of the Fox genes: origin of gene families and the ancestry of gene clusters. *Genomics* **95**, 256–260 (2010).
54. Layden, M. J., Meyer, N. P., Pang, K., Seaver, E. C. & Martindale, M. Q. Expression and phylogenetic analysis of the *zic* gene family in the evolution and development of metazoans. *Evodevo* **1**, 12 (2010).
55. Presnell, J. S., Schnitzler, C. E. & Browne, W. E. KLF/SP transcription factor family evolution: Expansion, diversification, and innovation in eukaryotes. *Genome Biol. Evol.* **7**, 2289–2309 (2015).
56. Mukhopadhyay, S. & Jackson, P. K. The tubby family proteins. *Genome Biol.* **12**, 225 (2011).
57. Larroux, C. et al. The NK homeobox gene cluster predates the origin of Hox genes. *Curr. Biol.* **17**, 706–710 (2007).
58. Wang, L., Tang, Y., Cole, P. A. & Marmorstein, R. Structure and chemistry of the p300/CBP and Rtt109 histone acetyltransferases: implications for histone acetyltransferase evolution and function. *Curr. Opin. Struct. Biol.* **18**, 741–747 (2008).
59. Petroni, K. et al. The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* **24**, 4777–4792 (2012).
60. Morrison, A. J. & Shen, X. Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. *Nat. Rev. Mol. Cell Biol.* **10**, 373–384 (2009).
61. Jones, M. H., Hamana, N., Nezu, Ji. & Shimane, M. A novel family of bromodomain genes. *Genomics* **63**, 40–45 (2000).
62. Song, W., Solimeo, H., Rupert, R. A., Yadav, N. S. & Zhu, Q. Functional dissection of a Rice Dr1/DrAp1 transcriptional repression complex. *Plant Cell* **14**, 181–195 (2002).
63. Matheos, D. P., Kingsbury, T. J., Ahsan, U. S. & Cunningham, K. W. Tcn1p/Crz1p, a calcineurin-dependent transcription factor that differentially regulates gene expression in *Saccharomyces cerevisiae*. *Genes Dev.* **11**, 3445–3458 (1997).
64. Rivera, A. S. et al. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC Evol. Biol.* **10**, 123 (2010).
65. Romanovskaya, E. V. et al. Transcription factors of the NF1 family: Possible mechanisms of inducible gene expression in the evolutionary lineage of multicellular animals. *J. Evol. Biochem. Physiol.* **53**, 85–92 (2017).
66. Leys, S. P. et al. Isolation of *Amphimedon* developmental material. *Cold Spring Harb. Protoc.* **2008**, prot5095 (2008).
67. Degnan, B. M. et al. *Evolutionary Developmental Biology of Invertebrates*, vol. 1 (Springer, 2015).
68. Larroux, C. et al. Whole-mount in situ hybridization in *Amphimedon*. *Cold Spring Harb. Protoc.* **2008**, prot5096 (2008).

Acknowledgements This study was supported by funds from the Australian Research Council (B.M.D. and S.M.D.). We thank I. Ruiz-Trillo for primary expression data for *Capsaspora* and *Creolimax* and N. Rhodes for assistance with computing and database management.

Author contributions B.M.D. and S.M.D. conceived and designed the project. S.S., D.S. and K.E.R. identified and isolated the cells and prepared the libraries. W.L.H., S.S. and K.M.K. performed gene expression and annotation and phylostratigraphic analyses with help from T.E.S., S.M.D, S.L.F.-V. and B.M.D. S.S. performed cell-lineage analyses. B.M.D., S.M.D. and S.S. wrote the manuscript with comments and contributions from all authors.

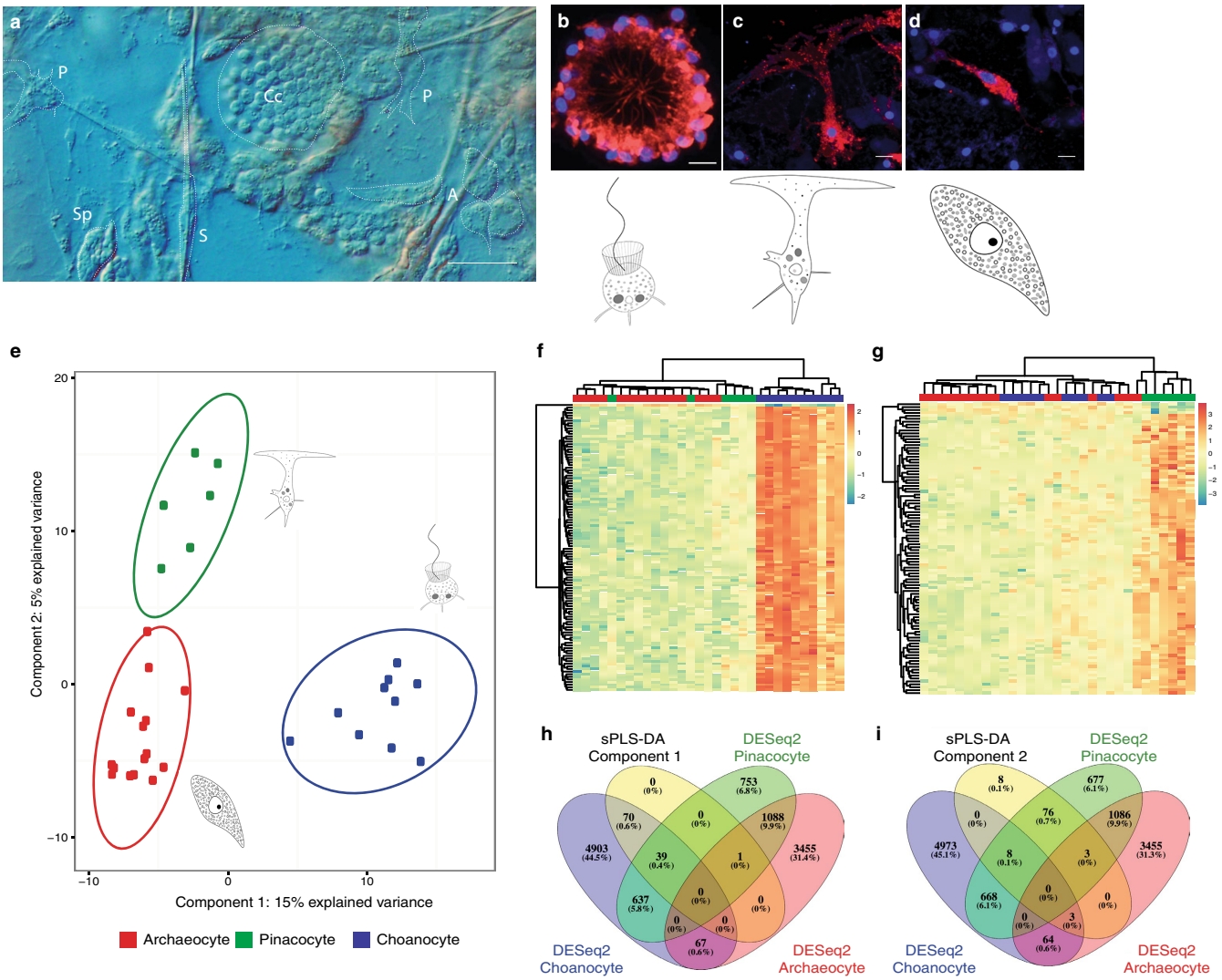
Competing interests The authors declare no competing interests.

