Pluripotency and the origin of animal multicellularity

Shunsuke Sogabe*[1†], William L. Hatleberg*[1†], Kevin M. Kocot[2], Tahsha E. Say[1], Daniel Stoupin[1†], Kathrein E. Roper[1†], Selene L. Fernandez-Valverde[1†], Sandie M. Degnan[1#] and Bernard M. Degnan[1#]

1. School of Biological Sciences, University of Queensland, Brisbane QLD 4072, Australia

2. Department of Biological Sciences and Alabama Museum of Natural History, The University of Alabama, Tuscaloosa, AL 35487 USA

* These authors contributed equally to this work

Corresponding authors

†Present addresses: The Scottish Oceans Institute, Gatty Marine Laboratory, School of Biology, University of St Andrews, East Sands, St Andrews, Fife KY16 8LB, UK (S.S.); Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213 USA (W.L.H.); BioQuest Studios, PO Box 603, Port Douglas QLD 4877, Australia (D.S.); Centre for Clinical Research, Faculty of Medicine, University of Queensland, Herston QLD 4029, Australia (K.R.); CONACYT, Unidad de Genómica Avanzada, Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, Guanajuato, Mexico (S.L.F.-V.).

The most 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 traditional view of animal origins by comparing the transcriptomes, fates and behaviours of the three primary sponge cell types – choanocytes, pluripotent mesenchymal archeocytes and epithelial pinacocytes – with choanoflagellates and other unicellular holozoans. Unexpectedly, we find 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. In contrast, pluripotent archeocytes up-regulate genes controlling 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 archeocytes, 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 offer no support for the homology of sponge choanocytes and choanoflagellates, nor for 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.

Main

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 (Fig. 1)[5-9]. Recent analyses reveal that unicellular holozoans use similar gene regulatory mechanisms to transit through the different cell states comprising their life cycles[2,5,6,10-12]. These observations suggest that early stem metazoans were more complex than generally 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 *Amphimedon queenslandica* with transcriptomes expressed during the life cycles of the choanoflagellate *Salpingoeca rosetta*, the filasterean *Capsaspora owczarzaki* and the ichthyosporean *Creolimax fragrantissima* (Fig. 1)[10-12]. We chose three sponge somatic cell types hypothesised 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 archeocytes, which inhabit the middle collagenous layer and have a range of other functions (Extended Data Fig. 1 and Supplementary Video 1)[2,14-16]. These three cell types were manually picked and frozen within 15 minutes 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 allowed visual verification of

the three cell types, minimised the time for transcriptional changes to occur after cell dissociation, and allowed for deep sequencing of cell type transcriptomes (Extended Data Table 1, and Supplementary Files S1 and S2).

Principle component analysis (PCA) 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 S3). Significant differences between cell types were independently corroborated by sPLS-DA, which highlighted a subset of 110 genes that explain 15% of the variance in the dataset and clearly discriminate the choanocytes from the other two cell types (Extended Data Fig. 1). This subset includes numerous putative immunity genes that typically encode multiple domains in unique configurations, including scavenger receptor cysteine-rich, tetratricopeptide repeat and epidermal growth factor domains (Supplementary File S4).

We find that archeocytes significantly up-regulate genes involved in the control of cell proliferation, transcription and translation, consistent with their function as pluripotent stem cells (Fig. 2c and Supplementary File S5). In contrast, choanocyte and pinacocyte transcriptomes are enriched for suites of genes involved in cell adhesion, signalling and polarity, consistent with their role as epithelial cells (Fig. 2d; Extended Data Figure 2 and Supplementary File S5).

The evolutionary age of all protein-coding genes in the *Amphimedon* genome, and specifically of genes significantly and uniquely up-regulated 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].

Amphimedon genes were classified as having evolved (i) before or (ii) after divergence of metazoan and choanoflagellate lineages (these are called pre-metazoan and metazoan genes, respectively), or (iii) after divergence of the sponge lineage from all other animals (sponge-specific genes). The *A. queenslandica* genome is comprised of 28% pre-metazoan, 26% metazoan and 46% sponge-specific protein-coding genes (Fig. 3a and Supplementary File S6). We find that 43% of genes significantly up-regulated in choanocytes are sponge-specific, which is similar to the entire genome (Fig. 3b). In contrast, 62% of genes significantly up-regulated in the pluripotent archeocytes belong to the evolutionarily oldest pre-metazoan category, which is significantly higher than 28% for the entire genome (Fig. 3c). As with archeocytes, 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 also into account relative transcript abundances (Extended Data Fig. 3 and Supplementary File S7), 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 archeocytes have a significantly similar transcriptome to the colonial stage of the choanoflagellate and the multinucleate stage of the ichthyosporean (Fig. 3e). Consistent with this result, the significantly up-regulated genes in the colonial or multinucleate stages of all three unicellular holozoans share the highest proportion of orthogroups with genes significantly up-regulated in

archeocytes (Extended Data Fig. 5). In 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 archeocytes (Fig. 3e and Extended Data Fig. 5a).