Additional information

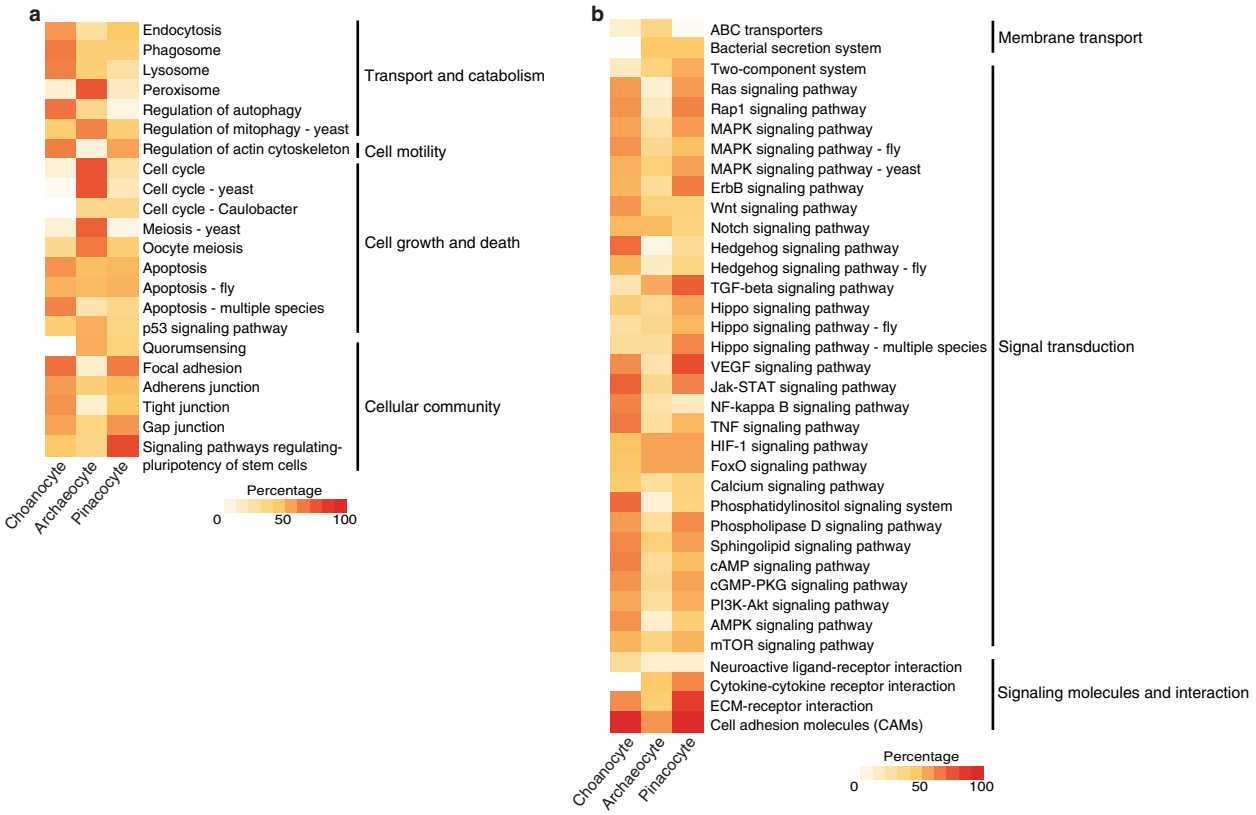
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1290-4.

Correspondence and requests for materials should be addressed to S.M.D. or B.M.D.