When we compare the 94 differentially up-regulated transcription factor genes in *A. queenslandica* choanocytes, pinacocytes and archeocytes, we find no marked difference in their phylostratigraphic age, suggesting that the gene regulatory networks in these cells are of an overall similar evolutionary age (Extended Data Fig. 6 and Supplementary File S8). We detected 20, 25 and 21 orthologues of the 43 evolutionarily-oldest (i.e. pre-metazoan) transcription factor genes expressed in the *Amphimedon* cells in the genomes of *Salpingoeca*, *Capsaspora* and *Creolimax* respectively, with 9 of these being present in all species (Supplementary File S8). Comparison of the expression profiles of the transcription factor genes shared among these unicellular holozoans and *Amphimedon* revealed no evidence of a conserved, co-expressed gene regulatory network (Extended Data Fig. 7 and Supplementary File S8). However, the proto-oncogene *Myc* and its heterodimeric partner *Max* are up-regulated in *A. queenslandica* archeocytes (Extended Data Fig. 6), as observed in other metazoan self-renewing pluripotent stem cells[23]. Myc and Max are present also in choanoflagellates, filastereans and ichthyosporeans, where they heterodimerise and bind to E-boxes just as they do in animals[10-12,24]. *Myc* is expressed in the proliferative stage of *Capsaspora*, where it regulates genes associated with ribosome biogenesis and translation[6]. Sponge archeocytes also have enriched expression of genes involved in translation, transcription and DNA replication (Fig. 2c). This suggests that Myc's role in regulating proliferation and differentiation predates its role in bilaterian stem cells and cancer[23,25], and was likely a cardinal feature of the first metazoan cell.

Given that *A. queenslandica* choanocytes and archeocytes express the most derived and ancient transcriptomes, respectively, we investigated the developmental role of these cell types. In *Amphimedon* and most other demosponges, archeocytes 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* archeocytes 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 3 day old juvenile *A. queenslandica* with CM-DiI (Fig. 4a) and followed their fate over 24 hours (Fig. 4b). Within 4 hours of labelling, many choanocytes dedifferentiated into archeocytes (Fig. 4c, d, Supplementary Video 3); this did not require prior cell division (Extended Data Fig. 8). By as little as two hours later, some of these CM-DiI labelled archeocytes had differentiated into pinacocytes (Fig. 4e); within 12 hours, multiple labelled cell types are present (Fig. 4e, f). Together, these results suggest that archeocytes are essential in the development and maintenance of the *A. queenslandica* body plan, as appears to be the case in other sponges[15]. Unlike archeocytes, choanocytes appear late in development and exist in a metastable state, sometimes lasting only a few hours before dedifferentiating back into archeocytes (Fig. 4g, 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, 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 where it could co-exist in multiple states of differentiation, giving rise to the first multicellular animal.

References

1 Cavalier-Smith, T. Origin of animal multicellularity: precursors, causes, consequences - the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20150286 (2015).

4 Nielsen, C. Six major steps in animal evolution: are we derived sponge larvae? *Evol. Dev.* **10**, 241-257 (2008).

5 Sebe-Pedros, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18,** 498-512 (2017).

6 Sebe-Pedros, 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., Tanurdzic, 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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **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. *Invert. Biol.* **123,** 1–22 (2004).

15 Ereskovsky, A. *The Comparative Embryology of Sponges.* Springer, Netherlands (2010).

16 Nakanishi, N., Sogabe, S. & Degnan, B. 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 Genom.* **16**, 387 (2015).

19 Le Cao, K. A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform.* **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**, 1-21 (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. & 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., Diolaiti, D., Conacci-Sorrell, M., Ruiz-Trillo, I., Eisenman, R. N. & King, N. Premetazoan ancestry of the Myc–Max network. *Mol. Biol. Evol.* **28,** 2961–2971 (2011).

25 Kress, T. R., Sabo, 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).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. (This submission includes eight Supplementary Information data files as well as additional material that is available on Dryad.)

Acknowledgements

This study was supported by funds from the Australian Research Council (B.M.D. and S.M.D.). We thank Iñaki Ruiz Trillo for primary expression data for *Capsaspora* and *Creolimax* and Nick 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.R. identified and isolated the cells, and prepared the libraries. W.H., S.S and K.M.K. undertook gene expression and annotation, and phylostratigraphic analyses with help from T.S., S.M.D, S. F.-V and B.M.D. S.S. undertook cell lineage analyses. B.M.D, S.M.D and S.S. wrote the manuscript with comments and contributions from all authors.

Author Information

Competing financial interests

The authors declare no competing financial interests.

Corresponding author

Correspondence and requests for materials should be addressed to
b.degnan@uq.edu.au or s.degnan@uq.edu.au.

Figures Legends

Figure 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.

Figure 2. Comparison of choanocyte, archeocyte and pinacocyte transcriptomes. **a,** PCA plot of CEL-Seq2 transcriptomes with 95% confidence level ellipse plots. Blue, choanocytes (n=10); red, archeocytes (n=15); green, pinacocytes (n=6). **b,** Venn diagram summary of the number of significantly up-regulated genes based on pairwise comparisons between each of the three cell types using a Negative Binomial distribution in DESeq2 with a false discovery rate (FDR) < 0.05. The percentages are of the total genes differentially up-regulated in all cell types. **c,** Percentage of KEGG Genetic Information Processing genes present in each cell type, corresponding to the number of components making up each KEGG category identified. **d,** Scaled heat map illustrating