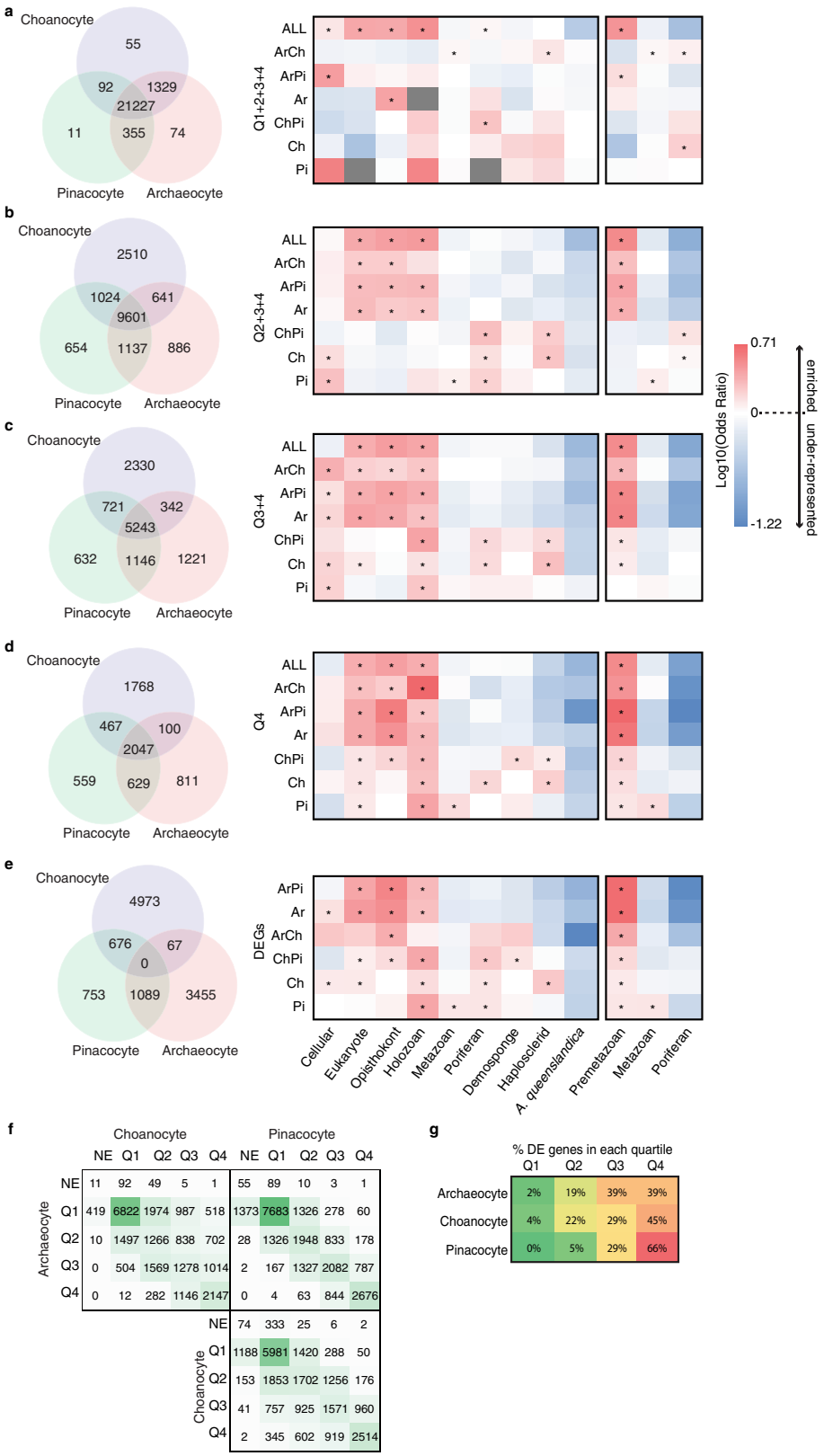
Peer reviewer information *Nature* thanks Alison Cole, Casey Dunn, Mark Martindale, Nori Satoh, Itai Yanai and the other anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | *A. queenslandica* cell types and sPLS-DA of choanocyte, archaeocyte and pinacocyte transcriptomes. a, Whole-mount internal view of a juvenile *A. queenslandica*. Cell types are outlined. A, archaeocyte (cluster of four outlined); Cc, choanocyte chamber; S, sclerocyte; Sp, spherulous cell; P, pinacocyte. **b**, Choanocyte chamber labelled with DiI with an illustration of a single choanocyte below. **c**, Pinacocyte labelled with DiI with illustration below. **d**, Archaeocyte labelled with DiI with illustration below. Scale bars, 10 μm (**b**), 5 μm (**c**, **d**). **e**–**i**, sPLS-DA identified the gene models that best characterize differences in choanocytes (blue, $n = 10$), archaeocytes (red, $n = 15$) and pinacocytes (green, $n = 6$). **e**, Sample plot for the optimal number of gene models that discriminate cell types on the first two components; ellipses indicate 95% confidence intervals. **f**, **g**, Hierarchically clustered heat maps show the expression of the 110 gene models selected for the first component (**f**) and the 98 gene models and 2 long non-coding RNAs selected for the second component (**g**), which accounted for 15% and 5% of explained variance, respectively. **h**, **i**, Venn diagrams summarize the significantly differentially expressed genes identified by the DESeq2 analyses for each cell type and the sPLS-DA on the first (**h**) and the second (**i**) sPLS-DA component. Percentages are of the total number of differentially expressed genes identified from all analyses.