293 the expression (Z-score) of *Amphimedon* epithelial cell polarity, junction and basal
294 lamina genes in each cell type. Expression based on collapsed count values using the
295 variance stabilising transformation (vst), which was blind to the experimental design.
296
297 **Figure 3. Analysis of gene age of choanocyte, archeocyte and pinacocyte**
298 **transcriptomes.**
299 **a,** Phylostratigraphic estimate of the evolutionary age of coding genes in the *A.*
300 *queenslandica* genome. **b-d,** Estimate of gene age of differentially-expressed genes in
301 choanocytes (b), archeocytes (c) and pinacocytes (d) and the enrichment of phylostrata
302 relative to the whole genome (bottom). Asterisks indicate significant difference (Two-
303 sided Fisher's exact test p-value <0.001) from the whole genome. The enrichment
304 values (log-odds ratio) for (b) choanocytes (n=10) are sponge specific (-0.0089, p-
305 value=0.7747), metazoan (-0.0361, p-value=0.9958) and premetazoan (0.0439, p-
306 value=0.0004) genes; (c) archeocytes (n=15) are sponge specific (-0.5634, p-
307 value=1.33e-133), metazoan (-0.1923, p-value=1.04e-18) and premetazoan (0.6772, p-
308 value=0); and (d) pinacocytes (n=6) are sponge specific (-0.2173, p-value=5.23e-13),
309 metazoan (-0.0008, p-value=0.5231) and premetazoan (0.2359, p-value=3.07e-36). **e,** A
310 heat map comparing uniquely up-regulated genes in *A. queenslandica* cell types
311 (n=18,774) that are orthologous (orthology group, OG) to genes expressed during
312 different life stages of *Salpingoeca rosetta* (n=10,350 OGs), *Capsaspora owczarzaki*
313 (n=9,492 OGs) and *Creolimax fragrantissima* (n=11,449 OGs). Colour indicates the
314 significance of overlap in transcriptional profiles based on the odds ratio. Values
315 indicate adjusted p-values and show significant resemblance only between the
316 archeocyte and the *S. rosetta* colonial stage and the *C. fragrantissima* multinucleate
317 stage transcriptomes. N.s., not significant.

Figure 4. Transdifferentiation of choanocytes in *Amphimedon queenslandica*.
a, b, Whole mount views of 4 day old juveniles labelled with CM-DiI. **a,** 30 min after CM-DiI labelling; arrows, representative labelled choanocyte chambers. **b,** 24 hours after labelling. CM-DiI labelling spread from choanocyte chambers at 30 min to throughout the juvenile at 24 hours 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,** 2 hours (c) and 4 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 archeocytes. **e,** 6 hours after labelling, CM-DiI labelled pinacocytes (arrow) with thin pseudopodia are present. **f,** 12 hours after initial labelling, labelled sclerocytes (arrow) and other cell types are present. The images presented in a-f represent the consensus 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: a, b, 200 μm; c-f, 10 μm.

Methods

Cell isolation

Three random adult *Amphimedon 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 based on morphology. Archeocytes 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 micromanipulator (MN-4, Narishige) connected to CellTram Oil (Eppendorf) (Supplementary video 2), flash frozen and stored at -80^{o}C. All cells were frozen within 15 min of dissociation. Samples used in CEL-Seq2 were comprised of pools of either five to six archeocytes or pinacocytes, or a single choanocyte chamber (~40-60 cells) (Extended Data Table 1). Based on differences in cell size, we estimated that these pools have similar amounts of total RNA. Three pinacocyte, and five archeocyte and choanocyte samples were randomly collected from each of three sponges (Supplementary File S2).

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 v2 reagents (Illumina); CEL-Seq2 libraries were randomised 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/). Read counts were obtained from demultiplexed

reads mapped to *A. queenslandica* Aqu2.1 gene models[18]. Samples with read counts less than 10^6 were removed and not included in subsequent analyses (Supplementary File S2). 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/). Genes that had read counts with a row sum of zero were removed. Principle component analyses (PCA) were performed on blind variance stabilising transformed (vst) counts obtained using DESeq2 and were visualised 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 (FDR) < 0.05. Venn diagrams were generated using VENNY (http://bioinfogp.cnb.csic.es/tools/venny) to visualise and compare the list of DEGs between each cell type. Heat maps were generated using the R-packages pheatmap[33] and RColorBrewer[34] to visualise the expression patterns between the cell types using the vst transformed 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 = 1e-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). 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 S3, as well as KEGG pathway enrichments in Supplementary File S5.

To identify the genes that best explain differences among cell type transcriptomes, we adopted the multivariate sparse Partial Least Squares Discriminant Analysis (sPLS-DA)[19], implemented in the mixOmics package[39] in R v3.3.1 (see additional supplementary data on Dryad: /sPLS-DA/README.txt). 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 optimised numbers of genes per component were obtained by training and correctly evaluating the performance of the predictive model using 5-fold cross-validation, repeated 100 times. A sample plot was used to visualise 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 visualise 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 visualise and compare the DEGs generated by DESeq2 and sPLS-DA.