Extended Data Fig. 2 | Percentage of KEGG cellular processes and environmental information processing (that is, cell signalling) genes present in each cell type, corresponding to the number of components that make up each KEGG category identified. a, Cellular processes genes. **b**, Environmental information processing (that is, cell signalling) genes.



f

Archaeocyte \ Choanocyte	NE	Q1	Q2	Q3	Q4
NE	11	92	49	5	1
Q1	419	6822	1974	987	518
Q2	10	1497	1266	838	702
Q3	0	504	1569	1278	1014
Q4	0	12	282	1146	2147

Archaeocyte \ Pinacocyte	NE	Q1	Q2	Q3	Q4
NE	55	89	10	3	1
Q1	1373	7683	1326	278	60
Q2	28	1326	1948	833	178
Q3	2	167	1327	2082	787
Q4	0	4	63	844	2676

Choanocyte \ Pinacocyte	NE	Q1	Q2	Q3	Q4
NE	74	333	25	6	2
Q1	1188	5981	1420	288	50
Q2	153	1853	1702	1256	176
Q3	41	757	925	1571	960
Q4	2	345	602	919	2514

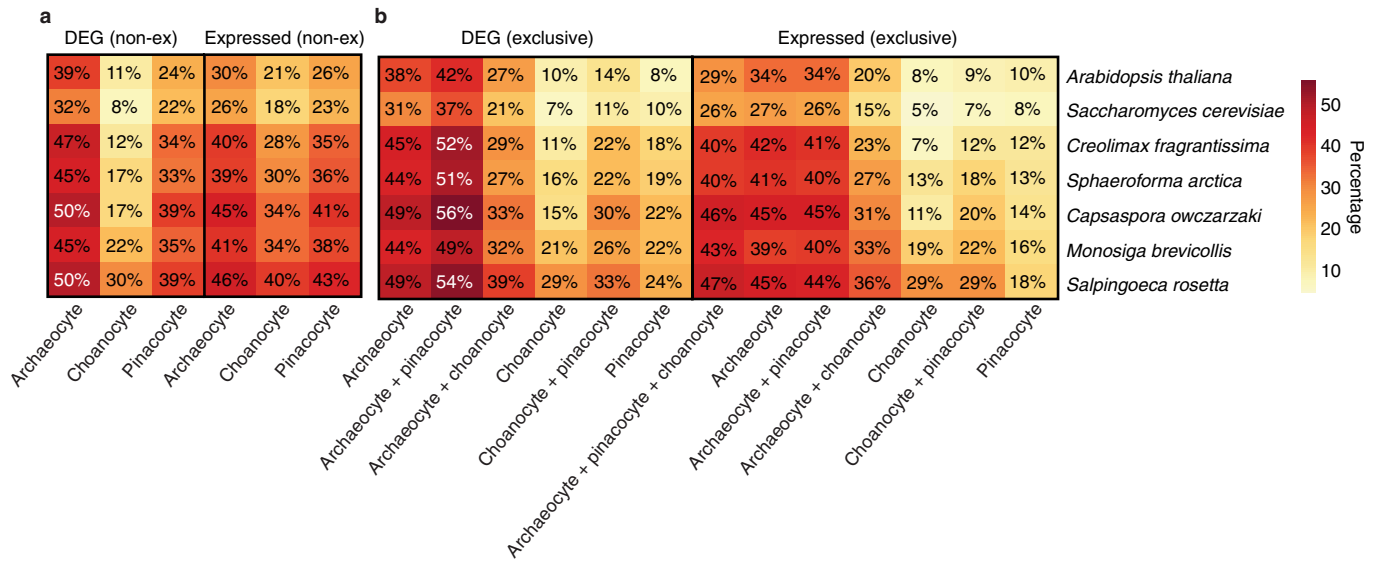
g

% DE genes in each quartile

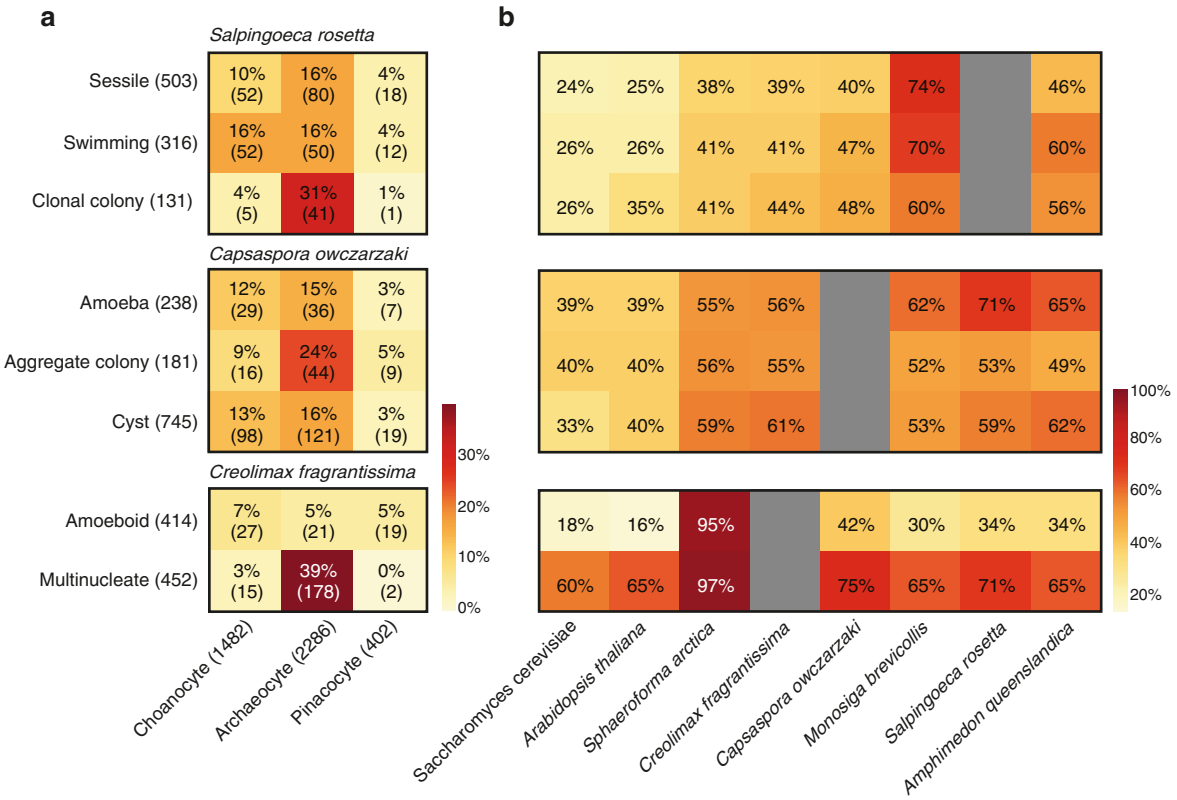
	Q1	Q2	Q3	Q4
Archaeocyte	2%	19%	39%	39%
Choanocyte	4%	22%	29%	45%
Pinacocyte	0%	5%	29%	66%