Phylostratigraphy

To estimate the evolutionary age of genes up-regulated 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 *A. queenslandica*'s phylogenetic position (i.e. 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 S6, see additional supplementary data on Dryad: /Phylostratigraphy annotations/). Every gene model in *A. queenslandica* was

418 blasted against each sequence in the database, and its age of gene origin was inferred
419 based on the oldest blast hit relative to a predetermined phylogenetic tree (see
420 additional supplementary data on Dryad: /Phylostratigraphy annotations/).
421 Phylostrata enrichments were performed using the Fisher's exact test[41] in the
422 BioConductor package, GeneOverlap[42] in R, to identify significant differences in gene
423 age of the cell type DEG lists relative to the genome (see additional supplementary data
424 on Dryad: /Fig.3b-d and /ED_Fig3_files). Enrichment (log odds ratio value above 0) and
425 under-representation (log odds ratio value below 0) of each phylostrata found in the
426 cell type DEG lists relative to the genome, were visualised using the R-packages
427 pheatmap[33] and RColorBrewer[34].
428
429 **Orthology analyses**
430 Orthology analyses were performed using FastOrtho[43] from a custom 'all-vs-all' blastp
431 database of coding sequences from the genomes of *Saccharomyces cerevisiae*[44],
432 *Arabidopsis thaliana*[45], *Creolimax fragrantissima*[12], *Sphaeroforma arctica*[46], *Capsaspora*
433 *owczarzaki*[47], *Monosiga brevicollis*[48], and *Salpingoeca rosetta*[10], using the following
434 configuration settings: pv_cutoff = 1e-5; pi_cutoff = 0.0; pmatch_cutoff = 0.0;
435 maximum_weight = 316.0; inflation = 1.5; blast_e = 1e-5 (see additional supplementary
436 data on Dryad: /FastOrtho/). FastOrtho classifies all of the genes present in each
437 genome into orthology groups (orthogroups, OGs), which contain all orthologous and
438 paralogous genes from each species. Genes that do not have any orthologues in other
439 species or paralogues within the same genome were not included in any orthogroups.
440 To compare the gene lists between species in all downstream analyses, species-specific
441 gene names were changed to the common orthogroup identifier.

442 Orthology analyses between *A. queenslandica* and *S. rosetta*, *C. fragrantissima*, and *C.*
443 *owczarzaki* cell types were performed using the cell type-specific DEG lists obtained
444 from previous studies on *S. rosetta*[10], *C. fragrantissima*[12], and *C. owczarzaki*[11]. The
445 BioConductor package, GeneOverlap[42], was used to identify (1) the number overlapping
446 OGs between species and cell type, and (2) the statistical significance of that overlap
447 based on list size and total number of OGs (see additional supplementary data on Dryad:
448 /Fig.3e). This function provided the odds ratio between the OG lists, where the null
449 hypothesis was no significant overlap (odds ratio value of 1 or smaller) and the
450 alternative being a significant overlap detected between the lists (odds ratio value over
451 1), as well as a p-value calculated for odds ratio values over 1.
452 To supplement phylostratigraphy analyses of *Amphimedon* cell-type specific gene
453 lists (Fig. 3 and Extended Data Fig. 3), the BioConductor package, GeneOverlap[42] was
454 used to identify the number and percentage of orthogroups that are also present in the
455 genomes of *Arabidopsis thaliana, Saccharomyces cerevisiae, Creolimax fragrantissima,*
456 *Sphaeroforma arctica, Capsaspora owczarzaki, Monosiga brevicollis, and Salpingoeca*
457 *rosetta* (Extended Data Fig. 4 and Extended data Fig. 5; see additional supplementary
458 data on Dryad: /ED_Fig4 and ED_Fig5)
459
460 **Classification of gene expression levels into quartiles**
461 In addition to differential gene expression analyses for *Amphimedon* transcriptomes, the
462 relative gene expression levels for all cell types were assigned to one of four expression
463 quartiles based on the number of reads that mapped to a given Aqu2.1 gene model
464 (Extended data Fig. 3). All zero read counts were discarded and the mean expression
465 value of the non-transformed normalised count values of all samples (from all cell
466 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_Fig3_files). 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 (OG) in each gene list was calculated (see additional supplementary data on Dryad: /ED_Fig4_files and ED_Fig5_files). In these analyses, exclusive lists refer to all of the regions in the Venn diagram being treated as a separate list (e.g. archeocyte only, common between archeocyte and choanocyte, common between archeocyte and pinacocyte, etc.), while non-exclusive lists collapse all of the lists containing a given cell type into one list (e.g. archeocyte non-exclusive DEG list includes, archeocyte DEGs + (archeocyte + pinacocyte DEGs) + (archeocyte + 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 DNA-binding domain (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[7,16,47,50-66] (Supplementary File S8). Third, we compared these lists to an unpublished in-house

database for *A. queenslandica* (Degnan *et al*. unpublished) 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 two of the three lists (Supplementary File S8).

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 S8). From this, each TF 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[67], left in FSW overnight and then placed in sterile 6-well plates with 10 ml of FSW for 1 hour in the dark with live coralline algae *Amphiroa fragilissima*. Postlarvae settled on *A. fragilissima* were removed using fine forceps (Dumont #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 hours[16,68]. 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 minutes to 1 hour. This minimised 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

517 epifluorescence microscopy (Nikon Eclipse Ti microscope) immediately after CM-DiI
518 was washed out, with juveniles detected with CM-DiI labelled cells outside of
519 choanocyte chambers discarded from the study. Juveniles were allowed to develop for
520 0, 2, 4, 6, 12 or 24 hours post-incubation (hpi) with CM-DiI, then washed in FSW three
521 times for 5 minutes and fixed[69] without dehydration in ethanol. Fixed juveniles were
522 washed three times in MOPST (1x MOPS buffer + 0.1% Tween). Nuclei were labelled
523 with DAPI (1:1,000, Molecular Probes) for 30 minutes, washed in MOPST for 5 minutes
524 and mounted using ProlongGold antifade reagent (Molecular Probes). All samples were
525 observed using the ZEISS LSM 710 META confocal microscope, and image analysis was
526 performed using the software ImageJ.
527 To visualise cell proliferation, the thymidine analogue EdU (Click-iT EdU AlexaFluor
528 488 cell proliferation kit, Molecular Probes C10337) was used as previously
529 described[16,26]. To label S-phase nuclei, juveniles were incubated in FSW containing 200
530 µM EdU for 6 hours, washed in FSW and immediately fixed as described above.
531 Fluorescent labelling of incorporated EdU was conducted according to the
532 manufacturer's recommendations prior to DAPI labelling and mounting in ProLong Gold
533 antifade reagent as described above.
534
535 **Data Availability Statement**
536 *Amphimedon queenslandica* genome sequence can be accessed at
537 (http://metazoa.ensembl.org/Amphimedon_queenslandica/ Info/Index).
538 All cell-type transcriptome data are available in the NCBI SRA database under the
539 BioProject PRJNA412708. Additional supplementary data are available from the Dryad
540 Digital Respository: https://doi.org/10.5061/dryad.hp2fr73.
541

References

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. Package 'pheatmap'. https://cran.r-project.org/package=pheatmap (2012).

34 Neuwirth, E. Package 'RColorBrewer'. 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**, D457-D462 (2015).

39 Rohart, F., Gautier, B., Singh, A. & Le 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. (2014).

43 Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581-591 (2014).

44 Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710-D716 (2016).

45 The 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**, D88-92 (2008).

50 Babonis, L. S. & Martindale, M. Q. Phylogenetic evidence for the modular evolution of metazoan signalling pathways. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 20150477 (2017).

590 51 Srivastava, M. *et al.* Early evolution of the LIM homeobox gene family. *BMC Biol.* **8**, 4
591 (2010).
592 52 Larroux, C. *et al.* Genesis and expansion of metazoan transcription factor gene
593 classes. *Mol. Biol. Evol.* **25**, 980-996 (2008).
594 53 Larroux, C. *et al.* Developmental expression of transcription factor genes in a
595 demosponge: insights into the origin of metazoan multicellularity. *Evol. Dev.* **8**, 150-
596 173 (2006).
597 54 Shimeld, S. M., Degnan, B. & Luke, G. N. Evolutionary genomics of the Fox genes:
598 origin of gene families and the ancestry of gene clusters. *Genomics* **95**, 256-260
599 (2010).
600 55 Layden, M. J., Meyer, N. P., Pang, K., Seaver, E. C. & Martindale, M. Q. Expression and
601 phylogenetic analysis of the *zic* gene family in the evolution and development of
602 metazoans. *EvoDevo* **1**, 12 (2010).
603 56 Presnell, J. S., Schnitzler, C. E. & Browne, W. E. KLF/SP transcription factor family
604 evolution: Expansion, diversification, and innovation in eukaryotes. *Genome Biol.*
605 *Evol.* **7**, 2289-2309 (2015).
606 57 Mukhopadhyay, S. & Jackson, P. K. The tubby family proteins. *Genome Biol.* **12**, 225
607 (2011).
608 58 Larroux, C. *et al.* The NK homeobox gene cluster predates the origin of Hox genes.
609 *Curr. Biol.* **17**, 706-710 (2007).
610 59 Wang, L., Tang, Y., Cole, P. A. & Marmorstein, R. Structure and chemistry of the
611 p300/CBP and Rtt109 histone acetyltransferases: Implications for histone
612 acetyltransferase evolution and function. *Curr. Opin. Struct. Biol.* **18**, 741-747
613 (2008).

60 Petroni, K. *et al.* The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* **24**, 4777-4792 (2012).

61 Morrison, A. J. & Shen, X. Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. *Nat. Rev. Mol. Cell Biol.* **10**, 373-384 (2009).

62 Jones, M. H., Hamana, N., Nezu, J. & Shimane, M. A novel family of bromodomain genes. *Genomics* **63**, 40-45 (2000).

63 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).

64 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).

65 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).

66 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).

67 Leys, S. P. *et al.* Isolation of *Amphimedon* developmental material. *Cold Spring Harb. Protoc.* **3**, 5095 (2008).

68 Degnan, B. M. *et al.* Porifera. *Evolutionary Developmental Biology of Invertebrates vol.1* (Springer, 2015).

69 Larroux, C. *et al.* Whole-mount in situ hybridization in *Amphimedon*. *Cold Spring Harb. Protoc.* **3**, 5096 (2008).