Extended Data Fig. 3 | See next page for caption.

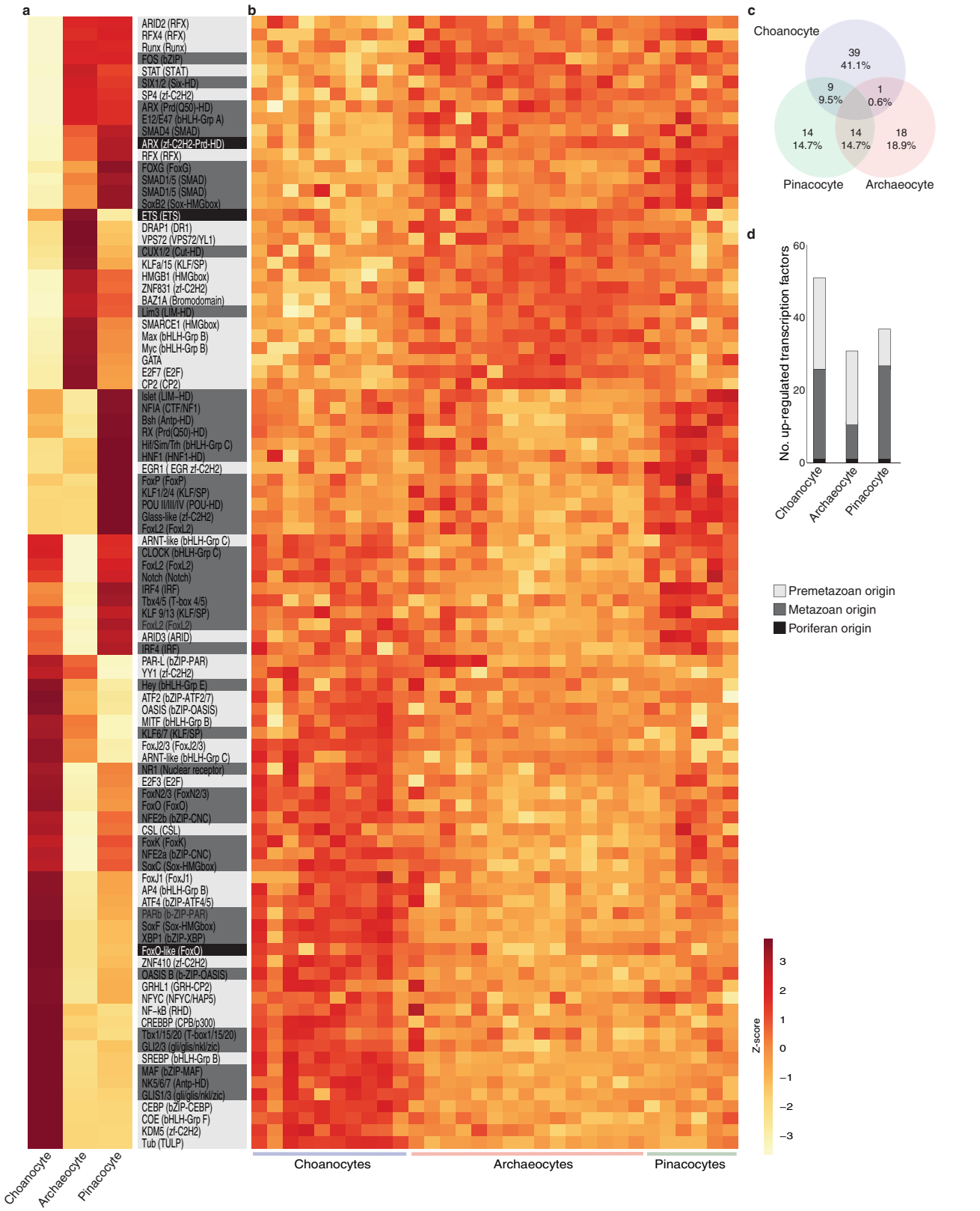
Extended Data Fig. 3 | Evolutionary age of genes expressed in *A. queenslandica* choanocytes, archaeocytes and pinacocytes using different expression thresholds. a–**e**, Phylostratigraphic enrichment of genes expressed in each cell type (Ar, archaeocyte; Ch, choanocyte; Pi, pinacocyte; ArCh, archaeocyte + choanocyte; ArPi, archaeocyte + pinacocyte; ChPi, choanocyte + pinacocyte; ALL, all three cell types combined) at different expression thresholds. Expressed genes are parsed into quartiles based on transcript abundance in each of the cell types. Quartile 1 (Q1) includes the least abundant transcripts and Q4 the most abundant. **a**, Phylostratigraphy enrichment of all genes expressed in each of the cell types (Q1–Q4). **b**, Phylostratigraphy enrichment of genes expressed in the top three quartiles (excluding Q1). **c**, Phylostratigraphy enrichment of genes expressed in the top 50% (Q3 and Q4). **d**, Phylostratigraphy enrichment of the most highly expressed genes (Q4). **e**, For comparison, the evolutionary age of differentially expressed genes identified using differential expression analysis, DESeq2. Heat maps indicate enrichment (log-odds ratio based on a two-sided Fisher's exact test) of phylostrata contained in each gene list in comparison to the *A. queenslandica* genome ($n = 44{,}719$). Asterisks mark significant ($P < 0.05$; Fisher's exact test) overlap between gene lists, indicative of phylostrata enrichment. The heat maps on the far right are collapsed versions of the heat maps on the left, in which the premetazoan category contains phylostrata from cellular to holozoan, and the poriferan category contains phylostrata from poriferan to *A. queenslandica*. To the left of each heat map is a Venn diagram, showing the number of genes in each cell type and set of cell types. Grey boxes on the heat map indicate that there were no genes in that particular gene list characterized by the given phylostrata. See additional supplementary data on Dryad: /ED_Fig. 3_files and /Fig. 3e. **f**, Pairwise comparison illustrating the number of overlapping genes for each of the quartiles between the three cell types. The numbers in the cells are the number of genes common between two cell types (for example, there are 1,569 expressed genes in common between Q2 in choanocytes and Q3 in archaeocytes). NE, not expressed. **g**, The percentage of differentially upregulated genes identified in each of the cell types using DESeq2 in the four quartiles.