Extended Data Figure Legends

Extended Data Figure 1: *Amphimedon queenslandica* cell types and sparse partial least squares discriminant analysis (sPLS-DA) of choanocyte, archeocyte and pinacocyte transcriptomes.

a, Whole mount internal view of a juvenile *Amphimedon queenslandica*. Cell types are outlined. A, archeocyte (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,** Archeocyte labelled with DiI with illustration below. Scale bars: b, 10 μm; c-e, 5 μm. **e-i,** A supervised multivariate analysis, sPLS-DA, identified the gene models that best characterise differences in choanocytes (blue, n=10), archeocytes (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 (f) the 110 gene models selected for the first component, and (g) the 98 gene models and 2 long non-coding RNAs selected for the second component, which accounted for 15% and 5% of explained variance, respectively. **h, i,** Venn diagrams summarise the significantly differentially expressed genes identified by the DESeq2 analyses, for each cell type, and the sPLS-DA on (h) the first and (i) the second sPLS-DA component. Percentages are of the total number of differentially expressed genes identified from all analyses.

Extended Data Figure 2: Percentage of KEGG cellular processes and environmental information processing (i.e. cell signalling) genes present in each

cell type, corresponding to the number of components making up each KEGG category identified.

a, Cellular processes genes. **b,** Environmental information processing (i.e. cell signalling) genes.

Extended Data Figure 3: Evolutionary age of genes expressed in *Amphimedon queenslandica* choanocytes, archeocytes and pinacocytes using different expression thresholds.

a-e, Phylostratigraphic enrichment of genes expressed in each cell type (Ar, archeocyte; Ch, choanocyte; Pi, pinacocyte; ArCh, archeocyte + choanocyte; ArPi, archeocyte + 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 (i.e. Q1-Q4). **b,** Phylostratigraphy enrichment of genes expressed in the top three quartiles (i.e. excluding Q1). **c,** Phylostratigraphy enrichment of genes expressed in the top 50% (i.e. Q3 and Q4). **d,** Phylostratigraphy enrichment of the most highly expressed genes (i.e. 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, where the premetazoan category contains phylostrata from cellular to holozoan, and the

688 poriferan category contains phylostrata from poriferan to *A. queenslandica*. To the left of
689 each heat map is a Venn diagram, showing the number of genes in each cell type and
690 sets of cell types. Grey boxes on the heat map indicate that there were no genes in that
691 particular gene list characterised by the given phylostrata. See additional
692 supplementary data on Dryad: /ED_Fig3_files and /Fig.3e. **f,** Pairwise comparison
693 illustrating the number of overlapping genes for each of the quartiles between the three
694 cell types. The numbers in the cells are the number of genes common between two cell
695 types (e.g. there are 1569 expressed genes in common between Q2 in choanocytes and
696 Q3 in archeocytes). NE, not expressed. **g,** The percentage of differentially up-regulated
697 genes identified in each of the cell types using DESeq2 in the four quartiles.
698
699 **Extended Data Figure 4: Orthologues shared between cell type-specific gene lists**
700 **and non-metazoan eukaryotes.**
701 Heat map showing the percentage of *A. queenslandica* genes with orthogroups (OGs)
702 shared with select eukaryotes. **a,** Percentage of genes with OGs shared between up-
703 regulated and total expressed genes from non-exclusive lists (i.e. all genes expressed in
704 each of the three cell types, not excluding genes that overlap between any two cell
705 types). **b,** Percentage of genes with OGs shared between DEG and total expressed genes
706 - exclusive lists (i.e. genes uniquely up-regulated or expressed in that cell type).
707
708 **Extended Data Figure 5: Orthologues found in *Salpingoeca rosetta*, *Capsaspora***
709 ***owczarzaki* and *Creolimax fragrantissima* life cycle stages, shared with *A.***
710 ***queenslandica* cell type transcriptomes and eukaryotic genomes.**
711 **a,** The percent and number (in parentheses) of differentially expressed OGs found in
712 *Salpingoeca rosetta*, *Capsaspora owczarzaki* and *Creolimax fragrantissima* life cycle

stages that are shared with *Amphimedon queenslandica* cell types. The numbers in parentheses alongside the unicellular holozoan cell states and sponge cell type names is the total number of OGs differentially expressed in that specific gene list. **b,** A heatmap showing the percentage of OGs shared between genes differentially expressed in *Salpingoeca rosetta*, *Capsaspora owczarzaki* and *Creolimax fragrantissima* life cycle stages, and genes present in other eukaryotic genomes.

Extended Data Figure 6: Heat map of transcription factor genes differentially expressed in choanocytes, archeocytes and pinacocytes.

94 transcription factor genes that are differentially expressed in *A. queenslandica* cell types are classified based on 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 variance stabilising transformation (vsd), from 10 choanocyte, 15 archeocyte 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 archeocyte and 6 pinacocyte). Rows in b correspond to the rows and genes in a. **c,** Venn diagram summary of differentially up-regulated transcription factor genes between the three cell types using DESeq2. Percentages are of the total transcription factor genes differentially up-regulated 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 Figure 7: Analysis of premetazoan transcription factors in *Amphimedon* cells and unicellular holozoan cell states.