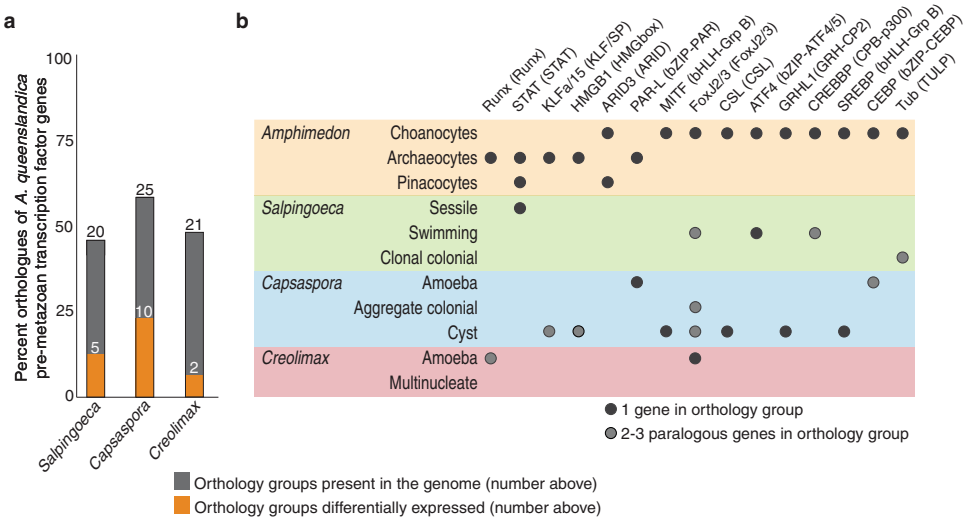
Extended Data Fig. 4 | Orthologues shared between cell-type-specific gene lists and non-metazoan eukaryotes. Heat map showing the percentage of *A. queenslandica* genes with orthogroups shared with select eukaryotes. **a**, Percentage of genes with orthogroups shared between upregulated and total expressed genes from non-exclusive lists (that is, all genes expressed in each of the three cell types, not excluding genes that overlap between any two cell types). **b**, Percentage of genes with orthogroups shared between DEG and total expressed genes-exclusive lists (that is, genes uniquely upregulated or expressed in that cell type).



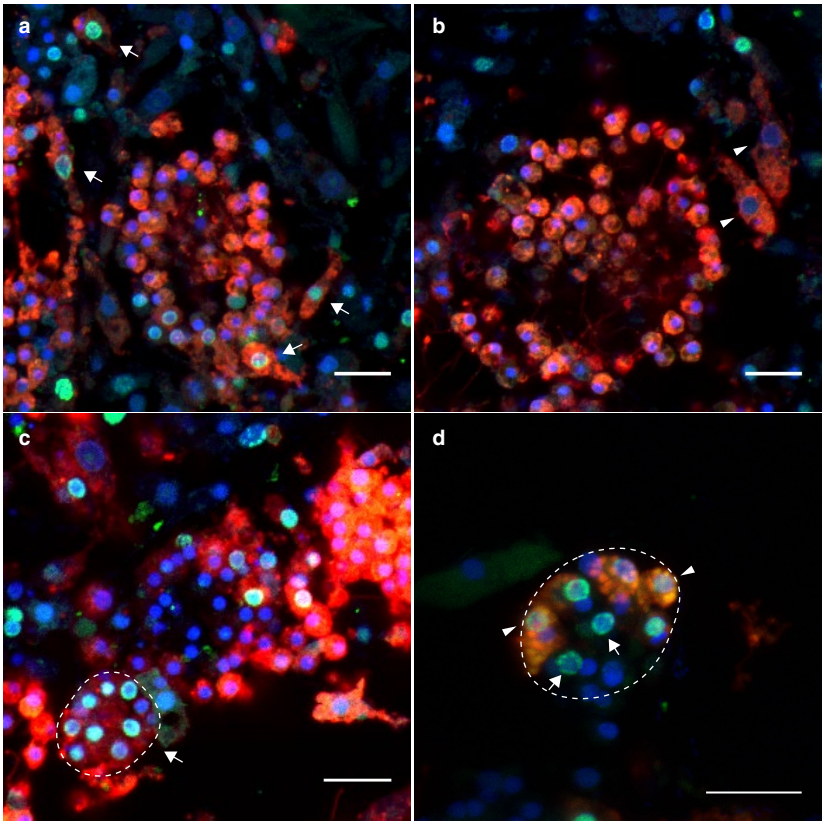
Extended Data Fig. 5 | Orthologues found in *S. rosetta*, *C. owczarzaki* and *C. fragrantissima* life-cycle stages, shared with *A. queenslandica* cell-type transcriptomes and eukaryotic genomes. a, The percentage and number (in parentheses) of differentially expressed orthogroups found in *S. rosetta*, *C. owczarzaki* and *C. fragrantissima* life-cycle stages that are shared with *A. queenslandica* cell types. The numbers in parentheses alongside the unicellular holozoan cell states and sponge cell-type names are the total numbers of orthogroups differentially expressed in that specific gene list. **b**, A heat map showing the percentage of orthogroups shared between genes differentially expressed in *S. rosetta*, *C. owczarzaki* and *C. fragrantissima* life-cycle stages, and genes present in other eukaryotic genomes.



Extended Data Fig. 6 | Heat map of transcription factor genes differentially expressed in choanocytes, archaeocytes and pinacocytes. Ninety-four transcription factor genes that are differentially expressed in *A. queenslandica* cell types are classified on the basis of phylostratum: premetazoan (light grey), metazoan (dark grey) and poriferan (black). **a**, Heat map of expression levels in the three cell types combining all analysed CEL-Seq2 data. Depicted values illustrate scaled (*z*-score) expression levels based on collapsed VST from 10 choanocyte, 15 archaeocyte and 6 pinacocyte transcriptomes. Gene names, families (in parentheses) and phylostrata shading are shown on the right. **b**, Heat map of uncollapsed expression levels (VST) of all transcriptomes (10 choanocyte, 15 archaeocyte and 6 pinacocyte). Rows in **b** correspond to the rows and genes in **a**. **c**, Venn diagram summary of differentially upregulated transcription factor genes between the three cell types using DESeq2. Percentages are of the total transcription factor genes differentially upregulated in all cell types. **d**, Bar graph of the number and distribution of transcription factor genes based on evolutionary age in the three cell types.



Extended Data Fig. 7 | Analysis of premetazoan transcription factors in *A. queenslandica* cells and unicellular holozoan cell states. a, The number and percentage of premetazoan transcription factor orthologues that are present in the genomes of *S. rosetta*, *C. owczarzaki* and *C. fragrantissima*. Percentages are based on the 43 premetazoan genes differentially expressed in the *A. queenslandica* cell types (Extended Data Fig. 5). The number of transcription factor orthologues in the genome is listed above the bar. The orange bar depicts the percent and number of unicellular holozoan premetazoan transcription factor orthologues that are significantly differentially upregulated in at least one cell state. **b**, The 15 premetazoan transcription factor orthology groups (listed along the top) that are significantly upregulated in at least one *A. queenslandica* cell type and one unicellular holozoan cell state. Dots correspond to the cell types and states this occurs. Black dots, orthology group with one gene member; grey dots, orthology group comprising two or more paralogues (see Supplementary File 8 for details).



Extended Data Fig. 8 | Choanocyte dedifferentiation into an archaeocyte does not require cell division. a, **b**, Four-day-old juveniles 6 h after CM-DiI and EdU labelling. **a**, CM-DiI-labelled archaeocytes with EdU incorporation (arrows) found near choanocyte chambers. **b**, Labelled archaeocytes without EdU incorporation (arrowheads), indicating dedifferentiation from choanocytes without cell division. **c**, **d**, Choanocyte-derived archaeocytes are capable of generating new choanocyte chambers. **c**, Four-day-old juvenile 6 h after CM-DiI and EdU labelling. Early choanocyte chamber (dotted line) completely labelled with CM-DiI and EdU, indicating that CM-DiI-labelled archaeocytes with large nuclei are forming this chamber. The absence of cilia and space at the centre of this structure indicates it is not yet a functional choanocyte chamber. **d**, Four-day-old juvenile 12 h after CM-DiI and 6 h after EdU labelling. Early choanocyte chamber (dotted line) with multiple EdU labelled cells, with both CM-DiI-labelled choanocytes (arrowheads) and non-CM-DiI-labelled choanocytes (arrows) indicate multiple cell lineages contributing to the formation of this chamber. The images presented in **a**–**d** represent the consensus of cell behaviours obtained from 10 independent labelling experiments, each comprising a minimum of 24 juveniles. Scale bars, 10 μm.

Extended Data Table 1 | Summary of CEL-Seq2 samples used in this study

Sample	Cell type	Individual	No. cells sequenced	No. reads	Percent reads mapped
1	Archaeocyte	A	5	1098454	21.4
2			5	10411743	67.6
3			5	5699424	60.3
4			6	6759553	72.6
5			5	5673223	64.7
6		B	5	14421299	65.5
7			5	9427170	64.0
8			5	8208828	65.3
9			5	13012311	71.1
10			5	11700365	71.7
11		C	5	25125367	69.8
12			5	15458602	69.7
13			6	16070906	70.6
14			5	20190551	71.0
15			6	22096837	71.7
16	Choanocyte	A	single chamber (40-60 cells)	9657992	49.2
17			single chamber (40-60 cells)	3864298	47.2
18			single chamber (40-60 cells)	7081396	59.7
19		B	single chamber (40-60 cells)	5177297	61.9
20			single chamber (40-60 cells)	6031263	64.9
21		C	single chamber (40-60 cells)	14879156	62.4
22			single chamber (40-60 cells)	12775312	67.0
23			single chamber (40-60 cells)	10569223	66.7
24			single chamber (40-60 cells)	17488774	64.1
25			single chamber (40-60 cells)	18808800	67.1
26	Pinacocyte	A	5	19146512	67.6
27			5	12081597	69.6
28			5	10798371	67.6
29		B	6	2906098	58.1
30		C	5	13792427	60.0
31			5	5184625	66.0