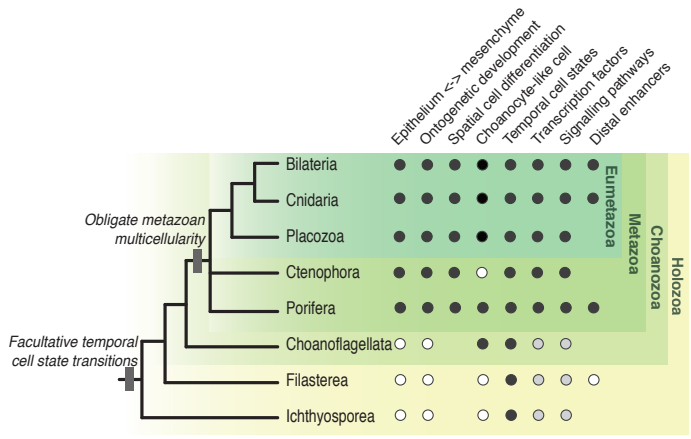
a, The number and percentage of premetazoan transcription factor orthologues that are present in the genomes of *Salpingoeca rosetta*, *Capsaspora owczarzaki* and *Creolimax fragrantissima*. Percentages are based on the 43 premetazoan genes differentially expressed in the *A. queenslandica* cell types (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 up-regulated in at least one cell state. **b,** The 15 premetazoan transcription factor orthology groups (listed along the top) that are significantly up-regulated in at least one *Amphimedon* 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 comprised of two of more paralogues (see Supplementary File S8 for details).

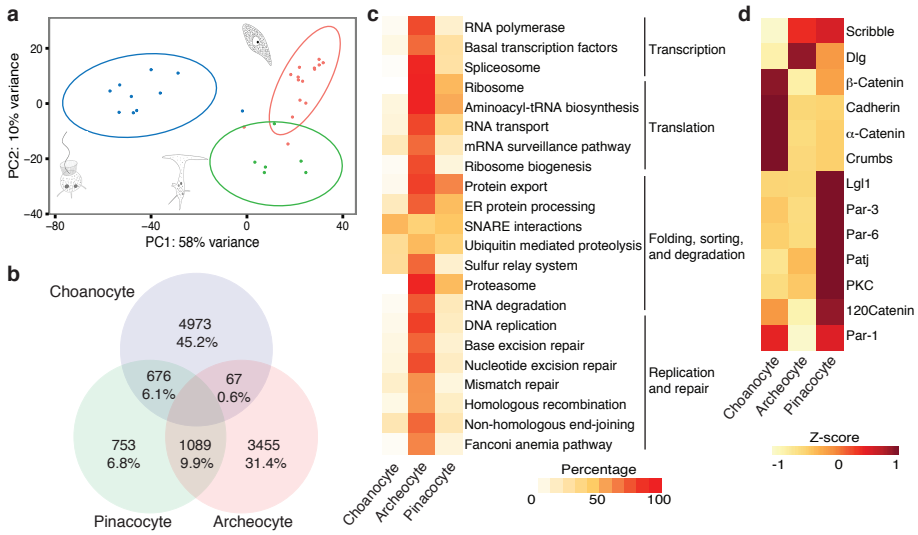
Extended Data Figure 8: Choanocyte dedifferentiation into an archeocyte does not require cell division.

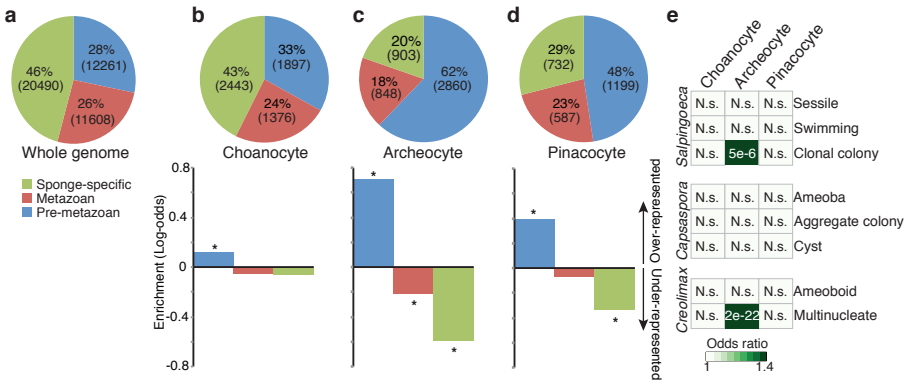
a, b, 4 day old juveniles 6 hours after CM-DiI and EdU labelling. **a,** CM-DiI labelled archeocytes with EdU incorporation (arrows) found near choanocyte chambers. **b,** Labelled archeocytes without EdU incorporation (arrowheads), indicating dedifferentiation from choanocytes without cell division. Scale bars: 10 μm.

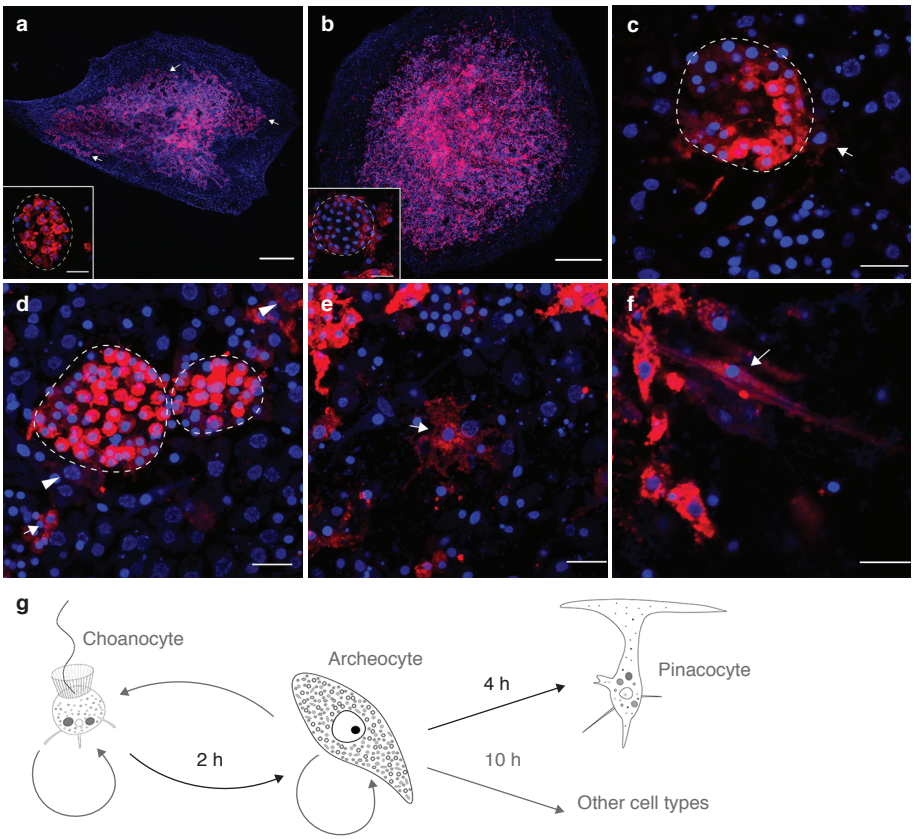
c, d, Choanocyte-derived archeocytes are capable of generating new choanocyte chambers. **c,** 4 day old juvenile 6 hours after CM-DiI and EdU labelling. Early choanocyte chamber (dotted line) completely labelled with CM-DiI and EdU, indicating CM-DiI labelled archeocytes, with large nuclei, are forming this chamber. The absence of cilia and space at the center of this structure indicates it is not yet a functional choanocyte chamber. **d,** 4 day old juvenile 12 hours after CM-DiI and 6 hours after EdU labelling.

763 Early choanocyte chamber (dotted line) with multiple EdU labelled cells, with both CM-
764 DiI labelled choanocytes (arrowheads) and non-CM-DiI labelled choanocytes (arrows)
765 indicate multiple cell lineages contributing to the formation of this chamber. The images
766 presented in a-d represent the consensus cell behaviours obtained from 10 independent
767 labelling experiments, each comprising a minimum of 24 juveniles. Scale bars: **a-d,** 10
768 μm.
769
770 **Extended Data Table 1: Summary of CEL-Seq2 samples used in this study.**
771
772
773
774
775


Epithelium <-> mesenchyme
Ontogenetic development
Spatial cell differentiation
Choanocyte-like cell
Temporal cell states
Transcription factors
Signalling pathways
Distal enhancers
Bilateria
Cnidaria
Placozoa
Ctenophora
Porifera
Choanoflagellata
Filasterea
Ichthyosporea
Obligate metazoan multicellularity
Facultative temporal cell state transitions
Eumetazoa
Metazoa
Choanozoa
Holozoa


a
PC2: 10% variance
PC1: 58% variance
b
Choanocyte
4973
45.2%
676
6.1%
67
0.6%
753
6.8%
1089
9.9%
3455
31.4%
Pinacocyte
Archeocyte
c
RNA polymerase
Basal transcription factors
Spliceosome
Ribosome
Aminoacyl-tRNA biosynthesis
RNA transport
mRNA surveillance pathway
Ribosome biogenesis
Protein export
ER protein processing
SNARE interactions
Ubiquitin mediated proteolysis
Sulfur relay system
Proteasome
RNA degradation
DNA replication
Base excision repair
Nucleotide excision repair
Mismatch repair
Homologous recombination
Non-homologous end-joining
Fanconi anemia pathway
Transcription
Translation
Folding, sorting, and degradation
Replication and repair
Choanocyte
Archeocyte
Pinacocyte
Percentage
d
Scribble
Dlg
β-Catenin
Cadherin
α-Catenin
Crumbs
Lgl1
Par-3
Par-6
Patj
PKC
120Catenin
Par-1
Choanocyte
Archeocyte
Pinacocyte
Z-score


a
46% (20490)
28% (12261)
26% (11608)
Whole genome
Sponge-specific
Metazoan
Pre-metazoan
b
43% (2443)
33% (1897)
24% (1376)
Choanocyte
c
20% (903)
18% (848)
62% (2860)
Archeocyte
d
29% (732)
48% (1199)
23% (587)
Pinacocyte
Enrichment (Log-odds)
0.8
0.4
0
-0.4
-0.8
Over-represented
Under-represented
e
Choanocyte
Archeocyte
Pinacocyte
Salpingoeca
N.s. N.s. N.s. Sessile
N.s. N.s. N.s. Swimming
N.s. 5e-6 N.s. Clonal colony
Capsaspora
N.s. N.s. N.s. Ameoba
N.s. N.s. N.s. Aggregate colony
N.s. N.s. N.s. Cyst
Creolimax
N.s. N.s. N.s. Ameboid
N.s. 2e-22 N.s. Multinucleate
Odds ratio
1
1.4


a
b
c
d
e
f
g
Choanocyte
Archeocyte
Pinacocyte
4 h
2 h
10 h
Other cell types