nature research

Corresponding author(s): Bernard Degnan

Last updated by author(s): May 3, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Randomly selected individual cells and choanocyte chambers derived from three randomly-field collected Amphimedon queenslandica were collected under an inverted microscope (Nikon Eclipse Ti microscope) using a micropipette mounted on micromanipulator (MN-4, Narishige) connected to CellTram Oil (Eppendorf). Videos of some of these colection were recorded using the same inverted microscope. Samples were prepared using the CEL-Seq2 protocol (Hashimshony et al., 2016) and sequenced at the Ramaciotti Centre for Genomics (UNSW), Sydney, Australia on Illumina HiSeq2500 on rapid mode using HiSeq Rapid SBS v2 reagents (Illumina). The samples were sequenced in two sequencing runs. Asymmetric paired-end sequencing was performed obtaining 15 bases of sample barcode sequence and 55 bases of mRNA sequence read (Hashimshony et al., 2016). The lipophilic cell tracker CM-DiI (Molecular Probes C7000) was used to label cells for cell-tracking and the thymidine analogue EdU (Click-iT EdU AlexaFluor 488 cell proliferation kit, Molecular Probes C10337) was used to visualize cell proliferation. For all samples, nuclei were labeled with the fluorescent dye 4',6-diamidino-2-phenylindole (DAPI; 1:1,000, Molecular Probes) and mounted using ProlongGold antifade reagent (Molecular Probes). All cell-tracking samples were observed using the ZEISS LSM 510 META confocal microscope, and image analysis was performed using the software ImageJ (versions 1.47-1.51).
Data analysis	CEL-Seq2 reads were processed using a publicly available pipeline for demultiplexing, mapping and counting of transcripts, from the Yanai group at Technion - Israel Institute of Technology (https://github.com/yanailab/CEL-Seq-pipeline). All downstream analyses including differential expression analysis (DESeq2), sparse partial least squares discriminant analysis (mixOmics package), shared orthology analysis (BioConductor package GeneOverlap) and heat map generation (pheatmaps package) was conducted on R or Rstudio. Orthology analysis was performed on FastOrtho. All microscopy images and videos were processed on ImageJ and/or NIS Elements software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Amphimedon queenslandica genome sequence can be accessed at (http://metazoa.ensembl.org/Amphimedon_queenslandica/ Info/Index).
All cell-type transcriptome data are available in the NCBI SRA database under the BioProject PRJNA412708. Additional supplementary data will be available on Dryad Digital Respository: https://doi.org/10.5061/dryad.hp2fr73.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As we sought to determine gene expression differences between three cell types in sponges (archeocytes, pinacocytes and choanocytes) using an RNA-Seq protocol, CEL-Seq2. We collected and sequenced 5 samples of archeocytes and choanocyte chambers, and 3 samples of pinacocytes from three individual sponges, giving a total of 39 samples (15 archeocyte and choanocyte, and 9 pinacocyte samples). Each archeocyte and pinacocyte sample consisted of a pool of 5-6 independently-isolated cells, while each choanocyte sample was a single choanocyte chamber comprised of 40-60 cells.
Data exclusions	As stated in the methods, CEL-Seq2 samples with read counts less than 10e6 were removed and not included in any analyses. Based on previous experience with CEL-Seq2 analyses, read counts below 10e6 yield false quantification of transcript abundance. Initial principal component analysis (PCA) of all CEL-Seq2 samples confirmed that all samples with read counts below 10e6 reads clustered together regarless of cell type, confirming these data should be removed from subsequent analyses.
Replication	Multiple cell type RNA samples for each of the three cell type pools were procured from a dissociated sponge (1-5 replicate cell type pools comprising of 5-40 cells/sponge). This was repeated on three different sponges, with all attempts at replication being successful.
Randomization	As stated in the methods, multiple archeocyte, choanocyte and pinacocyte RNA samples used to generate CEL-Seq2 libraries for sequencing were obtained from three randomly selected sponges from the field. The cells were picked randomly from dissociated sponges. RNA libraries from these samples were randomised over two lanes of Illumina sequencing.
Blinding	Yes, sequencing of CEL-Seq2 libraries (i.e. individual CEL-Seq2 libraries were distribute randomly over two lanes of Illumina HiSeq2500) and initial principle component analyses were performed blind.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	This study did not use laboratory animals.
Wild animals	This study uses cells and juveniles from the demosponge Amphimedon queenslandica collected from Heron Island Reef on the Great Barrier Reef. The sponges were collected under Great Barrier Reef Marine Park Authority permits G12/35053.1 and G16/38120.1 in the names of Bernard Degnan and Sandie Degnan. Ethics permits were not required to undertake this research.
Field-collected samples	Amphimedon queenslandica were maintained (1) at Heron Island Research Station in a flow-through aquaria where they were exposed to ambient water and light conditions; and (2) in a 8000 L closed aquarium in the School of Biological Sciences at the University of Queensland at 26oC in 12/12 light/dark system that mimic coral reef water conditions
Ethics oversight	Ethics permits were not required to undertake this research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.